Title

HOW CLOSE AND HOW MUCH? LINKING HEALTH OUTCOMES TO BUILT ENVIRONMENT SPATIAL DISTRIBUTIONS.

Permalink
https://escholarship.org/uc/item/1b18k8c4

Journal
The Annals of Applied Statistics, 17(2)

ISSN
1932-6157

Authors
Peterson, Adam
Berrocal, Veronica
Sanchez-Vaznaugh, Emma
et al.

Publication Date
2023-06-01

DOI
10.1214/22-AOAS1687

Peer reviewed

Published in final edited form as:
Ann Appl Stat. 2023 June ; 17(2): 1641–1662. doi:10.1214/22-AOAS1687.

HOW CLOSE AND HOW MUCH? LINKING HEALTH OUTCOMES TO BUILT ENVIRONMENT SPATIAL DISTRIBUTIONS

Adam T. Peterson[1], **Veronica J. Berrocal**[2], **Emma V. Sanchez-Vaznaugh**[3], **Brisa N. SÁnchez**[4]

[1]Department of Biostatistics, University of Michigan-Ann Arbor

[2]Department of Statistics, University of California-Irvine

[3]Department of Public Health, San Francisco State University

[4]Department of Biostatistics and Epidemiology, Drexel University

Abstract

Built environment features (BEFs) refer to aspects of the human constructed environment, which may in turn support or restrict health related behaviors and thus impact health. In this paper we are interested in understanding whether the spatial distribution and quantity of fast food restaurants (FFRs) influence the risk of obesity in schoolchildren. To achieve this goal, we propose a two-stage Bayesian hierarchical modeling framework. In the first stage, examining the position of FFRs relative to that of some reference locations - in our case, schools - we model the distances of FFRs from these reference locations as realizations of Inhomogenous Poisson processes (IPP). With the goal of identifying representative spatial patterns of exposure to FFRs, we model the intensity functions of the IPPs using a Bayesian non-parametric model, specifying a Nested Dirichlet Process prior. The second stage model relates exposure patterns to obesity. We offer two different approaches to carry out the second stage; they differ in how they accommodate uncertainty in the exposure patterns. In the first approach the odds of obesity at the school level is regressed on cluster indicators, each representing a major pattern of exposure to FFRs. In the second, we employ Bayesian Kernel Machine regression to relate the odds of obesity to the multivariate vector reporting the degree of similarity of a given school to all other schools. Our analysis on the influence of patterns of FFR occurrence on obesity among Californian schoolchildren has indicated that, in 2010, among schools that are consistently assigned to a cluster, there is a lower odds of obesity amongst 9th graders who attend schools with most distant FFR occurrences in a 1-mile radius as compared to others.

atpvyc@umich.edu .

SUPPLEMENTARY MATERIAL
We provide additional supporting plots and tables that show 1. Descriptive statistics of our data. 2. The Estimated Intensity Functions on the real line. 3. Spatial plot showing coclustering probabilities across California. 4. Mode GLM and BKMR results fit to full dataset 5. Posterior mode vs consensus glm probability of obesity estimates. 6. Simulation intensities and numerical summaries comparing traditional methods to our proposed methods across simulated scenarios. 7. Model Convergence Diagnostics. 8. Plot illustrating BKMR distance and co-clustering properties.
().

Keywords

Food environment; Child Obesity; Inhomogenous Poisson Process; Density Estimation; Nested Dirichlet Process

1. Introduction.

The dramatic increase in child obesity is one of the most pressing public health issues of the 21st century (Sacks, Swinburn and Xuereb, 2012). Indeed, obesity prevalence demonstrates linearly increasing trends among children of school age (Skinner et al., 2018). The need for population-level interventions, beyond individual-level treatments, has been strongly emphasized by the research community and policy makers alike (McGuire, 2012). Place-based interventions are one realm of population level approaches that seek to modify neighborhood environments in ways that can support residents' health promoting behaviors. Changes to the distribution of health-supportive (or detrimental) point-reference environmental characteristics within neighborhood environments have emerged as a possibility, given that the built environment – the human made space in which humans live, work and recreate on a day-to-day basis – constrains everyday health-relevant choices (Roof and Oleru, 2008).

The potential contribution of the food environment near schools (e.g., fast food restaurant availability) to child obesity has been studied extensively (Currie et al., 2010; Davis and Carpenter, 2009; Sanchez-Vaznaugh et al., 2019; Sánchez et al., 2012; Baek et al., 2016), as children spend large proportions of their waking hours and consume a large proportion of their food within and near the school environment. While the body of evidence supports these connections broadly, different approaches to conceptualize exposure make it challenging to more fully understand the health effects of environmental exposures, as well as identify where interventions may be especially needed. To assist policy makers with these challenges, methods need to be developed that both (i) identify different spatial patterns of exposure and (ii) link these patterns to health outcomes quantitatively. Exposure patterns, compared to continuous exposure measures, may make it more straightforward to identify places in higher need of interventions.

Previous work has approached these problems by first clustering some measure of built environment features (BEFs) and then incorporating cluster assignments as a categorical predictor into a second stage regression model. For example, Wall et al. (2012) used a spatial latent class analysis (LCA) to cluster multivariate measures of the built environment, including the density of food outlets within 1 mile of the subjects residential location, and subsequently estimated the association between cluster membership and adolescent obesity. However, clusters identified with these traditional buffer-based exposure counts ignore the spatial distribution of BEFs within the buffer. This spatial distribution is relevant from a mechanistic perspective because BEFs closer to schools are easier to access, as well as policy relevant since the distribution could inform built environment interventions such as zoning laws to curtail exposure. Finally, using an estimate of cluster membership as a predictor in a health outcome model does not account for the uncertainty in the estimated cluster assignment label, leading to potentially incorrect inference of the associated health effect.

Motivated by the need to better understand how proximity of FFRs to schools, beyond their number, is associated with child obesity, this paper has two complementary goals. First, we aim to develop a clustering procedure that provides interpretable groupings of BEF (e.g. FFR) exposure that is based on the spatial distribution of BEFs. For this goal, we work with the geographical coordinates of BEFs and schools, modeling the set of distances of each school to its nearby BEFs as a realization of a 1-dimensional Inhomogeneous Poisson process (IPP) with a school-specific intensity function (Diggle, 2013). Clusters of schools are formed by clustering the intensity functions using a Nested Dirichlet Process (NDP) (Rodriguez, Dunson and Gelfand, 2008). Working directly with point-level data and distances, instead of buffer-based counts, maintains the level of granularity needed to form clusters based only on the proximity of BEFs to schools. In turn, this allows us to separate the effect of the spatial distribution of BEFs around schools on children's obesity from the effect of the quantity of BEFs in the buffer. This is important because quantity can vary greatly across levels of urbanicity, and proximity is a separate dimension of accessibility to BEFs beyond quantity. This separation of effects may thus provide new insights compared to prior work. Second, we show two ways to use the output from the NDP clustering model to address cluster assignment uncertainty when using clusters as predictors in a regression that evaluates the association between FFR exposure near schools and students' obesity risk.

Clustering methods vary widely, from the model-based finite mixture models (FMM) (Diebolt and Robert, 1994) and LCA (Wall and Liu, 2009), to algorithmic K-means style methods (Hartigan, 1975; Friedman, Hastie and Tibshirani, 2001). Each of these have varying strengths and weaknesses according to the problem at hand. FMMs, K-means and LCA rely on pre-specifying the number of clusters that should be found in the data and make parametric assumptions about the relevant distribution or metric that define the clusters. Our use of the NDP (Rodriguez, Dunson and Gelfand, 2008) allows us to cluster schools without enforcing strong parametric constraints on the shape of the intensity functions in the IPP (introduced above and described in detail in Section 3) or prespecifying the number of clusters. Akin to Xiao, Kottas and Sansó (2015), our model factorizes the intensity functions in the IPP into the product of a normalized intensity function, modeled non-parametrically, and the total number of BEFs. This factorization enables clustering based on proximity of FFRs independently of FFR quantity. Our model differs from that of Xiao, Kottas and Sansó (2015)'s in our use of the IPP to model a spatial process instead of a temporal process, and a NDP to create clusters of subjects (schools) instead of using a dependent Dirichlet process (DDP) (MacEachern and Shen, 1999; MacEachern, 2000) to capture the temporal dependencies.

Additionally, in accordance with our conceptual objective (i), the NDP provides cluster assignment labels which can be processed and used in a second-stage regression analysis to estimate associations between the BEF's spatial distribution and a health outcome of interest. Second-stage models raise the need to accommodate uncertainty in estimated exposures, in this case cluster assignment (Chiang et al., 2017; Graziani, Guindani and Thall, 2015; Wall et al., 2012; Wade et al., 2018). We explore two approaches to using the output of our NDP clustering model in a second-stage analysis to handle the challenges of making cluster assignments, given that the NDP does not constrain the number of clusters to be fixed and consequently produces a varying number of cluster assignments across posterior samples. One approach relies on using a conservative “consensus of cluster assignments” determined from cluster-specific uncertainty bounds (Wade et al., 2018). The second approach avoids a single cluster assignment by using the matrix of co-clustering probabilities among schools to form an input to a Bayesian kernel machine (BKMR) regression model for the health outcome (Bobb et al., 2015; Valeri et al., 2017). This latter approach is an innovation in terms of expanding the applications of BKMR, as well as a way to utilize a clustering model’s output to address classification uncertainty.

