Published in final edited form as:

Health Place. 2015 September ; 35: 136–146. doi:10.1016/j.healthplace.2015.08.002.

Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments

Claudia Nau, PhD¹,

Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe Street, Baltimore, MD-21205, phone: 814-404-1066

Hugh Ellis, PhD^{1,2},

Johns Hopkins Whiting School of Engineering, 3400 North Charles Street, Baltimore MD-21218

Hongtai Huang, PhD¹,

Johns Hopkins Whiting School of Engineering, 3400 North Charles Street, Baltimore MD-21218

Brian S. Schwartz, MD^{1,3},

Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe Street, Baltimore, MD-21205

Annemarie Hirsch, PhD³,

Geisinger Center for Health Research, 100 North Academy Avenue Danville, PA-1728

Lisa Bailey-Davis, PhD³,

Geisinger Center for Health Research, 100 North Academy Avenue Danville, PA-1728

Amii M. Kress, PhD¹,

Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe

Jonathan Pollak, MA¹, and

Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe

Thomas A. Glass, PhD¹

Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe

Claudia Nau: cnau1@jhu.edu

Abstract

Correspondence to: Claudia Nau, cnau1@jhu.edu.

¹Johns Hopkins Bloomberg School of Public Health Global Obesity Prevention Center, 615 N Wolfe Street, Baltimore, MD-21205

²Johns Hopkins Whiting School of Engineering, 3400 North Charles Street, Baltimore MD-21218

³Geisinger Center for Health Research, 100 North Academy Avenue Danville, PA-1728

CONFLICT OF INTEREST: none to declare

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Past research has assessed the association of single community characteristics with obesity, ignoring the spatial co-occurrence of multiple community-level risk factors. We used conditional random forests (CRF), a non-parametric machine learning approach to identify the combination of community features that are most important for the prediction of obesogenic and obesoprotective environments for children. After examining 44 community characteristics, we identified 13 features of the social, food, and physical activity environment that in combination correctly classified 67% of communities as obesoprotective or obesogenic using mean BMI-z as a surrogate. Social environment characteristics emerged as most important classifiers and might provide leverage for intervention. CRF allows consideration of the neighborhood as a system of risk factors.

Keywords

obesogenic environments; childhood obesity; conditional random forest; physical activity features; food features; social features

INTRODUCTION

The concept of the “obesogenic environment” was first proposed in the late 1990’s (Hill and Peters, 1998; Poston and Foreyt, 1999; Swinburn et al., 1999) as a framework for understanding the joint impact of multiple dimensions of place on obesity risk. Through their physical, institutional, or social features, obesogenic environments impede healthy energy balance-related behaviors by promoting inactivity and excess caloric intake. Since the concept was proposed, a rich body of research has linked numerous environmental characteristics with obesity in a variety of populations.

Multilevel studies have shown associations of several features of the built environment such as land use mix and population density (Frank et al., 2007; Franzini et al., 2009; Rundle et al., 2009; Schwartz et al., 2011b), as well as food establishments (Casey et al., 2008; Cummins and Macintyre, 2006; Drewnowski, 2004; Fleischhacker et al., 2011; Franco et al., 2008; Fraser et al., 2012; Giskes et al., 2011; Inagami et al., 2006; Lake and Townshend, 2006; Mehta and Chang, 2008; Michimi and Wimberly, 2010; Morland et al., 2006) and physical activity features (Gordon-Larsen et al., 2006; Kipke et al., 2007) with body composition and obesity at the individual level. Some studies control for social and economic characteristics as potential confounders (Meyer et al., 2015). We followed the socio-ecological literature and conceptualized and modeled social and economic community characteristics as key features of the risk landscape for physical inactivity and caloric over-consumption as well as risk-regulators that influence the likelihood of exposure to other obesogenic features of environments (Block et al., 2004; Greves Grow et al., 2010; Janssen et al., 2006; Larson et al., 2009; Nau et al., 2015).

Despite the recognition that obesogenic environments represent a diverse cluster of spatially co-occurring features, many studies presume that there are separate environments for food, social factors, and physical activity-related features; this assumption has not been tested. Furthermore, most studies have used regression analysis to assess the independent effect of each “exposure” in isolation. Individuals, however, experience their community

environment as a unified ensemble of features that may act jointly to affect health. A variable-by-variable approach risks what the sociologist, Gordon, called the partialling fallacy (Gordon, 1968). That is, the effect of the obesogenic context cannot be fully determined because it involves multiple variables that measure different dimensions of the same construct. Many studies have found small effects across a wide range of community risk factors when examined in isolation. None captures the totality of the impact of the obesogenic environment because that impact represents what Marini and Singer call a conjunctive plurality of causes that cannot be represented by the independent effects detected in linear additive regression models (Marini and Burton, 1988). Further, many studies adjust for related features of environments that are facets of the spatially co-occurring structures that shape obesogenic environments. This exacerbates the partialling fallacy and biases observed associations toward the null. While theorists posit obesogenic environments as a complex multidimensional construct, standard regression analysis does not permit us to identify the combination of interacting risk factors that render an environment obesogenic. Researchers have begun to measure multiple environmental risk factors using factor analysis and latent class analysis (Adams et al., 2011; Meyer et al., 2015; Wall et al., 2012).

We expand this new body of work by demonstrating an innovative approach that allows us to identify from a large set of theoretically plausible risk factors those community features that are most important for rendering a community obesogenic. Our method allows us to reorient the focus from understanding *if* a particular risk factor matters to identifying the set of risk factors that *matter most*. We implement a method called conditional random forests (CRF) (Strobl et al., 2008) to analyze and identify community characteristics that together can predict observed rates of obesity at an ecological level. CRF is a supervised machine learning algorithm that has been used in biomedical research to identify, for example, the set of proteins associated with the presence or absence of a particular cancer (Izmirlian, 2004) or the combination of genetic and dietary factors that jointly increase the risk for the development of Metabolic Syndrome (de Edelenyi et al., 2008). RF and CRF have also been applied in engineering (Kaur and Malhotra, 2008), geography (Pal, 2005) and ecology. To our knowledge, Basu and Siddiqi (2014) are the only authors to date who have used RF to identify risk environments by identifying features of geographic regions with high mortality.

We define obesogenic and obesoprotective environments for children as communities that fall into the highest or lowest quartile of the community level mean BMI z-score distribution. We use community level BMI-z because obesogenic and obesoprotective environments are ecological constructs that cannot be reduced to individual characteristics. Our approach is ecological, with the strengths and limits that this approach implies. We adopt an ecological approach because our goal is primarily about identification of constellations of community features. Obesogenicity and obesoprotectiveness are much like other community features, such as “walkability” or deprivation, properties of larger aggregate ecologies, not a function of the individuals who reside in them. In this study, we limit the scope of our inferences to relations between community constructs and average BMI among children in a community. We avoid ecological fallacy by restricting our inferences to the group level (Schwartz, 1994). Further, this stage of the analysis is for proof

of concept in a new method. We plan on a subsequent analysis using a multilevel analytic framework to address how CRF can inform variation in individual risk for obesity. CRF considers the coordinated effect of a large set of community features in order to classify communities as either type. The CRF algorithm ranks the classifying variables in terms of their contribution to the classification success. It identifies the combination of risk factors that matter most for differentiating obesogenic from obesoprotective communities.

