EXTENDED ABSTRACT
Population Association of America Annual Meeting 2014

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

C. Nau[1], A. Hirsch[2], L. Bailey-Davis[2], B. S. Schwartz[1, 2], J. Pollak[1], A. Liu[1], T. A. Glass[1]

[1] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
[2] Geisinger Center for Health Research, Danville, PA

SHORT ABSTRACT

Past research has focused on assessing the association of single neighborhood characteristics with health ignoring spatial co-occurrence of multiple community-level risk factors. We demonstrate the use of random forests (RF), a non-parametric machine learning approach to identify the combination of community features that best predict obesegenic and obesoprotective environments for children. We use data from electronic health records on >160,000 children living in 1289 Pennsylvania communities and include a large number of contextual variables, previously linked to childhood obesity to analyze the joint, spatially co-occurring distribution of features of the food, land use, physical activity and social environments. This analysis allows us to (1) identify the combination of features that render an environment obesogenic, (2) determine their relative importance, and (3) provide evidence regarding the time-lag with which they operate. RF allows consideration of the neighborhood as a system of risk factors, an approach more likely to reflect residents' experiences.

INTRODUCTION

Most prior studies have assessed the influence of isolated neighborhood characteristics such as average neighborhood income, crime rates, or walkability scores, on a particular health outcome, treating each "exposure" in isolation, despite the well-known co-occurrence of multiple features of community risk. This approach runs the risk of committing what the sociologist Gordon called the "partialling fallacy" (Gordon 1968). Past research found that the effects of single community variables in isolation are often small (even if they are statistically significant) (Pickett and Pearl 2001; Robert 1999). Is it that neighborhoods do not matter, or, are we fragmenting the effect that neighborhoods have by looking at one part of a larger whole at a time? We sought to identify what combination of factors make up the experience of a "bad neighborhood." Combinations of features that have small effects individually may constitute a broader risk landscape, or what Rhodes has called a risk environment, in which the sum of accumulated risk is greater than its parts (Rhodes 2009).

We used "big data" from a large health care system to characterize the prevalence of obesity in a diverse set of communities. We assembled a range of variables on multiple dimensions of community features to demonstrate an innovative method that allows identification of a network of community characteristics that, in combination, characterize a high-risk community environment.

This study combined a theory driven approach with a data guided strategy to expand the consideration of neighborhood characteristics to a more holistic study of a set of factors that make up the experience of residents. We began with a large set of neighborhood features that had been linked in previous studies to childhood obesity and applied a machine learning technique called random forests that allowed ranking variables by their importance in differentiating high obesity from low obesity communities. We were thus able to examine the joint, spatially co-occurring distribution of features of the food, land use, physical activity and social environments. Results of this analysis allowed: (1) identification of the combination of features that render an environment obesegenic; (2) determination of the relative importance of particular environmental features compared to each other; and (3) generation of evidence regarding the time-lag with which these features were operating.

DATA

This study draws data from electronic health records of the Geisinger Health System. We used data on measured weight and height from 161,771 children aged 3-18 residing in 1289 communities in eastern and central Pennsylvania. The Geisinger population is approximately representative of the general population in the same geographic area (Liu et al. 2013). This large area of Pennsylvania, comprising approximately 40 counties, is characterized by communities that range from low-density rural places to high density urban neighborhoods. To operationalize community context, we use a mixed definition that consisted of census tracts in urban areas and minor civil division boundaries for townships and boroughs. We investigated four domains of community characteristics that have been linked to obesity prevalence in previous studies: community socioeconomic and demographic characteristics (Janssen et al. 2006; Matheson, Moineddin and Glazier 2008; Stafford et al. 2010), the built and land use environments (Frank et al. 2007; Franzini et al. 2009; Rundle et al. 2009; Schwartz et al. 2011), the food environment (Fleischhacker et al. 2011; Inagami et al. 2006) and the physical activity environment (Gordon-Larsen et al. 2006; Kipke et al. 2007). Results for the first two sets of factors are included here. For each domain we gathered and geocoded data from a variety of sources including InfoUSA and Dunn and Bradstreet for commercial establishments and the United States Census Bureau (see Table 1 for variables included in the preliminary analysis, variables on healthy and fast food access, for physical activity establishments, as well as the indicator variables included in the factors of social and physical disorganization will be added for the final analysis).