Section 2 discusses the data sources used in our analysis of child obesity in relation to FFR occurrence near their schools, including data processing needs for our proposed methods, as well as preliminary data analysis. Section 3 describes our proposed NDP clustering approach and the second stage health analysis models. Section 4 contains the results from fitting our models to the California data, as well as comparison of our proposed models to traditional modeling approaches with both simulated and California data. We finish with a discussion of our contribution to the built environment literature, limitations of our approach and possible methodological extensions.

2. Data on child obesity and food environment near schools in California.

2.1. Data sources and study sample.

By state mandate, California public schools collect data on the fitness status of 5^{th}, 7^{th} and 9^{th} graders using the Fitnessgram battery. The Cooper Institute’s sex-, age- and height-specific standards for body weight are used to classify each child as “meeting the standard”, “needs improvement”, or “needs improvement, high risk”, which correlate to normal, overweight, and obese classifications. We use the last two of these as “not meeting the standard”, and use the term obesity henceforth when referring to this outcome. We use data collected during academic year 2009-2010 on 9th graders only, since high school youth are more likely to be exposed to the food environment surrounding their school (e.g., students may leave the campus for lunch).

Fitnessgram as well as school-level characteristics are available through the California Department of Education (CDE) website (https://www.cde.ca.gov/ds/), including schools’ geocodes. We use geocodes to link schools to census tract level covariates and to calculate the distances between the school and the geocoded location of each FFR in California. FFRs were identified from the National Establishment Time Series (NETS) database (Walls, 2013), using a published algorithm that classifies specific food establishments as FFRs (Auchincloss et al., 2012). Only distances shorter than one mile were kept for this analysis, because previous work shows FFRs cease to have an effect on childhood obesity at approximately one mile from schools (Baek et al., 2016). Finally, we calculated the distance between all schools, to derive a data set of schools that are at least one-mile away from one another, to satisfy independence assumptions used in the analysis.

2.2. Preliminary analysis.

The data comprise 420,085 children who attended 1,193 high-schools. Forty percent of the children had obesity, and 64% of schools had ≥ 1 FFR within one mile. While the second stage analysis modeling obesity includes *all* schools, the first stage analysis that derives clusters of schools with similar spatial distribution of surrounding FFRs includes only the schools with ≥ 1 FFRs within one mile of their location.

Descriptive statistics of the schools are presented in Table S1 in the Supplementary Material, for the entire dataset and for the two subsets of schools with and without FFRs within a mile. Aside from having ≥ 1 FFR, schools in the first stage analysis are more likely to be located in urban areas (46%) compared to schools not included (27%). Schools included in the first stage analysis varied in terms of the number and spatial distribution of nearby FFRs. Among these schools, 45% had 1 to 4 FFRs within one mile, while the rest of the schools had at least 5; the median (Q1-Q3) distance to the first FFR was 0.4 (0.3-0.6) miles.

Examining the empirical cumulative distribution function (ECDF) of the distances between each school and its neighboring FFRs provides a richer understanding of schoolchildren's exposure to FFRs and illustrates how using buffer counts or distance to the closest FFR, may fail to incorporate meaningful aspects of spatial exposure. Figure 1A illustrates how schools with a similar number of FFRs within a given distance may be characterized by dramatically different spatial distributions of surrounding FFRs. Likewise, Figure 1B illustrates that while certain schools may have the same distance to the first outlet and/or have similar distribution of distances to FFRs, the total number of nearby FFRs might be completely different. Figure 1C shows the distribution of school-FFR distances for 100 randomly selected high schools, further demonstrating the wide variability in the spatial distribution of FFRs in our dataset. Figure 1 thus illustrates the need for improvements in characterizing spatial distribution (or proximity) and quantity of BEFs as separate, albeit related, dimensions of exposure.

3. Model and Estimation.

Our analysis to distinguish the association between obesity and the spatial distribution and quantity of FFRs is based on a two-stage approach: (i) a first stage model characterizes the main patterns of school-level exposure to FFR, by clustering the spatial distribution of FFRs near schools independently of their quantity; (ii) a second stage model uses the output from stage 1 in a regression model with child obesity as the outcome. Here we discuss each modeling stage and corresponding estimation strategy, as well as exposure summaries from the first stage that are input into the second stage model.

3.1. Stage I: Clustering the spatial distribution of FFRs near schools.

To characterize the food environment near schools, our clustering approach focuses on the point processes describing the relative locations of FFRs in the immediate vicinity of the schools, rather than the global 2-dimensional point process representing the location of FFRs across the entire state of California. Specifically, let r_{ij} be the distance between the ith school $(i = 1, \ldots, N)$ and the jth nearby FFR $(j = 1, \ldots, n_i)$: each r_{ij} belongs to the interval $(0, R) \subset \mathbb{R}$, with maximum distance R chosen on substantive grounds. Since the schools in the sample are relatively far from each other (by at least R), the distribution of distances for one school does not inform on the distribution of distances for another. Thus, for each school i, we model the random subset $\mathscr{D}_i = \{r_{ij}; j = 1, \ldots, n_i\}$ as a realization from a one-dimensional Inhomogeneous Poisson Process (IPP) with intensity function $\lambda_i(r), r \in (0, R)$. This realization consists of both the random number of FFR locations near a school as well as the distances to those locations, both of which are governed by the intensity function $\lambda_i(r)$. However, to accomplish our main purpose of separating the proximity of locations (distances) from the number of locations, we follow Xiao, Kottas and Sansó (2015) in decomposing the intensity function $\lambda_i(r)$ as $\lambda_i(r) = \gamma_i f_i(r)$ with γ_i representing the expected number of FFRs within radius R from school i and $f_i(r)$ denoting a normalized density. Thus, the ith school's contribution to the likelihood is:

$$p(\mathscr{D}_i \mid \gamma_i, f_i(r)) \propto \gamma_i^{n_i} \exp\{-\gamma_i\} \prod_{j=1}^{n_i} f_i(r_{ij}), \tag{1}$$

where $f_i(r_{ij})$ is the value of the density $f_i(r)$ evaluated at r_{ij}. Assuming independence among the N collections of distances between school i and the nearby fast food restaurants, $\{r_{ij}\}_{j=1}^{n_i}$ the full likelihood is obtained by taking the product over the N schools' likelihood contributions. Because of the normalization of the intensity function, we note that (1) separates into a component that handles the number, n_i, of FFRs for each school i, and a component that, given n_i FFRs surrounding school i, evaluates the density at each of the n_i distances. We consider $\gamma_i, i = 1, \ldots, N$ as nuisance parameters, as they do not affect the estimation or interpretation of the $f_i(r)$'s beyond what has been previously discussed. In the health outcome model we use the observed n_i's directly as a predictor, instead of their expected values γ_i's, in accordance with our aim to differentiate between the separate effects of the *observed* FFR quantity (a traditional exposure metric) and the FFRs' spatial distribution on child obesity.

Our goal is to simultaneously model and cluster the FFRs spatial density functions $f_i(r)$, $i = 1, \ldots, N$, in a non-parametric fashion. While the non-parametric estimation of a single $f_i(r)$ could be accomplished by using a Dirichlet Process (DP) mixture model (Gelman et al., 2013), we use a NDP modeling approach to accomplish our goal. The NDP uses two DP's simultaneously: one to estimate the normalized intensities, and another to cluster them. Specifically, we express each $f_i(r)$ as:

$$\begin{aligned} & f_i(r) = \int \mathscr{K}(r \mid \boldsymbol{\theta}) dG_i(\boldsymbol{\theta}) \\ & G_i \overset{iid}{\sim} Q \\ & Q \sim DP(\alpha, DP(\rho, G_0)), \end{aligned} \tag{2}$$

where, $\mathscr{K}(r \mid \boldsymbol{\theta})$ is a mixing kernel with parameter vector $\boldsymbol{\theta}$, and the distribution G_i is drawn from the random distribution Q on which we place a NDP prior. In (2), $DP(\rho, G_0)$

denotes a DP with concentration parameter $\rho, \rho > 0$, and parametric base measure G_0. The base measure G_0 is the distribution around which the DP is centered and the concentration parameter, ρ, reflects the variability around that base measure.