The present study uses data from a large electronic health record of measured height and weight on children geocoded to a diverse set of 1,288 communities in 37 counties in Pennsylvania. We assembled a large dataset of community features from secondary data sources that have been linked to obesity in prior research. This set consists of 44 characteristics that cover multiple domains of community risk for obesity including social factors, food availability, and physical activity-related features including land use characteristics and physical activity establishments.

Using CRF we are able to examine the joint, spatially co-occurring pattern of features that may constitute obesogenic and obesoprotective environments. Results of these analyses allow us to: (1) identify the combination of features that are most important in rendering an environment obesogenic; (2) determine the relative importance of particular environmental features; (3) identify factors that do not improve classification accuracy; (4) characterize environmental features that may be targets for policy intervention.

METHODS

Data and measures

For the outcome of interest, we used electronic health records from the Geisinger Health System on measured height and weight of children ages 10 to 18 during 2010 (N=22,497). Data was drawn from a dataset that provides information from the Geisinger Health System from 2001–2012 for children ages 0–18. Geisinger is the largest healthcare provider in Pennsylvania, serving patients in 37 counties in central and northeastern Pennsylvania. The study population has been found to be representative of the general population in the region (Schwartz et al., 2014). All children ages 2–18 with valid height and weight whose home address could be geo-coded via ArcGIS with a longitude and latitude were included in the sample. Our study population includes the subset of children who saw their healthcare provider in 2010 (Schwartz et al., 2014). Standardized BMI-z scores were computed using CDC growth charts to allow comparability across age and sex strata after removing implausible values (CDC, 2014a). Prior research has shown that effects of the community environment are stronger for teenagers than younger children (Schwartz et al., 2011a). Therefore, this analysis was limited to children 10–18 years of age.

Children's geo-coded street addresses were attributed to one of 1,288 communities. Communities were operationalized as minor civil divisions (townships and boroughs) in rural areas or census tracts in urban areas (Schwartz et al., 2014). This mixed definition of community has been applied successfully in several other studies (Liu et al., 2013; Nau et al., 2015; Schwartz et al., 2014; Schwartz et al., 2011b). Townships and boroughs provide meaningful political community definitions in rural areas (Schwartz et al., 2014). Census

tracts are the most frequently used community proxy in urban areas (Riva et al., 2007). Communities are assumed to be the dominant socio-geographic context within which an adolescent's life unfolds. Applying a mixed definition of place provides greatest face validity for our community measure. We calculated average community level BMIs and classified communities as high- and low-obesity based on the top and bottom quartile in the distribution of average BMI-z in each community. To have stable estimates of mean BMI-z, we only considered communities with at least 50 children with measured BMI in 2010 (number of communities with 50 or more children N=197, number of communities in highest and lowest quartile used in the analysis N=99).

To capture features of the environment across multiple dimensions, we assembled multiple secondary data sources including information on food, social, physical activity, and land use features of the environment. Data on physical activity establishments (e.g. exercise facilities, gyms, parks, outdoor recreational facilities) as well as both food service (e.g., full service restaurants, fast food restaurants) and food retail (e.g., convenience stores, grocery stores, specialty food stores) for 2010 were obtained from two secondary data sources (InfoUSA and Dun & Bradstreet) and were classified using standard NAICS codes. The use of two sources of commercial establishment data reduces error and increases validity relative to ground-truthing compared to using any one database (Liese et al., 2010). Establishments were deduplicated, merged across databases, geocoded and aggregated to communities to form counts. Data on the social characteristics and land use patterns came from the 2005–2009 estimates of the American Community Survey (ACS) and The Pennsylvania Department of Transportation, respectively. Table 1 presents mean/proportion and standard deviation of each of the 44 indicators classified by domain.

Mathematical analysis

The purpose of this analysis was to identify the set of variables that, in combination, best classified high- and low-obesity communities. A secondary goal was to evaluate the relative importance of classification variables and to address whether particular variables differentially identify high- and low-obesity communities. To achieve this goal we used the machine learning method Conditional Random Forest (CRF). Machine learning is a family of methods that uses a recursive algorithm and a potentially large number of predictor variables to derive a mathematical model that best predicts the target outcome of interest. In contrast to traditional parametric statistical methods where the analyst pre-defines the structure of the model and tests one or multiple hypotheses on independent or interaction effects of selected variables, machine learning uses an algorithm to search (or learn) the combination of variables that best classifies the outcome (supervisor) by improving a decision criteria in an iterative process (Breiman, 2001b). When using machine learning algorithms, the analyst is tasked with carefully selecting candidate predictors and the supervisor based on prior findings and plausible theoretical assumptions. In contrast to traditional statistical methods, the analyst is, however, not required to make a priori assumptions about the causal structure of relationships or distributional qualities of predictors and outcomes because most machine learning approaches are non-parametric (Breiman et al., 1984)

CRF is an extension of RF and improves classification success in the case of highly correlated predictors, such as community characteristics. The CRF algorithm creates increasingly homogeneous groups in terms of the outcome measure by using predictor variables to split cases. The resulting structure of the splitting process takes the shape of a data tree, as depicted in Figure 1. Figure 1 presents one of many thousands of trees that CRF grows. Please note that results from a single tree do not have a meaningful interpretation and only serve here to illustrate the iterative splitting process of the CRF algorithm. The CRF algorithm begins to grow each tree by randomly choosing about two thirds of the data as a training dataset. The remaining one third of the data, or the out-of-bag (OOB sample), is set aside for validation purposes once the tree is grown. After choosing a training dataset, CRF selects a small random sample of variables from all community characteristics. The size of this random sample is pre-defined and user specified. From this subset of variables, the algorithm uses Chi-Square tests to test the hypothesis whether or not a variable is statistically significantly associated with the outcome variable and selects the variable with the strongest statistical association with the outcome (Strobl et al., 2008; Strobl et al., 2007). In Figure 1 the variable with the strongest association among the first randomly chosen subset of variables is population density (pop_den00). Next, CRF uses an entropy reduction measure to identify a split value of population density that best differentiates high- and low-obesity communities. In our tree, obesogenic and obesoprotective communities are best differentiated by whether their population density is more or less than 2,555 per square mile. Communities are then split into two groups based on that value. Each of the two resulting nodes is further split according to the same procedure, each time choosing a new sub-sample of variables, until none of the randomly chosen variables has a statistically significant association with the outcome.