Each set of community features were measured in 2000 and 2010. We limited our analysis to children whose BMI's were measured in 2010 to assess the simultaneous and lagged effect of the community environment. Childhood BMIs, expressed as a z-score relative to the 2000 CDC population average growth curves, were used to compute an average BMI-z at the community level. To avoid unstable estimates of mean BMI-z, only communities with at least 50 children with valid BMI were included (N=285). Obesogenic communities were defined as communities with a mean BMI-z in the upper quartile of the community level BMI distribution, while obesoprotective communities were those that were in the lowest quartile.

METHOD

We use a non-parametric machine learning approach, random forests (RF), to identify the set of variables that were best at distinguishing obesogenic from obesoprotective communities. Random forests is a classification approach frequently used in engineering to identify characteristics of complex systems that are then incorporated into systems dynamics models. Its algorithm is supervised by an outcome, in our case the "obese" vs. "non-obese" community classification. Our RF model then used a large set of community characteristics to classify (or predicted) obesogenic and obesoprotective communities. In the process, RF generated a variable importance list that indicated the relative classification salience of each variable. Based on its classification results it then predicted whether a community was obese or non-obese and it calculated the error rates for the overall classification success of both types of environment as well as for the success of classifying each community type.

Methodologically, RF is a classification approach similar to that of classification trees (Malley, Malley and Pajevic 2011). Instead of growing one classification tree however, RF grows many trees on bootstrap samples of the initial dataset. RF bases each split on a random sample of a small, pre-defined number of variables (Liaw and Wiener 2002). Final results are calculated by averaging the results across all trees. The error rate is calculated by predicting, at each bootstrap iteration, the non-sampled data (out-of-bag sample) with the tree grown on the bootstrap sample. In other words, the result of each of our trees was used to predict the community type in the data that was not used to grow the tree. Results from the prediction on the unused dataset were compared with the actual class of each community and the error rate was calculated on the aggregate of all out-of-bag predictions across all trees (Liaw and Winer 2002). The variable importance ranking was, similarly, computed for each tree and then averaged across all trees. It was computed by assessing how much the prediction error increased when the values of a particular variable were randomly switched (Shih 2011). The bigger the error, the more important the variable. In this analysis we used the Gini-mean decrease, an importance measure commonly used in RF analysis. This measure was key to our analysis because it allowed ranking community-level variables on how important they were for predicting whether a community was obesogenic or not.

The key advantage of RF is that it is among the most accurate and robust learning algorithms. Also, because it uses a random sample of variables at each split, it can find predictors whose influence is small and would not be detected in conventional classification approaches, but may be important in improving the overall classification accuracy of a set of variables where at each split all predictors are used to decide on the split (Shih 2011).

We ran 5000 trees for each analysis and chose, as recommended by Strobl (2009), the number of variables at each split to be m, the square root of the number of variables. Sensitivity analysis showed that, with smaller or bigger m the classification success decreased. The analysis was conducted first using predictors for 2000 and 2010 simultaneously. Because results showed that for all variables the 10 year lag measures performed better than the current measures, the analysis was repeated with only the 2000 measures. Next, different sets of the most important variables were used to measure the joint influence of the less important variables. Each set of analyses was repeated with at least five different seed-numbers for the random number generator. We also varied the number of trees to assess the stability of the results. This was necessary because RF is a stochastic method whose results vary (usually only slightly) from run to run (Shih 2011).