The $f_i(r)$'s are clustered through the G_is as can be seen from the stick breaking construction representation of the NDP: $Q = \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*}(\cdot)$. In this representation, π_k^* represents the probability that a school is assigned to the k-th mixing measure, G_k^*, $\delta_{(\cdot)}(\cdot)$ is the Dirac delta function and $G_k^* = \sum_{l=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*}(\cdot)$ is itself composed of weights, w_{lk}^* and associated atoms $\boldsymbol{\theta}_{lk}^*$. This hierarchy of distributions, weights and atoms provides a framework that flexibly identifies clusters of schools, and also flexibly estimates the corresponding normalized intensity function (See Figure S1 in the Supplementary material for a helpful illustration (Peterson et al., 2020)).

Combined altogether, the hierarchical formulation of our model is:

$$
\begin{aligned}
&\{r_{ij}; j = 1, \ldots, n_i\} \overset{ind}{\sim} IPP(\lambda_i(r)), \quad i = 1, \ldots, N \\
&\lambda_i(r) = \gamma_i f_i(r) \\
&f_i(r) = \int \mathscr{K}(r \mid \boldsymbol{\theta}) dG_i(\boldsymbol{\theta}) \\
&G_i \overset{iid}{\sim} Q \\
&Q \sim DP(\alpha, DP(\rho, G_0)) .
\end{aligned}
\tag{3}
$$

As previously noted, given the separability of the likelihood contributions, (1), the γ_i are nuisance parameters that do not influence the estimation of the normalized intensity functions, which are of primary interest.

In our analysis of the California data, we transform the school-FFR distances from $(0, R) \rightarrow \mathbb{R}$ using a probit function to create the unrestricted distances $r_{ij}' = \Phi^{-1}(r_{ij}/R)$. This transformation enables us to use a normal mixing kernel and corresponding normalinverse-chi square base measure, $G_0 = N(0, \sigma) \times \text{Inv} - \chi^2(1,1)$, to facilitate computation. The Beta base measure could also be used here, as in the work of Xiao, Kottas and Sansó (2015), but it is not amenable to computation with large datasets. Additionally, in our preliminary simulations we found no difference in clustering estimates between the two base measures. Furthermore, an infinite mixture of normal distributions has been shown to be sufficient to approximate any distribution (Nguyen and McLachlan, 2019), lending further theoretical justification for our modeling choice. Completing our model specification, we place informative Gamma priors, on the concentration parameters, $\alpha, \rho \sim \text{Gamma}(10,10)$, to encode our *a priori* belief that there should be a small number of clusters (Ishwaran and James, 2001; Gelman et al., 2013; Rodriguez, Dunson and Gelfand, 2008).

3.1.1. Clustering model estimation.—As our modeling approach is specified within a Bayesian framework, inferences on all model parameters are obtained through the posterior distribution, which we approximate using a Markov chain Monte Carlo (MCMC) algorithm. Specifically, we use the blocked Gibbs sampler as described in Rodriguez, Dunson and Gelfand (2008), truncating the summations for the inner and outer DPs using

$G_k^* \approx \sum_{l=1}^{L} w_{lk}^* \delta_{\theta_{lk}^*}$ and $Q \approx \sum_{k=1}^{K} \pi_k^* \delta_{G_k^*}(\cdot)$, where $L = 30$ and $K = 35$. This model fitting routine is implemented within our `bendr` (Peterson, 2020) `R` package.

We drew 250,000 samples from the posterior distribution, with 240,000 discarded for burn-in and the last 10,000 iterations thinned by 3 to reduce auto-correlation. The length of the burn-in period and thinning were determined by inspecting traceplots for various model parameters and by computing Raftery and Lewis' diagnostic statistic (Raftery and Lewis, 1995).

Posterior medians, inter-quartile ranges (IQRs) and 95% credible intervals were calculated for the intensity function parameters, $\left(\mu_{lk}, \sigma_{lk}^2\right)^*, \pi_k^*$. The $\hat{f}(r)$'s were constructed over a fine grid of equally spaced values in $\mathbb{R}$ representing the distances of a BEF from a school, combining the K clusters and the L within-cluster components at each distance. Since we transformed the school-FFR distances from $(0, R)$ to $\mathbb{R}$, we back-transform the normalized intensities onto the $(0, R)$ domain using the inverse probit function, and rescale them by an empirically calculated proportionality constant.

3.1.2. Clustering summaries from stage I model.—

We create two summaries of the clustering model output that characterize the between-school similarity in the spatial distribution of FFRs near schools: (1) discrete cluster labels; (2) a matrix of co-clustering probabilities.

First, we assign cluster labels to the schools such that the set of schools is partitioned into mutually exclusive groups. We derive cluster assignment labels for each school using the posterior samples and a loss function in a decision theoretic framework (Wade et al., 2018). Specifically, we use the variation of information (VI) loss function to determine the optimal cluster configuration, which simultaneously identifies both the number of clusters and cluster labels for the schools. This approach finds the posterior sample that produces the minimal loss, and uses the number of clusters and cluster assignments in that posterior sample to assign labels to schools– thus deriving, essentially, a "point estimate" for the discrete/categorical cluster assignment. We refer to this point estimate as the "mode" cluster label. Cluster labels are derived using the `mcclust.ext2` package in `R` (Wade, 2015). Given our interest in characterizing clustering uncertainty, we note that Wade et al. (2018)'s method also produces 95% uncertainty bounds for both the number of clusters and for the cluster labels for each school, yielding three additional cluster configurations (for a total of four including the mode). Compared to "upper" and "lower" bounds that are typical of credible intervals for univariate parameters in $\mathbb{R}$, Wade et al. (2018) provide three bounds for the cluster configurations which are mapped from a lattice space. The bounds for the cluster configurations may differ from the point estimate in the number of clusters identified, and/or the members (schools) belonging to each cluster. We use the same `R` package to estimate the posterior assignment credible bounds with VI loss function as detailed in Wade et al. (2018), using the `minVI` and `credibleball` functions.

A second summary is the matrix of co-clustering probabilities, $\boldsymbol{P}$ with P_{ij} entry equal to the proportion of posterior samples where schools i and j were in the same cluster, with

P_{ii} equal to 1 by convention. When the underlying true clusters are well separated (i.e., the normalized intensity functions have little overlap), the values of P_{ij} will tend to be concentrated near 0 or 1. When $P_{ij} \approx 1$, we have high certainty that the two schools belong to the same (true) cluster and thus have a highly similar spatial distribution of FFRs near them. Conversely, when $P_{ij} \approx 0$, there is high certainty that the two schools do not belong in the same cluster and thus have different spatial distribution of FFRs nearby. The more likely scenario in practice is that most schools will have a non-zero probability of co-clustering with other schools, reflecting clustering uncertainty related to the amount of available information (e.g., sample size) as well as due to some overlap in the underlying (true) normalized intensity functions. The co-clustering probability matrix can be used to visualize the degree of clustering uncertainty by first re-ordering the rows and columns of $\boldsymbol{P}$ such that schools with high co-clustering probabilities have neighboring indices. We use the algorithm described in Rodriguez, Dunson and Gelfand (2008)'s Supplementary Material to re-order the indices. After this reordering, the heat-map of the co-clustering probability matrix will have distinct blocks when the NDP is able to identify underlying clusters; well-defined blocks in the heat-map reflect less clustering uncertainty.

3.2. Stage II: Health Outcomes Model.

In the second stage of the analysis, we examine whether the spatial distribution of FFRs around schools are associated with obesity of children in the school. We propose two alternative approaches that separately use the NDP clustering summaries described in 3.1.2 and that seek to deal with clustering uncertainty.

3.2.1. Consensus generalized linear model (CGLM).—The first approach *controls* (or reduces) uncertainty in the cluster labels by using, in the health outcome model, only the subset of schools for which the cluster label is known with greater certainty. To identify this subset, we take the intersection over the four cluster configurations that can be obtained from the NDP output using Wade et al. (2018)'s method - the mode, and the three bounds as described in Section 3.1.2 - to arrive at the consensus cluster assignment. Identifying the consensus cluster assignment is possible when clusters are well identified and posterior samples do not exhibit label-switching across iterations – as is our case – or a post-processing step that adjusts for label switching has been run (Gelman et al., 2013; Rodríguez and Walker, 2014; Stephens, 2000; Papastamoulis, 2016). These conditions ensure that cluster labels are consistent across configurations and, consequently, taking the intersection has a consistent meaning. While using each of the four assignments in separate health outcome regression models, and subsequently fuse together their results may be possible, that would entail fusing models with potentially a different number of clusters. We name our proposed approach *consensus generalized linear model* (CGLM); and use the term 'mode GLM' (MGLM) to refer to a model that simply uses the mode (or "point estimate") cluster labels.