Figure 1 illustrates that the recursive partitioning allows CRF to account for complex interactions of variables that go far beyond the two- and three-way interactions routinely used in regression analysis. Once a tree is grown, it is used to classify its respective out-of-bag sample – this is the validation step. Figure 1 shows the tree's prediction success for each node. Nodes vary in their prediction rate from 100 to 60% (50% would be expected if only chance would be used to predict the outcome). From these predictions, error rates are calculated for each tree (Breiman, 2001a). CRF builds several thousand trees. OOB overall and class-specific error rates are calculated by averaging the prediction errors across all trees (Strobl et al., 2009). In addition to providing classification errors, CRF ranks each variable based on its relative importance in the classification process. The variable importance score of a variable is calculated by re-computing the overall error rate after the values of that variable have been randomly permuted (Liaw and Wiener, 2002). The more the classification error increases after permutation of the values of a variable, the higher the variable's importance score. Permutation therefore mimics the absence of that variable. CRF provides conditional variable importance scores, which are one of the features that render CRF robust with respect to correlated classifying variables (Strobl et al., 2008). Conditional variable importance scores are computed by stratifying, for each tree, the sample according to the split points of the variables used to construct that tree (Strobl et al., 2008). The permutation of values is then executed within the strata produced by this grid; error rates are averaged across all trees. This process assures that the importance score of a particular

variable is conditioned on the influence of all the other variables that contribute to the classification accuracy. CRF can handle any number of predictors and is particularly apt to handle small n large p problems (Matthews, 2011). Predictors may be continuous, categorical or nominal. Similar to traditional regression methods, there is ongoing research on the best number of predictors for CRF and RF (James et al., 2013) that explores the tradeoffs of potential under- (high variance) and over-fitting (high bias).

The CRF constructed 5,000 trees for each analysis. We follow Strobl (2009) and Breiman (2001a) who used simulation experiments to empirically derive a recommendation for the number of variables to be evaluated at each split. They arrive at a number close to the square root of the number of classifying variables. For analyses involving the full set of 44 variables however, we used 9 variables for evaluation at each split based on a sensitivity analysis that indicated that 9 variables decreased the classification error maximally. For all other analysis we follow Breiman's and Strobl's recommendations. To assure stable variable importance rankings and OOB error rates, we ran a Monte Carlo ensemble of 50 random forests with varying random seeds for each analysis. We then created a variable importance list based on the mean rank of each variable across the 50 runs. OOB error rates were averaged across all 50 forests.

According to Strobl (2008) a variable can be said to contribute to the classification if its importance score is larger than the absolute value of the smallest negative score of all variable importance scores. Importance scores of variables that do not contribute to the classification are expected to vary randomly around zero; therefore, Strobl's rule is a conservative measure for assessing which variables are most important. We identified the set of variables that contributed most to the classification by identifying those factors that consistently scored above Strobl's threshold.

To assess whether a particular domain of characteristics—social, food, physical activity-related features—played a larger role in predicting high- and low-obesity communities, we ran CRF analyses separately for each of these domains. Social conditions have been found to be upstream risk regulators of other environmental risk factors related to food consumption and physical activity (Glass and McAtee, 2006). To assess whether these downstream characteristics operated to translate the effect of social characteristics, we also ran an analysis that used food and physical activity-related features as classifiers but excluded social characteristics. We assessed the direction of the association of selected variables with the outcome via Spearman correlations. All analyses were conducted in R using the *party* package (Strobl et al., 2009).

RESULTS

The analysis was based on prediction of 50 high-obesity and 49 low-obesity communities (communities in the upper and lower quartile of the BMI-z distribution of communities with at least 50 children). The average community BMI-z across all communities was 0.69 (standard deviation (SD)=0.18); the highest quartile had an average BMI z-score of 0.92 (SD=0.1) while the lowest quartile had an average BMI z-score of 0.47 (SD=0.09). Table 2

shows descriptive characteristics of the four quartiles of the community BMI z-score distribution.

Table 3 presents the model specifications and overall and class-specific OOB error rates for five different analyses. For the analysis of the full set of 44 variables (Model 1), the overall error rate for the Markov ensemble was 0.33. In other words, 33% of communities were misclassified; in consequence, CRF classified 67% of communities correctly. The class-specific error rates showed that the full set of variables (Model 1) classified high- and low-obesity communities equally well (33% and 32% were misclassified, respectively). Figure 2 presents the average conditional variable importance rankings across 50 runs ordered by decreasing importance; the bar length indicates the average rank (shorter bars indicate higher importance ranking). The first number following each bar is the average conditional variable importance rank, the number in parentheses indicates the number of runs (out of 50) that each variable contributed to the classification according to Strobl's rule.

Thirteen variables contributed consistently (49 or 50 runs out of 50) to the classification. Beyond these 13 variables, the number of times a variable contributed to the classification dropped sharply (from 50 down to 36 and 33 runs). Several variables—such as the count of full-service restaurants or grocery stores or the number of physical activity establishments—did not contribute to the classification in any run. The 13 most important variables included a mix of social, food, land use, and physical activity features. Social environment characteristics (unemployment, social disorganization, percent population with less than a high school degree, not owning a car and on public assistance) and land use characteristics (population density, population change, place type, vehicle miles travelled) appeared to be most important in classifying the obesogenicity of places. In addition, the count of physical activity establishments and two food variables (count of snack food and chain fast food outlets) were among the 13 most important variables.

Table 3 also shows classification error rates for the analysis of variables of the food and physical activity related variables without social variables (Model 2) and for analyses of each of the domains separately (Models 3–5). Social features (Model 3) classified communities approximately as well as the full set of variables (Model 1: OOB=31% compared to 33% for full set). Social characteristics were, however, less successful in classifying low-obesity communities (OOB=34% and 27% for low- and high-obesity communities, respectively). The CRF model that used only food and physical activity variables (Model 2) showed that these two domains jointly performed about as well as Model 1 that included social characteristics (OOB=31% and 31%, respectively).

Table 4 lists all variables that contributed at least once to the classification of variables when social features were excluded (Model 2). The third and fourth column list average rankings in the conditional variable importance and the number of significant runs of each variable from the analysis of the full set of variables for comparison. Overall, the relative ranking of variables is comparable across the full Model 1 and Model 2 without social features. Once social features are removed, variables that were contributing less consistently in Model 1 become more consistent contributors in Model 2. The models with only food or the physical activity features (Model 4 and Model 5) yielded higher classification errors (each 35%).

Food variables better classified high-obesity communities (average OOB error=29%) than low-obesity communities (OOB error=41%). In contrast, physical activity-related features were stronger classifiers for low-obesity communities (OOB=30%) than high-obesity communities (OOB=40%).

A pre-planned sensitivity analysis was conducted to determine whether results would differ using the proportion of children above the 85th percentile on BMI-z as the supervisor/ outcome, rather than average BMI-z. The results proved to be consistent and similar.