Preliminary results reported below used community characteristics measured in 2000 and 2010 to classify the community mean BMI-z in 2010. We will extend this analysis further by predicting community BMI z-scores derived from a multi-level analysis that controls for compositional differences in communities according to the distribution of race/ethnicity and a proxy-measure for family-level deprivation.

PRELIMINARY RESULTS

The goal of this preliminary analysis was to demonstrate the use of RF models for identification of the set of characteristics that rendered a community high-risk for obesity and to use it to identify the time lag over which these characteristics operated. The final analysis for this paper will also include information on the food and physical activity environments.

Table 2 presents the sample of 285 communities that in 2010 had at least 50 resident children on whom we collected at least one valid measured BMI. Obesogenic communities, had an average BMI-z of 0.84 and those that were obesoprotective had a notably lower, yet relatively high, average BMI-z of 0.36.

We began our analysis by including all predictors for each of the two time periods 2000 and 2010 (results not shown). Across all predictors the variables measured in 2000 were stronger classifiers of obese communities than were 2010 measures. Table 3 presents the classification errors for the analysis that included all 22 community characteristics measured in 2000. Overall, the 22 characteristics, in combination, correctly classified 64% of communities. Obesogenic places were more successfully classified; 70% of these communities could be classified with the set of 22 predictors. In comparison, 61% of obesoprotective places were correctly classified, suggesting that obesoprotective places were either more heterogeneous and/or that they were characterized by additional factors not yet captured in our analysis. Figure 1 presents the Gini-mean decrease, our variable importance measure, and illustrates that indicators of socioeconomic deprivation, namely the percent of the population without a high school diploma, percent population in poverty, unemployed and on public assistance were the most powerful characteristics for differentiating obesogenic and obesoprotective places. These variables were followed by the factor score for social disorganization and three demographic characteristics of communities: population size, population change from 1990 to 2000, and population density. The latter two might suggest that economic conditions that lead to migration, or alternatively, to community differences in urbanization and sub-urbanization, may be at work. The role of population density requires further investigation. Its effect did not seem to be linked to urbanicity or the community type (census tract, borough or township) since these variables scored very low in terms of their variable-importance.

Figure 1 also shows that variables are grouped into sets with similar importance scores. We used the 10 most important variables, all of which were indicators of socioeconomic deprivation or population characteristics, to assess if they would be sufficient to reproduce the preliminary classification. Table 3 shows that the overall error rate increased to 38.95 %, which indicates that the reduced set of 10 variables was insufficient to reproduce the classification success achieved with 22 variables. Addition of the next set of three predictors, road connectivity, vehicle miles travelled per capita and per square mile improved the overall error rate and both class error rates. For this preliminary analysis we retained a set of 13 predictors from three domains: traffic characteristics, socioeconomic measures, and population characteristics that allowed identification of 66% of all obesogenic places and 60% of obesoprotective places.

CONCLUSION

We used a machine learning technique on an initial set of candidate variables that have been linked to obesity by prior research. We found that variables measured with a 10 year time lag were better predictors of current community-level childhood BMI than were those concurrently measured. We identified three sets of characteristics, socioeconomic, demographic and traffic characteristics that, in this preliminary analysis, differentiated 66% of obesogenic communities and 60% of obesoprotective communities. The differential success in classifying communities further suggests that obesoprotective environments may be more diverse than the obesogenic environments. We are extending this analysis to incorporate variables for the food environment that include food establishment density and food accessibility as well as density and accessibility to physical activity establishments including parks, health and fitness clubs, and sports facilities. The final analysis will be conducted on predicted values from a multilevel analysis to eliminate potential confounding by individual-level compositional effects.

Random forests offer an innovative and flexible modeling tool for operationalizing risky environments in an ecological manner. RF allows considering the residential environment as a system of factors, an approach that is more likely to resemble the resident’s experience than a variable-by-variable exploration of the environment.