To define the CGLM and enable us to distinguish the association between the quantity of FFRs and obesity from that of the FFRs' spatial distribution, let $C_{i,k} = I$(ith school belongs to cluster k), $k = 1, \ldots, K$, be an indicator variable equal to one when the i-th school is assigned to cluster k and zero otherwise. Note that the cluster indicators are available only for schools where the number of FFR $n_i > 0$; however, we bring back into consideration the schools that had zero FFRs within 1 mile. Thus, the schools included in the CGLM model are both those with $n_i = 0$ or those with $n_i > 0$ that are determined by the consensus approach to have a cluster label with relatively higher certainty. This set of schools is denoted as $\mathbb{D}_{Consensus}$. In addition, define $Q_{i,m}, m = 0, \ldots, 5$ a set of indicator variables that treat the number of FFRs as a categorical variable. In particular $Q_{i,m} = I(n_i \in B_m)$, with $B_0 = \{0\}$ and the other categories being defined as $B_1 = \{2\}$; $B_2 = \{3\}$; $B_3 = \{4\}$; $B_4 = \{5, 6, 7\}$, and $B_5 = [8, \infty) \in \mathbb{N}$. This categorical representation is used in our analysis given the lack of linearity in the association between FFR quantity and the odds of obesity, though other parametrizations of n_i could be used.

The CGLM outcome model is thus:

$$\mathrm{logit}(\mathrm{p}_{i'}) = \left\{\sum_{m=0}^{5} Q_{i',m}\zeta_m\right\} + I(n_{i'} > 0)\left\{\sum_{k=1}^{K} C_{i',k}\xi_k\right\} + \boldsymbol{Z}_{i'}^T\boldsymbol{\beta} \quad i' \in \mathbb{D}_{consensus} \tag{4}$$

where $\mathrm{p}_{i'}$ denotes the proportion of obese 9th grade students at the i'th high school, ζ_m and ξ_k are quantity- and cluster-specific coefficients, respectively, and $\boldsymbol{Z}_{i'}$ is a vector of school characteristics (shown in Table 1) without an intercept term. The outcome model specification is completed by specifying flat improper priors for $\boldsymbol{\beta}, \zeta_m, m = 1, \ldots, 5$ and $\xi_k, k = 1, \ldots K$.

In summary, the CGLM approach mitigates the uncertainty in the cluster label assignment by taking the intersection of the four cluster labels to arrive at a more conservative (less uncertain) estimate of the schools' cluster assignment. This reduction of uncertainty in cluster assignment comes at the cost of sample size, which we anticipate will lead to wider credible intervals for the coefficients of the health outcome model. The key advantages of this approach are that it enables a straightforward analysis, and the cluster assignment is more precise than using the single "point estimate" (that is, the cluster mode label) for all the school in the sample. This increase in precision in the cluster label assignment reduces miss-classification error, thereby reducing bias in cluster effect estimates.

3.2.2. Bayesian Kernel Machine Regression (BMKR).—In the second approach, we draw from the Bayesian kernel machine regression literature and include, in the health outcome model, all the clustering information contained in the matrix **P** of co-clustering probabilities through a function $h(\cdot)$ on which we place a Gaussian Process prior. In contrast to using discrete cluster labels, this approach incorporates clustering uncertainty since the co-clustering probabilities are not always 0 or 1. Heuristically, assigning discrete cluster labels is equivalent to thresholding the co-clustering probabilities in **P**, such that schools with co-clustering probabilities above some threshold are assigned to the same cluster, and subsequently discarding both the threshold and the remaining variation of the P_{ij} within and between clusters thereby discarding clustering uncertainty. To define the second outcome model, recall that a BKMR relies on: (a) each subject having separate ‘exposure information’ (e.g., exposure vector); (b) the availability of a metric that captures the similarity between the exposure information of any two subjects; and (c) a covariance function that takes as input the similarity in exposure information between pairs of subjects (Bobb et al., 2015).

Thus, to take into account clustering uncertainty, let the ‘exposure vector’ for each school $i, i = 1, \ldots, N$ be the N-dimensional vector of co-clustering probabilities $\mathbf{P}_i$ defined as the i^{th} row of the co-clustering matrix $\mathbf{P}$. Clearly, $\mathbf{P}_i$ is not a typical exposure vector since none of its entries are interpretable as a higher or lower degree of exposure. While this vector does not quantify exposure in an absolute sense, by definition of co-clustering this vector as a whole indirectly captures information on the FFR distribution near the ith school as it relates to exposure of other schools and thus can serve as an exposure vector for the i^{th} school. Moreover, when schools i and j have a similar spatial distribution of FFRs near them, the vectors $\mathbf{P}_i$ and $\mathbf{P}_j$ will tend to be similar. We use the Euclidean distance between these vectors as the metric measuring their similarity. Supplementary Figure S10 demonstrates that when the co-clustering probability for schools i and j, P_{ij}, is high, the Euclidean distance between $\mathbf{P}_i$ and $\mathbf{P}_j$ is small.

More formally, the vector $\mathbf{P}_i$ is mapped to a scalar $h(\mathbf{P}_i)$, through the function $h(\cdot)$. This function $h(\cdot)$ is provided with a GP prior with mean 0 and Gaussian covariance function $\kappa\left(\cdot, \cdot; \phi, \sigma^2\right)$ that evaluates the covariance between $h(\mathbf{P}_i)$ and $h(\mathbf{P}_j)$. The covariance function depends on the Euclidean distance between the two N-dimensional vectors $\mathbf{P}_i$ and $\mathbf{P}_j$, and the parameters σ^2 and ϕ, which encode, respectively, the marginal variance of each $h(\mathbf{P}_i)$, and the Euclidean distance between $\mathbf{P}_i$ and $\mathbf{P}_j$ at which the correlation between $h(\mathbf{P}_i)$ and $h(\mathbf{P}_j)$ is negligible. Thus, given the inverse relationship between the Euclidean distance of these vectors and the co-clustering probability, the model implies that, *a priori*, $h(\mathbf{P}_i)$ and $h(\mathbf{P}_j)$ are very similar if the two schools have a high co-clustering probability.

As with the CGLM, we incorporate all observations with 0 FFRs into the outcome model and use the indicators for quantity of FFRs nearby, $Q_{i,m}$, to distinguish the effect on obesity associated with the quantity of FFRs from their spatial distribution. Altogether, the second health outcome model is:

$$\begin{aligned} &\operatorname{logit}(\mathrm{p}_{i'}) = Q_{i',0}\zeta_0 + I(n_{i'} > 0)\left\{\tilde{\alpha} + h(\boldsymbol{P}_i) + \sum_{m=0}^{5} Q_{i',m}\zeta_m\right\} + \boldsymbol{Z}_{i'}^T\boldsymbol{\beta} \quad i' \in \mathbb{D}_{Full} \\ &h(\cdot) \sim \mathscr{G}\mathscr{P}(\mathbf{0}, \kappa(\cdot, \cdot \mid \phi, \sigma^2)) . \end{aligned} \tag{5}$$

In (5), $\mathbb{D}_{Full}$ is the set of schools with zero FFRs in addition to the full set of N schools used in the first stage model less those without outcome data. Additionally, $\tilde{\alpha}$ denotes the intercept for schools with at least one FFR, $h(\boldsymbol{P}_{i'}), i' \in \mathbb{D}_{Full}$ is as described above, and $\boldsymbol{\beta}$, and $\boldsymbol{Z}_{i'}$ have the same definition and interpretation as in the CGLM.

To complete the BKMR outcome model specification, coefficients $\boldsymbol{\beta}$, and $\zeta_m, m = 1, \ldots, M$ are given flat priors, while ϕ and σ^2 are each given informative folded $\mathrm{Normal}(1, 3)$ priors to accommodate known identifiability issues (Zhang, 2004). These informative priors were chosen after initial runs with uniform priors on larger intervals of $\mathbb{R}^+$ for both parameters showed that posterior samples were contained in the $(0, 1)$ interval.

3.2.3. Estimation and inference for health outcome models.—For comparative purposes, we fit the BKMR and the model with categorical cluster assignments to both datasets, $\mathbb{D}_{Consensus}$ and $\mathbb{D}_{Full}$. In the latter case, the mode cluster assignment was used to determine cluster specific indicators; the mode or MGLM previously discussed. The models are fit using the Hamiltonian Monte Carlo variant sampler implemented in `stan` (Carpenter et al., 2016) via `rstan` (BKMR) (Stan Development Team, 2020) and `rstanarm` (CGLM,MGLM) (Goodrich et al., 2020). All model fitting was performed within `R` (v3.6.1) (R Core Team, 2019) on a Linux Centos 7 operating system with $2\mathrm{x}3.0$ GHz Intel Xeon Gold 6154 processors.

For each outcome model, we ran 4 independent chains, using different initial values, each ran for 2000 iterations. For each chain, we kept 1000 samples after burn-in, for a total of 4000 posterior samples. Convergence was assessed via split $\hat{R}$ (Gelman et al., 2013) and visual inspection of traceplots.