DISCUSSION

This study combined a theory-driven approach to candidate variable selection with a data-driven analysis strategy. We believe this is the first study to use Conditional Random Forests (CRF), a machine learning technique, to consider the risk landscape of obesogenic environments as a diverse set of spatially co-occurring risk factors. Our results suggest that environments represent a unified risk topography that is not easily divided along dimensional lines. We identified 13 variables that contributed consistently to the classification success; these 13 variables included social, food and physical activity-related features. Among the physical activity-related features, land use characteristics stood out in particular. While high- and low-obesity communities were best described by a diverse set of risk factors, our results point to the particular importance of social characteristics as key features of high- and low-obesity communities. Six of the thirteen variables that ranked highest in conditional importance were social characteristics. Excluding social characteristics from the analysis, however, yielded comparable overall classification accuracy. The conditional variable importance ranking without social features was comparable to that of the full set of variables with variables further down the list contributing more consistently after social features were excluded, suggesting that food and physical activity features may be more proximate risk factors through which social characteristics operate. This also casts doubt on the frequent conceptualization of the social domain as distinct and separable from the food or physical activity domains.

One limitation of the CRF analysis is that it does not indicate clearly the direction or intensity of association between a predictor and the supervisor. The recursive splitting algorithm provides relief from parametric assumptions, but it also makes interpretation of direction of associations more challenging. Correlation coefficients only assess the independent association of a single variable with the outcome, and do not reflect the pattern of recursive variable splits that informed the CRF. They can, however, provide a sense of the direction of the association. For heuristic purposes we provide Table 5 with Spearman's correlations between the top 13 variables and the classification supervisor.

Social features present moderately strong positive associations with high-obesity communities. In prior research we found that living in well-off communities is obesoprotective. Children born into these communities have BMIs that are on average 1.5 points lower by age 18 than their counterparts living in deprived communities (Nau et al., 2015). Population change and vehicle miles travelled are negatively correlated with high-obesity communities. These trends likely reflect the suburbanization of Pennsylvania

townships with better-off families moving into more sparsely settled communities that are also more likely to be traversed by major traffic arteries and highways (authors' data queries). Negative correlations of the counts of snack stores and fast food chains with high-obesity communities add to the evidence that obesoprotective communities are those that are economically more viable. This short review of bivariate associations demonstrates the importance of considering community features as an ensemble. Finding a negative effect of fast food chains on BMI in a regression model by itself would have been difficult to interpret without information on the importance of other community features.

Domain-specific analyses further suggested that high- and low-obesity communities were best characterized by different sets of risk factors. Classification accuracy by domain varied for high- and low-obesity communities. Greater classification accuracy of one class compared to another points to differential heterogeneity of each community type in terms of the classifiers. The high-obesity communities were classified with greater accuracy by social and food features; low-obesity communities were classified with greater accuracy by physical activity features.

Our results support and add to the literature on ecological studies of environments and obesity rates (Drewnowski et al., 2007; Holtgrave and Crosby, 2006; Reidpath et al., 2002; Vandegrift and Yoked, 2004). A number of studies have shown that physical activity features are either negatively associated with obesity or are unrelated to average bodyweight (Ewing et al., 2003; Lopez-Zetina et al., 2006; Vandegrift and Yoked, 2004). Similarly, the presence of certain types of food establishments have shown weak or no associations with community level BMI (Maddock, 2004; Simmons et al., 2005). Our method and results add to studies that have aimed at modeling patterns of neighborhood features via regression, factor, and latent class analysis. One study used a regression analysis to predict adolescents' moderate to vigorous physical activity with 50 community predictors and found that most predictors were not statistically significantly associated (Graham et al., 2014). Only personal and social factors showed significant associations. While results confirm in part our findings, a regression study of this kind is hampered by highly correlated predictors, inability to assess complex interactions, and risks of violating several regression assumptions with its large set of variables. Another study that used theoretically constructed indices and latent class analysis found that certain types of neighborhood conditions such as high walkability and activity supportive environments were associated with physical activity in an 11-country study (Adams et al., 2013). A comprehensive study by Meyer et al. (2015) used individual level observations and latent class analysis to identify obesogenic neighborhood profiles in high- and low-density residential areas based on the combination of neighborhood physical and food resources. Meyer et al identified two different latent classes that were predictive of diet quality in high- and low-population areas, respectively. Tseng et al. (2014) created a conceptually based aggregate measure of obesogenicity and found a positive association of their indices with BMI in urban low income communities in Australia but inverse associations in rural communities. These studies have pioneered a comprehensive approach to measuring and modeling community effects on obesity and obesity-related behaviors and outcomes. Our work contributes to this body of research by providing a novel, non-parametric approach to identifying a best fitting model by directly using the outcome of interest as supervisor.

Our study showed that machine learning algorithms, when used with variables that have been linked in prior studies to obesity, can be used to generate new insights into the nature of obesogenic and obesoprotective communities. Regression analysis assumes that risk factors affect BMI linearly and independently (Breiman, 2001b; Fox, 1997). While regression models can accommodate a host of interactions and non-linearities, CRF and other non-parametric machine learning methods are more powerful and flexible. CRF can jointly describe the patterning of community characteristics without requiring parametric assumptions of linearity, additivity, normality, and independence among features, and can accommodate a large number of community risk factors that are often highly collinear and non-normally distributed with strong skew. Unlike regression analysis, CRF does not assume that the data are generated from a specific causal model. Instead, the recursive partitioning algorithm of CRF allowed us to consider an ensemble of 44 environmental risk factors without strong assumptions about their functional form. Breiman (2001a) also found that minimizing classification error in the out-of-bag sample provided better predictive accuracy than conventional model fitting in regression analysis, thus reducing the risk of potentially misleading conclusions from ill-fitting models.

CRF is similar to other methods more commonly used in public health research such as single classification and regression trees (CART) (Goel et al., 2009; Kitsantas and Gaffney, 2010; Lemon et al., 2003; Marshall, 2001). However, CRF models have been shown to yield higher classification success and significantly more stable results than CART. Results of CART tend to be more unstable across model iterations because split-point selection can be influenced by variability induced by the random selection of cases from the population (Reidpath et al., 2002). CRF yields stable results by growing numerous trees on different sub-samples of the data and averaging across these trees. Strobl et al. (2002) and Breiman (2001a) agree that growing a large number of trees does not, however, lead to over-fitting. Breiman suggests that averaging across a large number of trees that were constructed using random draws of variables and bootstrap samples of the data increases precision of estimates when the number of trees is increased because of the law of large numbers. CRF also has advantages over other data-reduction techniques such as exploratory factor analysis (EFA). The procedures and the results are, however, fundamentally different. Instead of using the correlation structure of the data to discover common factors, CRF uses a criterion variable as a prediction target to supervise the classification algorithm. Thus, variable selection is directly linked to the outcome. CRF also provides some practical advantage over EFA in that it can handle small sample sizes and a very large number of variables, is insensitive to the measurement scale of variables and, by virtue of being non-parametric, does not require assumptions about the number of factors, their correlations or multivariate normality of indicators (Breiman, 2001a).