TABLES AND FIGURES

Table 1: Community features by domain (preliminary analysis)

Socioeconomic Deprivation	Demographic Characteristics	Street Network/Traffic	Land use/Urbanization
-Percent no HS -Percent out of labor force -Percent in poverty -Percent unemployed -Percent no car -Percent Pub Assist -Social disorganization factor score -Physical disorganization factor score	-Population Change 90-00 -Population Density Change 90-00 -Population 00 -Household count	-Road length -Road Intersect density -Road connectivity -Vehicle Miles Traveled (VMT)/sq mile -Vehicle Miles Travelled/ per capita	-Average block size -Population density -Household Density -Urban density -Community type

Table 2: Descriptive of Obesogenic/Obesoprotective Communities

	N/mean
Number of communities	285
Average number of children per community	215 (range: 50-1130)
Average community BMI-z all 285 communities	0.6
Average BMI-z obesoprotective communities	0.36
Average BMI-z obesogenic communities	0.84

Table 3: Error Rate (across both community types) and Class Error Rate (for each type of community) for three random forest analyses

	22 characteristics	10 most important characteristics	13 most important characteristics
Out of Bag (OOB) Error Rate	35.79	38.95	37.19
Class Error Obesogenic	30.69	34.97	33.56
Class Error Obesoprotective	41.54	42.95	40.80

Figure 1: Variable importance ranking of 22 community characteristics linked to obesogenic environments

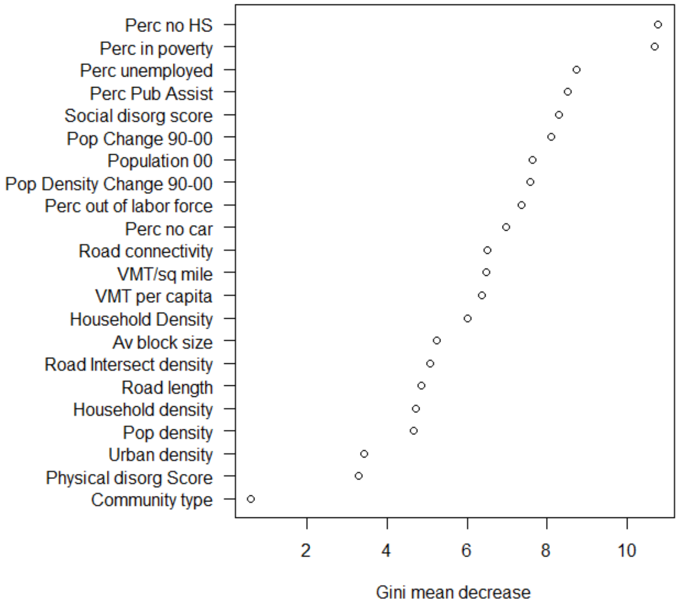
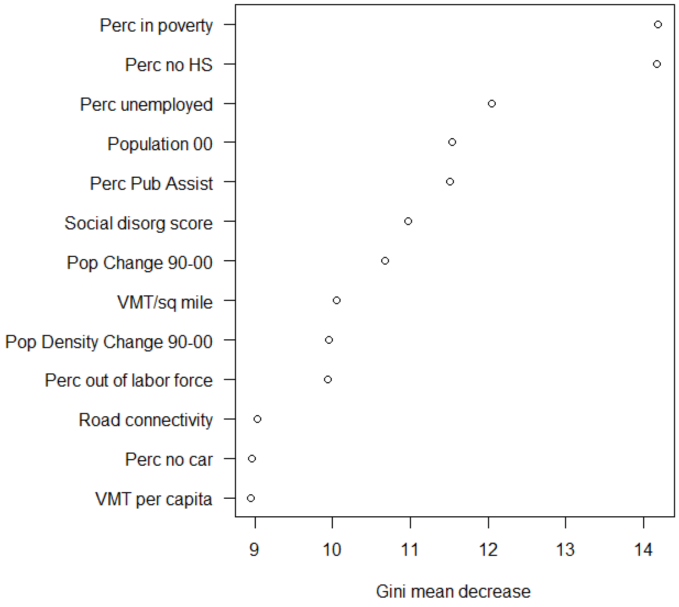


Figure 2: Variable importance ranking of 13 community characteristics linked to obesogenic environments



REFERENCES

Fleischhacker, S., K. Evenson, D. Rodriguez, and A. Ammerman. 2011. "A systematic review of fast food access studies." *Obesity reviews* 12(5):e460-e471.