Posterior median and 95% credible intervals are calculated for regression coefficients $\boldsymbol{\beta}$, cluster effects $\xi_k, k = 1, \ldots, 6$, and $h(\boldsymbol{P}_i), i = 1, \ldots, N$. Additionally, given that schools with $n_i = 0$ cannot be assigned a cluster for obvious reasons, the CGLM includes the non-standard interaction terms between quantity $n_{i'} > 0$ and cluster assignment. As in any model with interactions,the "main effect" of the quantity of FFRs thus depends on the cluster to which schools with $n_{i'} > 0$ were assigned. To estimate this FFR `quantity effect`, for each category of FFR quantity $Q_{i',m}$ with $m > 0$, we marginalize over the cluster assignment to define the probability of obesity given category $Q_{i',m}$, holding $\boldsymbol{Z}_i = 0$ and taking the median across the $s = 1, \ldots, S$ post burn-in MCMC samples, $P\big(\mathrm{Obesity} \mid Q_m, \mathrm{Data}) = \mathrm{median}_s \sum_{k=1}^{K} w_k^s \mathrm{inv} - \mathrm{logit}(\zeta_m^s + \xi_k^s)$, where w_k is the probability of a school being assigned to cluster k in the $\mathbb{D}_{CGLM}$ dataset. Note that for $n_i = 1$, there is no corresponding effect, ζ_1, as this is defined as the average cluster effect by the construction of the model in (4). Similarly, for the BKMR we calculate the FFR quantity effect on schoolchild obesity, now averaging over the $h(\boldsymbol{P}_i)$'s: $P(\mathrm{Obesity} \mid \zeta_m, \mathrm{Data}) = \mathrm{median}_s \frac{1}{|\mathbb{D}_{Full}|} \sum_{i=1}^{|\mathbb{D}_{Full}|} \mathrm{inv} - \mathrm{logit}(\zeta_m^s + h(\mathbf{P}_i)^S)$.

3.3. Model Comparison and Validation.

We compare our proposed methods to traditional modeling approaches using the California data as well as simulations. In separate logistic regressions, we include traditional predictors, namely: (1) the count of FFRs within the 1 mile buffer radius, (2) the distance to the nearest FFR, or (3) both of these. We compare these models on our California data using the Widely Applicable Information Criteria (WAIC) (Vehtari, Gelman and Gabry, 2017). WAIC is asymptotically equal to leave-one-out cross validation and thus represents the ability of the model to predict obesity while penalizing model complexity. In addition, we include MSE as a traditional model performance metric that balances estimation bias and variance in the school-level obesity prevalence. To better evaluate each model's prediction performance, we calculated MSE for the estimates of the proportion of obese students on 20% held-out observations. Thus, these metrics evaluate the models in terms of their predictive ability and estimation accuracy and, consequently, how well the models may help identify schools with greater need for interventions (e.g., high estimated obesity prevalence and/or high number of children with obesity).

Second, in order to further validate our model results in a more general setting, we conduct simulation experiments. Clearly, since the first-stage model relies on the substantive belief that there is heterogeneity in the spatial distribution of BEF locations around schools (i.e., some degree of clustering), our NDP proposed approach will not work when the FFR locations are uniformly distributed, as the NDP will identify only one cluster, resulting in unidentifiable consensus and BKMR models. In real data, this is unlikely to occur, given the strong spatial patterning of businesses more broadly. Hence, our simulations focus on demonstrating that when heterogeneity in the spatial distribution of FFR is present and the outcome is associated with it and/or the quantity of FFRs, our models either outperform or compare well with traditional approaches. In the cases when there is no association between the outcome and the spatial distribution of environmental features, our proposed models perform similarly to traditional models.

In our simulations we assume that three underlying clusters govern the spatial distribution of FFRs near schools, where the three clusters' number and distance of FFRs are generated from a 1-dimensional IPP with cluster-specific intensity functions. We simulate first simulate the number of FFRs within a mile of a school from a Poisson distribution with a mean of 10 for each of three intensity functions. The intensity functions for these clusters have some degree of overlap (see Supplementary Figure S8, Peterson et al. (2020)), given that we found overlap in the estimated intensity functions in our analysis of the CA data. These functions were used to generate distances between FFRs and schools for 150 schools evenly assigned to each cluster (50 schools each). Each school is then assigned a_i children, where $a_i \sim \text{Poisson}(100)$, to reflect the observed variability in the number of students across schools. Given the exposure data, we then generated the proportion of obese children in each school i according to four models for p_i: (a) Cluster Model: $\text{logit}(p_i) = -1.4 + I(\text{Cluster}_i = 1) \times .4 + I(\text{Cluster}_i = 3 \times .2$, (b) Distance Model: $\text{logit}(p_i) = -1.4 + \text{Minimum Distance}_i \times .4$, (c) Quantity Model: $\text{logit}(p_i) = -1.4 + \text{Num FFR}_i \times .1$ and (d) Quantity & Distance Model: $\text{logit}(p_i) = -1.4 + \text{Minimum Distance}_i \times .4 + \text{Num FFR}_i \times .1$. To obtain the number Y_i of obese children in each school, we used the binomial model: $Y_i \sim Binomial(a_i, p_i)$.

We first fit the NDP to the distance data and obtain consensus clusters and co-clustering matrix $\boldsymbol{P}$ as described above, before generating 6 datasets in each of the four settings described above. We fit both our CGLM and BKMR models in each setting and compare against the traditional modeling approaches that employ simple buffer counts, minimum distance to the nearest FFR, or both counts and minimum distance to estimate the exposure risk. For each model fitted to each outcome data set, we construct plots of the density of residuals calculated as follows for a random dataset of the 6 simulated. For each school, at each MCMC iteration, we obtain a predicted number of children with obesity. We then obtain the median of these predicted values across the posterior samples ($\hat{y}_{median}$), and calculate the residual as the difference between this quantity and the observed number of children with obesity at each school. We also tabulate the average WAIC rank as well as the average out-of-sample MSE of the proportion of obesity estimated for each model in the full and consensus dataset setting.

4. Results.

4.1. Spatial Intensity Functions.

All models converged according to visual inspection of traceplots and numerical assessment of the split $\hat{R}$ statistic. Summary statistics for this diagnostic and effective sample size are included in Supplementary Table S3 (Peterson et al., 2020).

The clustering model estimates six clusters with high probability, with the estimates of the cluster-assignment probabilities, π_k^*, beyond the first six effectively negligible when rounding to the hundredths place. The median cluster normalized density estimates, representing the likelihood of finding an FFR at a given distance from a school, are presented in Figure 2, along with the proportion of schools in each cluster. Clusters are labeled according to their mode's proximity to the school, i.e. the cluster which estimates most FFRs are located nearest to schools is labeled cluster 1, and so forth. Supplementary Figure S2 shows the estimated densities along with 95% credible intervals (Peterson et al., 2020). Figure 3 presents a heat map of the co-clustering probability matrix **P**, after re-ordering the rows and columns such that indices for schools with high co-clustering probabilities are near each other.The six clusters are evident, viewing the figure from left to right, followed by the remaining schools which the model cannot cluster consistently.

Table 1 presents summary statistics for the characteristics of the schools included in the six clusters identified, as well as schools that have no FFRs within one mile of their location (labeled "Cluster 0"). There is a weak association between schools' socioeconomic characteristics and cluster membership. For example, Cluster 1's (closest FFRs) median census tract median income is $55,200, Cluster 6's median census tract median income is $67,400. However, this patterning does not include Cluster 0, which has a lower median census tract median income of $53,900. Also, 44% of schools in Cluster 1 have a majority of white students, whereas 38% have predominantly Latino students; for cluster 6, these percentages change to 58% and 16%, respectively. Notably, all clusters contain schools across all urbanicity classification, and include schools with a varying number of FFRs. Thus, the mode cluster is not driven by FFR quantity or broader context (e.g., urbanicity) of the schools.

To assess whether the six identified clusters were geographically concentrated in one or more sub regions of California, and to investigate whether schools tended to co-cluster with nearby schools, we produced spatial plots of the co-clustering probabilities for a randomly selected school – Figure S4 in the Supplementary Material (Peterson et al., 2020). Schools that are more likely to be co-clustered with the other school are not necessarily located nearby.

4.2. Health Outcomes Models.

The proportion of students that are obese is similar in both the consensus and full datasets (Table 2), which is encouraging since schools are not excluded on the basis of the outcome. Schools used to fit the consensus model are less likely to have few FFRs around them - only 21% have 1-4 FFRs vs. 45% in the full dataset.

We discuss both second stage approaches on both datasets, starting with the consensus dataset. Since the BKMR results mirror those of the CGLM, we focus on how these second-stage models reinforce one another rather than describing each individually.