It is important to note that variable importance rankings only assess the relative importance of a variable in prediction. They do not provide the direction and absolute strength of associations (such as a measure of the variance explained or a predictor significance level); however, as we showed above, the direction can be easily ascertained through the sign of the Spearman's correlation coefficient between predictor and response. Like any analysis, CRF depends on the reliability and validity of both predictors and the supervisor variable. The importance ranking of a variable can be influenced not only by error in the measurement of

that particular variable but also by error in the measurement of all other variables in the analysis. We are confident that systematic measurement error is not related to error in measurement of the outcome since the data sources are entirely separate (electronic health records vs. ACS and other secondary data sources). Similarly, conclusions about the importance of particular domains assume that these domains are adequately represented through the variables included in the analysis. Further, we assume that the upper and lower quartile of the distribution of community-level average BMI-z is a valid surrogate measure of obesogenic and obesoprotective environments. Sensitivity analysis with the distribution of proportion of children above the 85th percentile provided similar results. Other measures such as quintiles or deciles on data pooled over several years might provide better supervisors and will be tested. Also, because we chose communities with 50 or more children, these results may not generalize to smaller, sparsely populated communities. Furthermore, our results cannot confirm that identified environmental features are causally related to risk of obesity. Like all observational study designs, we are not able to rule out the possibility that associations between community characteristics and obesity rates are an artifact of obese children (and their parents) moving to certain environments. In a future analysis, we will explore a longitudinal application to explore this question further. This study is ecological in design and interpretation. We limit our inferences about patterns of association between environmental characteristics and rates of obesity in the aggregate. We do not infer that these associations imply that individuals living in obesogenic communities are at higher risk at the individual level.

We suggest that machine learning methods can make an important contribution to policy discussions and suggest new directions for further research. To implement efficient policies we need to understand the entire set of risk factors that render an environment obesogenic or obesoprotective. We identified a cluster of 13 variables that, in combination, were able to categorize communities as high- or low-obesity. Our results suggest that obesity-prevention efforts should consider interventions aimed at a coordinated set of environmental features. Several risk factors that have garnered attention in prior studies, such as grocery stores and parks, do not rank among the most important variables if considered as a system of risk factor (Gordon-Larsen et al., 2006; Inagami et al., 2006). Policy efforts that target individual components of the environment may be inefficient. A better understanding of how overlapping features of environments synergistically interact may provide further directions for research into the best levers for comprehensive cost-effective interventions. This analysis points, for example, to the importance of social features and economic viability of communities for determining community level obesogenicity. Interventions such as land use and tax policies might provide broader and more sustainable change that affects community-level BMIs. In a next step, we plan to use further machine learning methods to build “diagnostic tools” to identify obesogenic and obesoprotective communities to direct policy and intervention efforts towards communities and combination of community factors that “matter most.” Our research is the first step to adapting this promising suite of methods for creating a public health research and practice that accounts for the complexities of the environment.

CRF offers an innovative and flexible modeling tool for operationalizing risk environments in an ecological manner. The method is easily accessible through the open-source R-package

with abundant clearly written technical documentation. CRF provides a strategy for analyses that acknowledges complexity instead of reducing it by imposing assumptions that may be unreasonable. Current research has been hampered by a one-variable-at-a-time approach that has yielded inconsistent results. Obesogenic and obesoprotective environments are structured risk landscapes involving multiple factors that combine synergistically and that are difficult to tease apart. Individually, correlations between single variables and the outcome may be small to modest. We identified a set of factors that combined to create a fabric of risk that may shape BMI in children. While prior research has identified a wealth of community-level risk factors that matter for obesity, CRF is an analysis technique that allows approaching the next important questions: “What combination of risk factors matters most?”

Acknowledgments

The project described was supported by Grant Number U54HD070725 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD). The project is co-funded by the NICHD and the Office of Behavioral and Social Sciences Research (OBSSR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NICHD or OBSSR. Dr. Nau was supported by the training core of the Johns Hopkins Global Obesity Prevention Center.

References

- Adams MA, Ding D, Sallis JF, Bowles HR, Ainsworth BE, Bergman P, Bull FC, Carr H, Craig CL, De Bourdeaudhuij I. Patterns of neighborhood environment attributes related to physical activity across 11 countries: a latent class analysis. *Int J Behav Nutr Phys Act*. 2013; 10:10.1186. [PubMed: 23351329]
- Adams MA, Sallis JF, Kerr J, Conway TL, Saelens BE, Frank LD, Norman GJ, Cain KL. Neighborhood environment profiles related to physical activity and weight status: a latent profile analysis. *Preventative Medicine*. 2011; 52:326–331.
- Basu S, Siddiqi A. Geographic disparities in US mortality: “hot-spotting” large databases. *Epidemiology*. 2014; 25:468–470. [PubMed: 24713886]
- Block JP, Scribner RA, DeSalvo KB. Fast food, race/ethnicity, and income: a geographic analysis. *American journal of preventive medicine*. 2004; 27:211–217. [PubMed: 15450633]
- Breiman L. Random forests. *Machine learning*. 2001a; 45:5–32.
- Breiman L. Statistical modeling: The two cultures. *Statistical Science*. 2001b; 16:199–231.
- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software; Monterey, CA: 1984.
- Casey AA, Elliott M, Glanz K, Haire-Joshu D, Lovegreen SL, Saelens BE, Sallis JF, Brownson RC. Impact of the food environment and physical activity environment on behaviors and weight status in rural U.S. communities. *Preventative Medicine*. 2008; 47:600–604.
- CDC. Control, C.f.D. Cut-offs to define outliers in the 2000 CDC Growth Charts. 2014a.
- CDC. Prevention, C.f.D.C.a. A SAS Program for the CDC Growth Charts. Atlanta, GA: 2014b.
- Cummins S, Macintyre S. Food environments and obesity--neighbourhood or nation? *International Journal of Epidemiology*. 2006; 35:100–104. [PubMed: 16338945]
- de Edelenyi FS, Goumidi L, Bertrais S, Phillips C, MacManus R, Roche H, Planells R, Lairon D. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest. *Genes & nutrition*. 2008; 3:173–176. [PubMed: 19034549]
- Drewnowski A. Obesity and the food environment: dietary energy density and diet costs. *American journal of preventive medicine*. 2004; 27:154–162. [PubMed: 15450626]
- Drewnowski A, Rehm DC, Solet D. Disparities in obesity rates: analysis by ZIP code area. *Social science & medicine*. 2007; 65:2458–2463. [PubMed: 17761378]