Frank, L.D., B.E. Saelens, K.E. Powell, and J.E. Chapman. 2007. "Stepping towards causation: do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity?" *Social science & medicine (1982)* 65(9):1898.

Franzini, L., M.N. Elliott, P. Cuccaro, M. Schuster, M.J. Gilliland, J.A. Grunbaum, F. Franklin, and S.R. Tortolero. 2009. "Influences of physical and social neighborhood environments on children's physical activity and obesity." *Journal Information* 99(2).

Gordon-Larsen, P., M.C. Nelson, P. Page, and B.M. Popkin. 2006. "Inequality in the built environment underlies key health disparities in physical activity and obesity." *Pediatrics* 117(2):417-424.

Gordon, R.A. 1968. "Issues in multiple regression." *American Journal of Sociology* 73(5):592-616.

Inagami, S., D.A. Cohen, B.K. Finch, and S.M. Asch. 2006. "You are where you shop: grocery store locations, weight, and neighborhoods." *American journal of preventive medicine* 31(1):10-17.

Janssen, I., W.F. Boyce, K. Simpson, and W. Pickett. 2006. "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* 83(1):139-145.

Kipke, M.D., E. Iverson, D. Moore, C. Booker, V. Ruelas, A.L. Peters, and F. Kaufman. 2007. "Food and park environments: neighborhood-level risks for childhood obesity in east Los Angeles." *Journal of Adolescent Health* 40(4):325-333.

Liu, A.Y., F.C. Curriero, T.A. Glass, W.F. Stewart, and B.S. Schwartz. 2013. "The contextual influence of coal abandoned mine lands in communities and type 2 diabetes in Pennsylvania." *Health & Place*.

Malley, J.D., K.G. Malley, and S. Pajevic. 2011. *Statistical learning for biomedical data*: Cambridge University Press.

Matheson, F.I., R. Moineddin, and R.H. Glazier. 2008. "The weight of place: a multilevel analysis of gender, neighborhood material deprivation, and body mass index among Canadian adults." *Social science & medicine* 66(3):675-690.

Pickett, K.and M. Pearl. 2001. "Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review." *British Medical Journal* 55(2):111.

Rhodes, T. 2009. "Risk environments and drug harms: a social science for harm reduction approach." *Int J Drug Policy* 20(3):193-201.

Robert, S. 1999. "Socioeconomic Position and Health: The Independent Contribution of Community Socioeconomic Context " *Annual review of sociology* 25(1):489-516.

Rundle, A., K.M. Neckerman, L. Freeman, G.S. Lovasi, M. Purciel, J. Quinn, C. Richards, N. Sircar, and C. Weiss. 2009. "Neighborhood food environment and walkability predict obesity in New York City." *Environmental Health Perspectives* 117(3):442.

Schwartz, B.S., W.F. Stewart, S. Godby, J. Pollak, J. DeWalle, S. Larson, D.G. Mercer, and T.A. Glass. 2011. "Body mass index and the built and social environments in children and adolescents using electronic health records." *American journal of preventive medicine* 41(4):e17-e28.

Shi, Stephanie. “Random Forests for classification Trees and Categorical Dependent Variables: An informal Quick Start R Guide”. (2011): http://www.stanford.edu/~stephsus/R-randomforest-guide.pdf

Stafford, M., E.J. Brunner, J. Head, and N.A. Ross. 2010. "Deprivation and the development of obesity: a multilevel, longitudinal study in England." *American journal of preventive medicine* 39(2):130-139.