As shown in Figure 4, we observe a monotonic decrease in the probability of obesity as a function of the proximity of FFRs to the school, after adjusting for 1 mile radius quantity of FFRs. Specifically, according to the CGLM, children attending schools consistently assigned to Cluster 6 have a 35% (95% CI: [33%;38%]) probability of being overweight or obese, while, for other clusters, the lower bound estimate of the probability of obesity ranges from 37% to 40%. These results are consistent with the substantive expectation that students who are exposed to FFRs in the immediate environment around schools are more likely to be obese than they would be otherwise: the density of FFRs for schools in Cluster 6 is greatest after 3 quarters of a mile, in explicit contrast to the other clusters. This finding supports prior work suggesting that zoning laws that restrict the placement of fast food restaurants could serve as possible population-level strategies to reduce child obesity (Austin et al., 2005). Given that our model separates proximity and quantity of FFRs, this finding and potential policy recommendation applies independently of the overall number of FFRs across different urban areas.

Figure 4 overlays the results of the CGLM and the BKMR models, when both are fitted to the consensus data set. The figure demonstrates general agreement across both models. However, in addition to the central tendency of obesity risk across the clusters as estimated by the CGLM, the BKMR provides additional information regarding the probability of obesity for children in each school. Thus, beyond potential policy implications of the average obesity risk for children across schools' food environment clusters, the school-level estimates can be used to prioritize individual schools for additional interventions.

Figure S5 in the Supplementary Material shows the estimated probability of obesity as a function of the number of FFRs within a 1-mile radius of a school, calculated as described in Section 3.2.3, from both the CGLM and BMKR fitted to the consensus dataset(Peterson et al., 2020). As the figure indicates, there is a general agreement between the CGLM and BKMR models with respect to the negligible effect of the number of FFRs on obesity after adjusting for the FFRs' spatial distribution and other covariates. The only estimate that stands out from these analyses is the BKMR's estimate of lower obesity for children in schools with 5-7 FFRs nearby, as compared to zero FFRs - a counter intuitive result. However, it is possible that the greater number of FFRs implies greater variety of food choices, including healthier options. The data set in this analysis does not contain information on the specific types of FFRs, beyond the number and location, thus not allowing us to examine this possibility.

As with the consensus data set, the results from fitting a GLM (using only the median cluster assignment) and the BKMR in the full data set are in agreement with each other, as shown in Figures S3 and S6 of the Supplementary Material (Peterson et al., 2020). However, the results from the analysis on the full data set instead of the consensus data identify Cluster 2 as having the lowest probability of obesity, at 37% (95% CI: 36%, 38%), with the probability for all other clusters near or above 40%. Differences in the association between the spatial distribution of FFRs near schools on child obesity, comparing the full and consensus data sets are likely due to the fact that the full dataset contains schools with more uncertain cluster assignments, and thus potentially more prone to miss-classification errors and thus bias in the associations. The quantity effects are similar in the consensus and full data set, and again agree between methods (see Figures S5 and S6 in the Supplementary Material (Peterson et al., 2020)).

4.3 Results of Model Comparison.

We now examine the results from comparing our model to more traditional methods. Considering the CA data first, Table 3 shows that BKMR and CGLM consistently outperform the traditional models in terms of WAIC and out-of-sample MSE.

Figure 5 presents the simulation results. In scenario (a), where cluster indicators determine obesity risk, we see that the BKMR and CGLM perform better than other models, as expected; this is evidenced by the fact that the residuals' densities have sharper peaks at zero and thinner tails for the proposed models compared to other competitor models. Thus, the proposed models have improved prediction. However, even when the model for risk is determined by the distance to the first FFR (minimum distance), (b), or both distance to the first FFR and quantity of FFRs, (d), the BKMR and CGLM continue to perform just as well or better than the fitted model that matches that scenario's generative model. The reasoning for this performance derives from the association between the distance to the first FFR and the cluster intensity functions, through which the BKMR and CGLM gain their information. In the scenario where obesity risk depends only on the count of FFRs in the buffer, the BMKR and CGLM perform similarly to the model that matches the generative model. These results are further substantiated in Table S2 and Figure S9 which we include in the Supplementary Material (Peterson et al., 2020). Both the BKMR and CGLM are found to have consistently better or equivalent performance by WAIC and effectively equivalent in-sample MSE scores.

5. Discussion.

We presented a two-stage modeling strategy that aims to provide built environment investigators with tools to disentangle the contribution of two interrelated but distinct dimensions of availability of BEFs to health outcomes, namely quantity vs. proximity, and identify subjects at greatest risk of negative health outcomes. By implementing methods for point pattern data in a Bayesian non-parametric clustering approach, the first stage model characterizes subjects' exposure in terms of BEFs' spatial distribution near subjects. This approach is an innovative contribution to built environment science because it introduces a new tool to characterize exposures in this field, beyond exposure assessment methods based on the quantity of BEFs as previously done. The second stage links the spatial distribution of BEFs around subjects to a health outcome, while adjusting for the number of FFRs nearby. This enables researchers to estimate the independent effects of quantity and proximity of BEFs, which is challenging to do with traditional modeling approaches. In particular, the introduction of kernel machine regression to the built environment research gives researchers in this area a new tool that has improved prediction performance, and is thus better equipped to identify subjects at greater risk associated with BEF exposures. The twostage modeling strategy allowed us to identify clusters of high schools in California that have FFRs within closer proximity relative to peer schools, and those that have FFRs farther way. Moreover, we found that after accounting for the spatial distribution of FFRs nearby, the quantity of FFRs did not have an independent effect on child obesity in our ecological analysis. This two-stage modeling strategy can be easily adapted to answer questions involving the association between other point-referenced environmental characteristics and health outcomes, for example, availability of parks and measures of physical activity (Evenson et al., 2016), depression (Bojorquez and Ojeda-Revah, 2018), or availability of social engagement destinations and cognition, among others (Besser et al., 2018).

Both second stage modeling approaches are innovative as they offer new ways to incorporate output from a clustering method into a second stage regression model while considering clustering uncertainty, although each with advantages and disadvantages. The CGLM is a novel combination of the idea of using a cluster label derived by a BNP model as a regressor in a health outcome model (Graziani, Guindani and Thall, 2015) with the notion of characterizing uncertainty in cluster assignment via credible balls (Wade et al., 2018). Thus, the CGLM is advantageous because it controls the uncertainty in cluster assignments and ultimately reduces misclassification error and potential bias in regression coefficients that would otherwise be present in an outcome model that uses a single point estimate of the cluster labels as a regressor. Naturally, the approach loses efficiency given the reduced sample size and has the potential to suffer from selection bias by using only the subjects with highest certainty in their class assignment. In our analysis, although the schools with higher uncertainty tended to have fewer outlets nearby, the excluded schools did not differ in terms of the outcome, thus minimizing selection bias concern upon conditioning by the number of FFRs in the second stage. Inverse probability of selection weights could be incorporated into the second stage when selection bias is a concern. Furthermore, while our modeling approach sought to estimate the independent health effects of the quantity of amenities vs. their proximity to study subjects, the sizes of the consensus clusters we identified did not permit us to examine interactions among these two factors. Estimating synergistic effects of quantity and proximity may be of interest in future studies with larger sample sizes. The BKMR is advantageous as it can use data from all subjects, handles cluster assignment uncertainty by using the co-clustering probability matrix, and can provide more granular information about health outcome risk for each subject in a school through the posterior estimates of $h(\mathbf{P}_i)$ instead of for discrete clusters. However, visualizing/interpreting the BKMR's rich set of output could be challenging. In our case, we compare the results of the analyses between the GCLM and BMKR methods and thus we used the mode cluster label to visualize the BKMR results. Other visualizations of the results may include displays of plots of the $\widehat{h(\boldsymbol{P}_i)}$ as a function of the L^2 norm of the co-clustering probabilities, $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ for a reference school j or visualizations of the estimated BEF normalized intensities among subgroups of subjects with similar values or $\widehat{h(\boldsymbol{P}_i)}$. Importantly, high values of $\widehat{h(\boldsymbol{P}_i)}$ imply higher outcome risk, thus these values can be used to identify subjects that may need further interventions.