- Ewing R, Schmid T, Killingsworth R, Zlot A, Raudenbush S. Relationship between urban sprawl and physical activity, obesity, and morbidity. *American Journal of Health Promotion*. 2003; 18:47–57. [PubMed: 13677962]
- Fleischhacker S, Evenson K, Rodriguez D, Ammerman A. A systematic review of fast food access studies. *Obesity reviews*. 2011; 12:e460–e471. [PubMed: 20149118]
- Fox, J. *Applied regression analysis, linear models, and related methods*. Sage Publications, Inc; 1997.
- Franco M, Diez Roux AV, Glass TA, Caballero B, Brancati FL. Neighborhood characteristics and availability of healthy foods in Baltimore. *American journal of preventive medicine*. 2008; 35:561–567. [PubMed: 18842389]
- Frank LD, Saelens BE, Powell KE, Chapman JE. Stepping towards causation: do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity? *Social science & medicine* (1982). 2007; 65:1898–1914. [PubMed: 17644231]
- Franzini L, Elliott MN, Cuccaro P, Schuster M, Gilliland MJ, Grunbaum JA, Franklin F, Tortolero SR. Influences of physical and social neighborhood environments on children's physical activity and obesity. *American journal of public health*. 2009; 99:271–278. [PubMed: 19059864]
- Fraser LK, Clarke GP, Cade JE, Edwards KL. Fast food and obesity: a spatial analysis in a large United Kingdom population of children aged 13–15. *American journal of preventive medicine*. 2012; 42:e77–85. [PubMed: 22516506]
- Giskes K, van Lenthe F, Avendano-Pabon M, Brug J. A systematic review of environmental factors and obesogenic dietary intakes among adults: are we getting closer to understanding obesogenic environments? *Obesity Reviews: an official journal of the International Association for the Study of Obesity*. 2011; 12:e95–e106. [PubMed: 20604870]
- Glass TA, McAtee MJ. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. *Social science & medicine*. 2006; 62:1650–1671. [PubMed: 16198467]
- Goel R, Misra A, Kondal D, Pandey RM, Vikram NK, Wasir JS, Dhingra V, Luthra K. Identification of insulin resistance in Asian Indian adolescents: classification and regression tree (CART) and logistic regression based classification rules. *Clinical endocrinology*. 2009; 70:717–724. [PubMed: 18778399]
- Gordon-Larsen P, Nelson MC, Page P, Popkin BM. Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics*. 2006; 117:417–424. [PubMed: 16452361]
- Gordon RA. Issues in multiple regression. *American Journal of Sociology*. 1968; 73:592–616.
- Graham DJ, Wall MM, Larson N, Neumark-Sztainer D. Multicontextual correlates of adolescent leisure-time physical activity. *American journal of preventive medicine*. 2014; 46:605–616. [PubMed: 24842737]
- Greves Grow HM, Cook AJ, Arterburn DE, Saelens BE, Drewnowski A, Lozano P. Child obesity associated with social disadvantage of children's neighborhoods. *Social science & medicine*. 2010; 71:584–591. [PubMed: 20541306]
- Hill JO, Peters JC. Environmental contributions to the obesity epidemic. *Science*. 1998; 280:1371–1374. [PubMed: 9603719]
- Holtgrave DR, Crosby R. Is social capital a protective factor against obesity and diabetes? Findings from an exploratory study. *Annals of epidemiology*. 2006; 16:406–408. [PubMed: 16246583]
- Inagami S, Cohen DA, Finch BK, Asch SM. You are where you shop: grocery store locations, weight, and neighborhoods. *American journal of preventive medicine*. 2006; 31:10–17. [PubMed: 16777537]
- Izmirlian G. Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the Setting of a Cancer Prevention Trial. *Annals of the New York Academy of Sciences*. 2004; 1020:154–174. [PubMed: 15208191]
- James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*. Springer; 2013.
- Janssen I, Boyce WF, Simpson K, Pickett W. Influence of individual-and area-level measures of socioeconomic status on obesity, unhealthy eating, and physical inactivity in Canadian adolescents. *The American journal of clinical nutrition*. 2006; 83:139–145. [PubMed: 16400062]

- Kaur, A.; Malhotra, R. Application of Random Forest in Predicting Fault-Prone Classes, *Advanced Computer Theory and Engineering*, 2008. ICACTE'08. International Conference on. IEEE; 2008. p. 37-43.
- Kipke MD, Iverson E, Moore D, Booker C, Ruelas V, Peters AL, Kaufman F. Food and park environments: neighborhood-level risks for childhood obesity in east Los Angeles. *Journal of Adolescent Health*. 2007; 40:325–333. [PubMed: 17367725]
- Kitsantas P, Gaffney KF. Risk profiles for overweight/obesity among preschoolers. *Early human development*. 2010; 86:563–568. [PubMed: 20716472]
- Lake A, Townshend T. Obesogenic environments: exploring the built and food environments. *Journal of the Royal Society of Health*. 2006; 126:262–267.
- Larson NI, Story MT, Nelson MC. Neighborhood environments: disparities in access to healthy foods in the US. *American journal of preventive medicine*. 2009; 36:74–81. e10. [PubMed: 18977112]
- Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*. 2003; 26:172–181. [PubMed: 14644693]
- Liaw A, Wiener M. Classification and Regression by randomForest. *R news*. 2002; 2:18–22.
- Liese AD, Colabianchi N, Lamichhane AP, Barnes TL, Hibbert JD, Porter DE, Nichols MD, Lawson AB. Validation of 3 food outlet databases: completeness and geospatial accuracy in rural and urban food environments. *American Journal of Epidemiology*. 2010; 172:1324–1333. [PubMed: 20961970]
- Liu AY, Curriero FC, Glass TA, Stewart WF, Schwartz BS. The contextual influence of coal abandoned mine lands in communities and type 2 diabetes in Pennsylvania. *Health & Place*. 2013; 22:115–122. [PubMed: 23689181]
- Lopez-Zetina J, Lee H, Friis R. The link between obesity and the built environment. Evidence from an ecological analysis of obesity and vehicle miles of travel in California. *Health & Place*. 2006; 12:656–664. [PubMed: 16253540]
- Maddock J. The relationship between obesity and the prevalence of fast food restaurants: state-level analysis. *American Journal of Health Promotion*. 2004; 19:137–143. [PubMed: 15559714]
- Marini, MM.; Burton, S. Causality in the Social Sciences. In: Clogg, C., editor. *Sociological Methodology*. American Sociological Association; Washington, DC: 1988. p. 347-409.
- Marshall RJ. The use of classification and regression trees in clinical epidemiology. *Journal of clinical epidemiology*. 2001; 54:603–609. [PubMed: 11377121]
- Matthews, SA. *Communities, Neighborhoods, and Health*. Springer; 2011. Spatial polygamy and the heterogeneity of place: studying people and place via egocentric methods; p. 35-55.
- Mehta NK, Chang VW. Weight status and restaurant availability a multilevel analysis. *American journal of preventive medicine*. 2008; 34:127–133. [PubMed: 18201642]
- Meyer KA, Boone-Heinonen J, Duffey KJ, Rodriguez DA, Kiefe CI, Lewis CE, Gordon-Larsen P. Combined measure of neighborhood food and physical activity environments and weight-related outcomes: The CARDIA study. *Health and Place*. 2015; 33:9–18.
- Michimi A, Wimberly MC. Associations of supermarket accessibility with obesity and fruit and vegetable consumption in the conterminous United States. *Int J Health Geogr*. 2010; 9:49. [PubMed: 20932312]
- Morland K, Diez Roux AV, Wing S. Supermarkets, other food stores, and obesity: the atherosclerosis risk in communities study. *American journal of preventive medicine*. 2006; 30:333–339. [PubMed: 16530621]
- Nau C, Schwartz BS, Bandeen-Roche K, Liu A, Pollak J, Hirsch A, Bailey-Davis L, Glass TA. Community socioeconomic deprivation and obesity trajectories in children using electronic health records. *Obesity*. 2015; 23:207–212. [PubMed: 25324223]
- Pal M. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*. 2005; 26:217–222.
- Poston WS 2nd, Foreyt JP. Obesity is an environmental issue. *Atherosclerosis*. 1999; 146:201–209. [PubMed: 10532676]

- Reidpath DD, Burns C, Garrard J, Mahoney M, Townsend M. An ecological study of the relationship between social and environmental determinants of obesity. *Health & Place*. 2002; 8:141–145. [PubMed: 11943585]
- Riva M, Gauvin L, Barnett TA. Toward the next generation of research into small area effects on health: a synthesis of multilevel investigations published since July 1998. *Journal of epidemiology and community health*. 2007; 61:853–861. [PubMed: 17873220]
- Rundle A, Neckerman KM, Freeman L, Lovasi GS, Purciel M, Quinn J, Richards C, Sircar N, Weiss C. Neighborhood food environment and walkability predict obesity in New York City. *Environmental Health Perspectives*. 2009; 117:442–447. [PubMed: 19337520]
- Schwartz B, Stewart W, Pollak J, Mercer D, DeWalle J, Glass T. PS2-35: Associations of Body Mass Index with Measures of the Built and Social Environments in Children and Adolescents: The Environmental Health Institute Child and Adolescent Obesity Study. *Clinical Medicine & Research*. 2011a; 9:160–160.
- Schwartz BS, Bailey-Davis L, Bandeen-Roche K, Pollak J, Hirsch AG, Nau C, Liu AY, Glass TA. Attention Deficit Disorder, Stimulant Use, and Childhood Body Mass Index Trajectory. *Pediatrics*. 2014; 133:668–676. [PubMed: 24639278]
- Schwartz BS, Stewart WF, Godby S, Pollak J, DeWalle J, Larson S, Mercer DG, Glass TA. Body mass index and the built and social environments in children and adolescents using electronic health records. *American journal of preventive medicine*. 2011b; 41:e17–e28. [PubMed: 21961475]
- Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *American Journal of Public Health*. 1994; 84:819–824. [PubMed: 8179055]
- Simmons D, McKenzie A, Eaton S, Cox N, Khan MA, Shaw J, Zimmet P. Choice and availability of takeaway and restaurant food is not related to the prevalence of adult obesity in rural communities in Australia. *International journal of obesity*. 2005; 29:703–710. [PubMed: 15809667]
- Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC bioinformatics*. 2008; 9:307. [PubMed: 18620558]
- Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*. 2007; 8:25. [PubMed: 17254353]
- Strobl C, Hothorn T, Zeileis A. Party on! A new, conditional variable-importance measure for random forests available in the Party package. *The R Journal*. 2009; 1:14–17.
- Swinburn B, Egger G, Raza F. Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Preventative Medicine*. 1999; 29:563–570.
- Vandegrift D, Yoked T. Obesity rates, income, and suburban sprawl: an analysis of US states. *Health & Place*. 2004; 10:221–229. [PubMed: 15177197]
- Wall MM, Larson NI, Forsyth A, Van Riper DC, Graham DJ, Story MT, Neumark-Sztainer D. Patterns of obesogenic neighborhood features and adolescent weight: a comparison of statistical approaches. *American Journal of Preventive Medicine*. 2012; 42:e65–75. [PubMed: 22516505]

HIGHLIGHTS

- obesogenic environments (OGE) consist of a diverse cluster of spatially co-occurring risk factors
- we used a machine learning algorithm to identify the risk factors that matter most for childhood obesity
- OGEs are best described by a diverse set of social, physical activity, and food features
- Social features are of particular importance
- Obesogenic and obesoprotective environments are best defined by different sets of risk factors

Figure 1. Illustrative CRF analysis tree for all 44 variables. Note that results from a single tree are not used directly; CRF uses averages over many trees.

Figure 2.

Average ranks of conditional permutation variable importance with number of runs above Strobl's threshold for significant contribution in paranthesis, Monte Carlo ensemble of 50 CRF on full set of variables

Bar length indicates ranking of conditional variable importance averaged over all 50 runs (shorter bars equal higher ranking). Number to the right of each bar indicates the average variable importance ranking across 44 variables and, in brackets the number of runs (out of 50) in which that variable was found to contribute significantly to classification accuracy according to Strobl's threshold.

Health Place. Author manuscript; available in PMC 2016 September 19.

Social features ^d		Food establishments ^b		Physical activity-related features		
	OPE ^c	OGE ^d		OPE	OGE	
Prop of pop un-employed	0.04 (0.02)	0.05 (0.02)	Cnt convenience stores	2.36 (2.99)	2.31 (1.94)	5658.60 (6473.19)
Prop of pop with less than high school	0.10 (0.05)	0.15 (0.05)	Cnt fast food chain restaurants	1.06 (1.84)	0.37 (0.70)	1278.91 (1956.60)
Prop of pop in poverty	0.12 (0.10)	0.17 (0.07)	Cnt snack stores	1.98 (2.543)	1.12 (1.82)	504.18 (997.56)
Prop of pop without car	0.08 (0.02)	0.13 (0.08)	Cnt limited service restaurants	3.14 (5.30)	1.67 (2.35)	Community Type (proportion)
						Boroughs
						Tracts
						Townships
Prop of pop on public assistance	0.02 (0.02)	0.04 (0.03)	Diversity food estab.	5.86 (3.24)	5.96 (2.51)	Street connectivity ^h
Prop of pop not in labor force	0.38 (0.07)	0.39 (0.07)	Cnt any food establishment	21.48 (27.65)	16.84 (14.61)	Vehicle miles travelled ⁱ
Social disorg score ^e	0.30 (2.67)	2.72 (2.98)	Cnt pharmacies/drug stores	1.52 (2.06)	1.84 (2.26)	<u>Physical activity establishments</u>
Physical disorder score ^f	3.68 (2.46)	4.67 (2.77)	Cnt bars/hotels with foods	1.42 (2.49)	1.67 (1.86)	Cnt any physical activity estab
			Cnt full service restaurants	7.06 (10.24)	5.35 (4.41)	Cnt outdoor fitness clubs
			Cnt retailers	10.00 (11.35)	8.14 (7.60)	Cnt outdoor recreation clubs
			Cnt specialty food	0.34 (0.66)	0.37 (0.60)	Cnt outdoor attractions
<i>Food estab.</i>			<i>Physical activity estab.</i>			
				OPE	OGE	OPE
			Cnt other food services	0.22 (0.79)	0.06 (0.24)	Diversity phy act estab
			Cnt discount stores	0.02 (0.14)	0.00 (0.00)	Cnt outdoor public activity spaces
			Cnt diet/weight reduction businesses	0.04 (0.20)	(0.00) (0.00)	Cnt indoor fitness club
			Cnt fruit/vegetable markets	0.14 (0.41)	0.08 (0.28)	Cnt entertainment/leisure estab