We note several future research directions. First, as a comment on our application, future research could conduct this analysis on subject level data and avoid the possibility of introducing ecological bias, as our analysis of aggregate data does here. Methodologically, it is desirable to pursue joint estimation or other methods to more comprehensively propagate the uncertainty associated with the use of DP clustering results as an input in a health outcome analysis, as neither the BKMR or CGLM fully do so. Our current method is unable to easily embrace such a joint modeling approach due to both label switching and the varying number of assigned clusters across the MCMC iterations, yielding identifiability problems for the health outcomes models (Gelman et al., 2013). One possible solution is to incorporate the health outcome at the level at which the cluster is constructed by, for example, adapting the Logistic Stick Breaking Process (Ren et al., 2011). While the two-stage approach proposed here does not fully propagate uncertainty in cluster assignment in a standard fashion, defining exposure clusters independent of the outcome ensures a greater level of interpretability and conforms to substantive understanding of clusters being specifically about exposure. Furthermore, estimating clusters separately from an outcome means they can be used for more than one health outcome analysis. Second, we note our clustering objective focused on identifying meaningful patterns of FFR exposures and using these patterns as predictors of health outcomes, but not on explaining FFR prevalence. Thus, we do not model the distribution of FFRs as a function of related covariates. Researchers interested in identifying predictors of these environment clusters may find the work of Nylund-Gibson, Grimm and Masyn (2019) useful. Third, future researchers using NDP for clustering BEFs spatial distributions, or other uses, will need to be mindful of emerging theoretical results that cast some doubts on the quality of clustering results in some cases (Camerlenghi et al., 2019; Miller and Harrison, 2013, 2014, 2018). The specific causes of concern raised by these theoretical results may be less relevant when clustering BEF data and/or can be addressed by the dual use of CGLM and BMKR in the second stage. Finally, it is of interest to consider various extensions of the first stage model. One is to consider alternative modeling strategies for the point pattern of the fast food restaurants that allow for dependence among their locations that is not incorporated by the IPP. Another is to incorporate the spatial distribution of more than one type of BEF amenity and incorporate the spatial proximity of subjects to one another.

In summary, this work offers new tools to characterize the built environment in which humans live and approaches to use novel exposure summaries to disentangle the health effects of two dimensions of accessibility to amenities in the built environment. Disentangling the effect of proximity from that of availability of amenities may help informing decisions as to how new built environments may be constructed in the future to reduce health risks within community environments (e.g., zoning regulations for school neighborhoods), as well as to identify places where to target interventions to ameliorate risks associated with existing built environment characteristics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

This research was partially supported by NIH grants R01-HL131610 (PI: Sánchez) and R01-HL136718 (MPIs: Sanchez-Vaznaugh and Sánchez). We also thank the efforts of the anonymous reviewers and editors whose helpful comments greatly improved this paper.

REFERENCES

Auchincloss AH, Moore KA, Moore LV and Roux AVD (2012). Improving retrospective characterization of the food environment for a large region in the United States during a historic time period. Health & Place 18 1341–1347. [PubMed: 22883050]

Austin SB, Melly SJ, Sanchez BN, Patel A, Buka S and Gortmaker SL (2005). 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 1575–1581. [PubMed: 16118369]

Baek J, Sánchez BN, Berrocal VJ and Sanchez-Vaznaugh EV (2016). Distributed lag models: examining associations between the built environment and health. Epidemiology (Cambridge, Mass.) 27 116. [PubMed: 26414942]

Besser LM, Rodriguez DA, McDonald N, Kukull WA, Fitzpatrick AL, Rapp SR and Seeman T (2018). Neighborhood built environment and cognition in non-demented older adults: the multi-ethnic study of atherosclerosis. Social Science & Medicine 200 27–35. [PubMed: 29355828]

Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ and Coull BA (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16 493–508. [PubMed: 25532525]

Bojorquez I and Ojeda-Revah L (2018). Urban public parks and mental health in adult women: Mediating and moderating factors. International Journal of Social Psychiatry 64 637–646. [PubMed: 30112960]

Camerlenghi F, Dunson DB, Lijoi A, Prünster I, Rodríguez A et al. (2019). Latent nested nonparametric priors (with discussion). Bayesian Analysis 14 1303–1356. [PubMed: 35978607]

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A et al. (2016). Stan: A probabilistic programming language. Journal of Statistical Software 20 1–37.

Chiang S, Guindani M, Yeh HJ, Dewar S, Haneef Z, Stern JM and Vannucci M (2017). A hierarchical Bayesian model for the identification of PET markers associated to the prediction of surgical outcome after anterior temporal lobe resection. Frontiers in Neuroscience 11 669. [PubMed: 29259537]

Currie J, DellaVigna S, Moretti E and Pathania V (2010). The effect of fast food restaurants on obesity and weight gain. American Economic Journal: Economic Policy 2 32–63.

Davis B and Carpenter C (2009). Proximity of fast-food restaurants to schools and adolescent obesity. American Journal of Public Health 99 505–510. [PubMed: 19106421]

Diebolt J and Robert CP (1994). Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society: Series B (Methodological) 56 363–375.

Diggle PJ (2013). Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. CRC press.

Evenson KR, Jones SA, Holliday KM, Cohen DA and McKenzie TL (2016). Park characteristics, use, and physical activity: A review of studies using SOPARC (System for Observing Play and Recreation in Communities). Preventive Medicine 86 153–166. [PubMed: 26946365]

Friedman J, Hastie T and Tibshirani R (2001). The Elements of Statistical Learning 1. Springer series in statistics New York.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB (2013). Bayesian Data Analysis. Chapman and Hall/CRC.

Goodrich B, Gabry J, Ali I and Brilleman S (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.19.3.

Graziani R, Guindani M and Thall PF (2015). Bayesian nonparametric estimation of targeted agent effects on biomarker change to predict clinical outcome. Biometrics 71 188–197. [PubMed: 25319212]

Hartigan JA (1975). Clustering Algorithms. John Wiley & Sons, Inc.

Ishwaran H and James LF (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96 161–173.

MacEachern SN (2000). Dependent dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University 1–40.

MacEachern SN and Shen X (1999). Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior: Comment. Journal of the American Statistical Association 94 799–802.

McGuire S. (2012). Institute of Medicine (IOM) Early Childhood Obesity Prevention Policies. Washington, DC: The National Academies Press; 2011. Advances in Nutrition 3 56–57. [PubMed: 22332102]

Miller JW and Harrison MT (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In Advances in Neural Information Processing Systems 199–206.

Miller JW and Harrison MT (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. The Journal of Machine Learning Research 15 3333–3370.

Miller JW and Harrison MT (2018). Mixture models with a prior on the number of components. Journal of the American Statistical Association 113 340–356. [PubMed: 29983475]

Nguyen HD and McLachlan G (2019). On approximations via convolution-defined mixture models. Communications in Statistics-Theory and Methods 48 3945–3955.

Nylund-Gibson K, Grimm RP and Masyn KE (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. Structural Equation Modeling: A Multidisciplinary Journal 26 967–985.

Papastamoulis P. (2016). label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs. Journal of Statistical Software, Code Snippets 69 1–24.

Peterson A. (2020). bendr: Built Environment Nested Dirichlet Processes in R. R package version 0.1.0-alpha.

Peterson A, Berrocal V, Sanchez-Vaznaugh E and Sanchez B (2020). Supplementary material for "How Close and How Much?: Linking Health Outcomes to Built Environment Spatial Distributions".

Raftery AE and Lewis SM (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. Practical Markov Chain Monte Carlo 7 763–773.

Ren L, Du L, Carin L and Dunson D (2011). Logistic stick-breaking process. Journal of Machine Learning Research 12 203–239. [PubMed: 25258593]

Rodriguez A, Dunson DB and Gelfand AE (2008). The nested Dirichlet process. Journal of the American Statistical Association 103 1131–1154.

Rodríguez CE and Walker SG (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. Journal of Computational and Graphical Statistics 23 25–45.

Roof K and Oleru N (2008). Public health: Seattle and King County's push for the built environment. Journal of Environmental Health 71 24–27.

Sacks G, Swinburn B and Xuereb G (2012). Population-based approaches to childhood obesity prevention.

Sánchez BN, Sanchez-Vaznaugh EV, Uscilka A, Baek J and Zhang L (2012). Differential associations between the food environment near schools and childhood overweight across race/ethnicity, gender, and grade. American Journal of Epidemiology 175 1284–1293. [PubMed: 22510276]

Sanchez-Vaznaugh EV, Weverka A, Matsuzaki M and Sánchez BN (2019). Changes in fast food outlet availability near schools: Unequal patterns by income, race/ethnicity, and urbanicity. American Journal of Preventive Medicine 57 338–345. [PubMed: 31377084]

Skinner AC, Ravanbakht SN, Skelton JA, Perrin EM and Armstrong SC (2018). Prevalence of obesity and severe obesity in US children, 1999–2016. Pediatrics 141.

Stephens M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62 795–809.

R Core Team (2019). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.19.3.

Valeri L, Mazumdar MM, Bobb JF, Claus Henn B, Rodrigues E, Sharif OI, Kile ML, Quamruzzaman Q, Afroz S, Golam M et al. (2017). The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: evidence from rural Bangladesh. Environmental Health Perspectives 125 067015. [PubMed: 28669934]

Vehtari A, Gelman A and Gabry J (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27 1413–1432.

Wade S. (2015). mcclust.ext: Point estimation and credible balls for Bayesian cluster analysis R package version 1.0

Wade S, Ghahramani Z et al. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). Bayesian Analysis 13 559–626.