Social features ^a	Food establishments ^b		Physical activity-related features				
	OPE ^c	OGE ^d	OPE	OGE	<i>Land use features^g</i>	OPE	OGE
					Cnt indoor phys act estab	2.28 (2.84)	1.75 (2.59)
					Count outdoor fitness clubs	2.08 (2.72)	1.41 (1.98)
						</	

^a Social characteristics of population (proportion of population, prop of pop) come from the ACS 2005–2009.

^b Counts (Cnt) of food and physical activity establishments per community derive from a merged database of InfoUSA and Dun & Bradstreet.

^c OPE: Obesoprotective environment; communities in the lowest quartile of the average BMI-z distribution

^d OGE: Obesogenic environments; communities in the highest quartile of the average BMI-z distribution

^e The social disorganization index is a summary measure constructed by summing 4 census variables including (1) percent of female headed households, (2) the percent of heads of household who lived in a different county 5 years ago, (3) the percent of the adult population divorced or separated, and (4) 100 – percent of households that are owner occupied. Each variable was z-score transformed to reflect its rank position across all 1288 communities.

^f The physical deprivation score has been computed using information on (1) the proportion of vacant buildings and (2) households without kitchen, (3) phone or (4) plumbing. Similarly to the social disorganization score, all variables have been z-transformed and summed to form the summary score.

^g Information on land use features are drawn from the ACS 2005–2009 and the Pennsylvania Department of Transportation.

^h Street connectivity is the percentage of road intersections that are connected (that are not dead ends or cul-de-sacs). Road connectivity was calculated using geographic information systems software ArcGIS.

ⁱ Vehicle miles travelled are the vehicle miles travelled by all cars on state-maintained roads during the year 2010. Information is drawn from the Pennsylvania Department of Transportation.

^j Parks are included in the count of public activity spaces.

Table 2

Description of BMI z-score quartiles of Pennsylvania communities with more than 50 children aged 10–18 years.

Class of communities	N	Mean BMI z-score (Std. dev.)
Obesogenic (highest quartiles)	49	0.92 (0.10)
3 rd quartile (not used in analysis)	49	0.75 (0.03)
2 nd quartile (ibidem)	49	0.65 (0.04)
Obesoprotective (lowest quartile)	50	0.47 (0.09)

Table 3
Conditional Random Forrest specification details and out-of-bag (OOB) error rates for five models

CRF parameters	Model 1: Full set ^a	Model 2: Food and PA ^b	Model 3: Social only ^c	Model 4: Food only ^d	Model 5: PA only ^e
Number of trees in each forest	5000	5000	5000	5000	5000
Number of variables tried at each split	9	6	3	4	4
Number of variables used to build each forest	44	36	8	19	17
Number of forests in Monte Carlo ensemble	50	50	50	50	50
OOB Error Rates					
Overall Error	0.33	0.31	0.31	0.35	0.35
OPE Class Error ^f	0.32	0.31	0.34	0.41	0.30
OGE Class Error ^g	0.33	0.31	0.27	0.29	0.40

^aModel 1, Full set: Full set of variables includes social, food, land use and physical activity establishment variables.
^bModel 2, Food & PA: Includes food establishment variables and physical activity related features (land use features and physical activity establishments).
^cModel 3, Social only: Social features, only.
^dModel 4, Food only: Food establishment variables, only.
^eModel 5, PA only: Physical activity related features (land use features and physical activity establishments), only.
^fOPE: obesoprotective environments
^gOGE: obesogenic environments

Table 4

Average ranks of conditional permutation variable importance with number of runs above Strobl's threshold for significant contribution, Monte Carlo ensemble of 50 CRF on variable set without socioeconomic features (Model 2) and average ranks and sig ranks of full analysis (Model 1) for comparison. Only variables that contributed at least once to the classification of Model 2 are shown

Variables contributing at least once to analysis without socioeconomic features	Avg ranking, analysis without SES Model 2	Sig runs, Analysis without SES Model 2	Comparison Avg ranking, full analysis Model 1	Comparison Sig runs, full analysis Model 1
Pop density	1	50	1.58	50
Pop change	2	50	4.6	50
Community type	3	50	10.08	50
Vehicle miles travelled	4.3	50	11.98	49
Cnt fast food stores	5.04	50	8.6	50
Cnt snack stores	6.2	50	8	50
Cnt convenience/gas	7.08	50	17.16	27
Street connectivity	7.54	50	16.84	33
Cnt bars hotels with foods	9.54	50	22.88	4
Cnt outdoor fitness club	9.84	50	15.88	36
Diversity food establishments	10.84	50	19.28	6
Cnt pharmacies/drug stores	13.18	41	23.26	4
Cnt any phys act establishment	13.32	39	13.24	50
Cnt outdoor rec clubs/org	13.86	33	41.7	10
Cnt any food establishment	15.46	14	21.5	6
Cnt limited service restaurant	16.78	9	20.56	11
Cnt retail stores	16.78	6	19.72	15
Cnt specialty food	17.74	2	30.34	0
Population	18.42	2	26.44	3

Table 5

Spearman's Correlation of supervisor (obesogenic- and obeso-protective environments) and the 13 variables with highest variable importance ranking

Variables	Spearman's Correlation coefficient
Unemployment	0.42
Pop Density	0.35
Social Disorganization	0.4
Less than High School	0.4
Pop Change	-0.42
No Car Ownership	0.35
Public Assist	0.33
Cnt Snack Stores	-0.21
Cnt Fast Food Chain	-0.13
Poverty	0.35
Vehicle Miles Travelled	-0.31
Any Physical Activity Establishment	0.18
Outdoor fitness establishment	-0.23