Wall MM and Liu X (2009). Spatial latent class analysis model for spatially distributed multivariate binary data. Computational Statistics & Data Analysis 53 3057–3069. [PubMed: 20161235]

Wall MM, Larson NI, Forsyth A, Van Riper DC, Graham DJ, Story MT and Neumark-Sztainer D (2012). Patterns of obesogenic neighborhood features and adolescent weight: a comparison of statistical approaches. American Journal of Preventive Medicine 42 e65–e75. [PubMed: 22516505]

Walls D. (2013). National establishment time-series (NETS) database: 2012 database description. Oakland: Walls & Associates.

Xiao S, Kottas A and Sansó B (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. The Annals of Applied Statistics 353–382.

Zhang H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. Journal of the American Statistical Association 99 250–261.

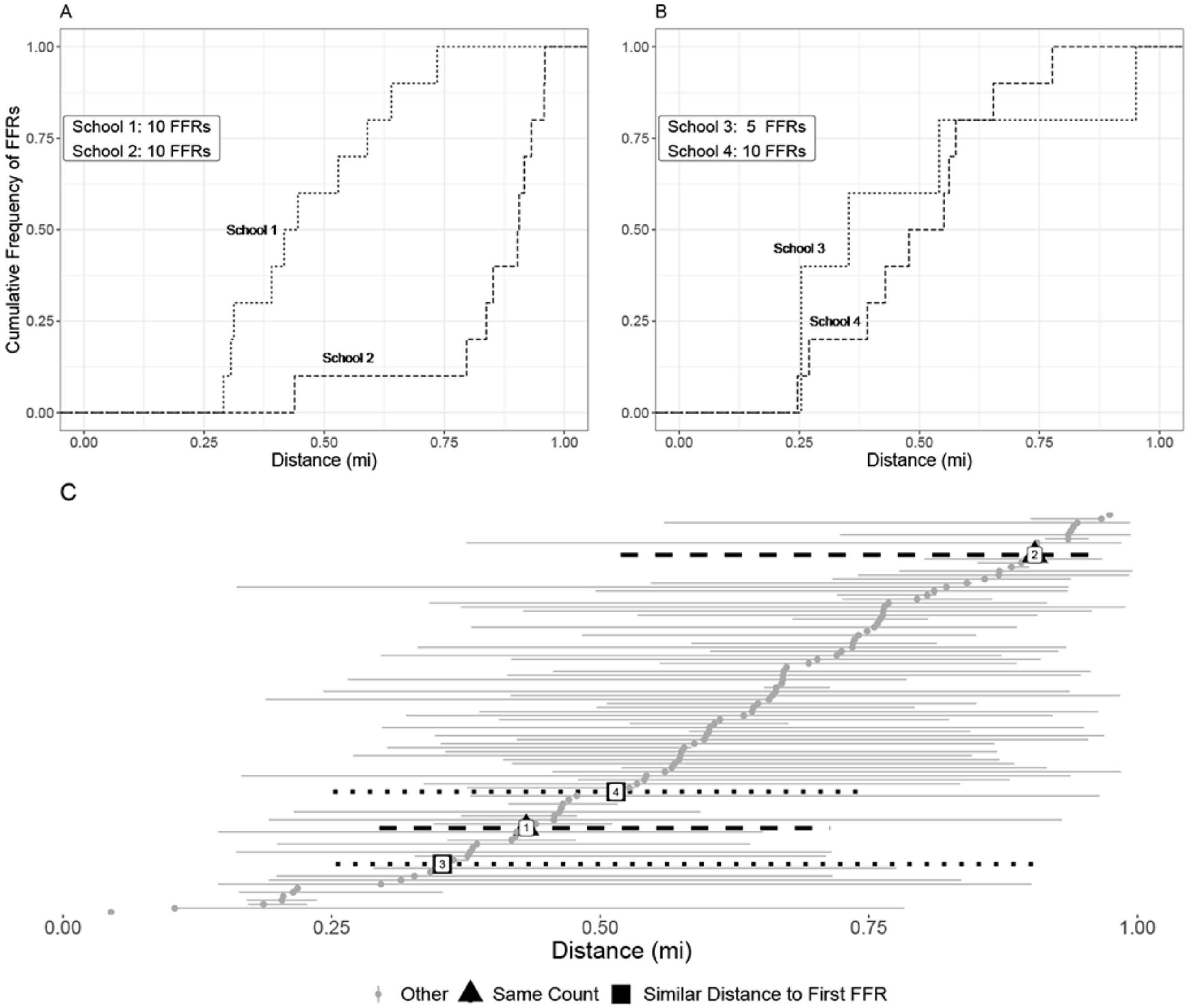

Fig 1.

Panel A: Distribution of distances from the school to nearby FFRs for two schools with 10 fast food restaurants (FFRs) within a 1 mile radius. Panel B: Distribution of distances from the school to nearby FFRs for two schools that have the same distance to the closest FFR. Panel C: distribution of distances to FFRs for a sample of 100 schools. For each school the plot shows the range of distances between the 2.5th and the 97.5th percentile. Schools are sorted by median distance to FFR. Darker dashed and dotted lines represent the four schools depicted in panels A and B of this figure.

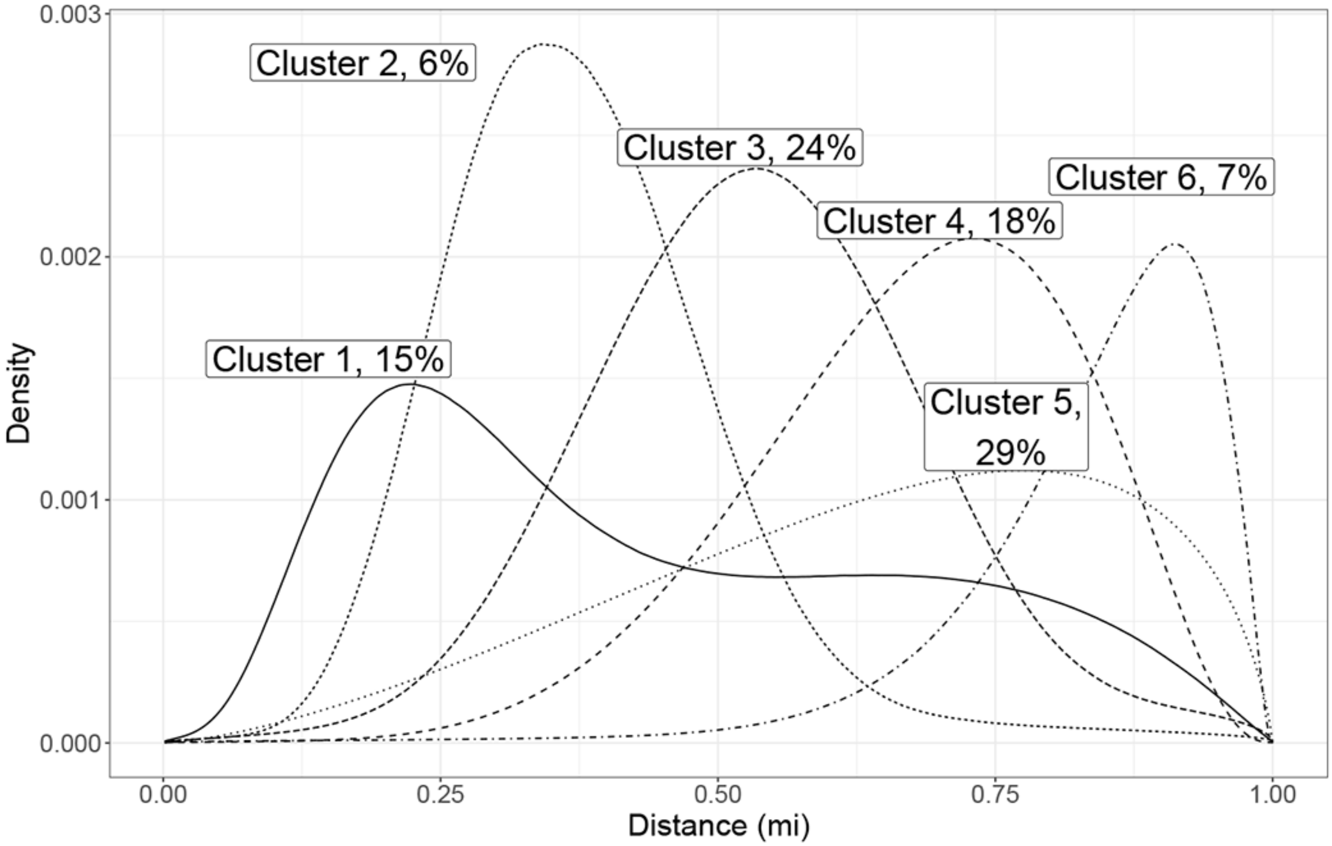

Fig 2.

Estimate of cluster density functions $f_k^*(r), k = 1, \ldots, 6$, with the estimated percent of schools within each cluster, π_k^*. The estimate here is taken to be the posterior median. The IQR for the percent of schools in each cluster are, for clusters 1 to 6, respectively: 3, 2, 4, 5, 5, and 2%

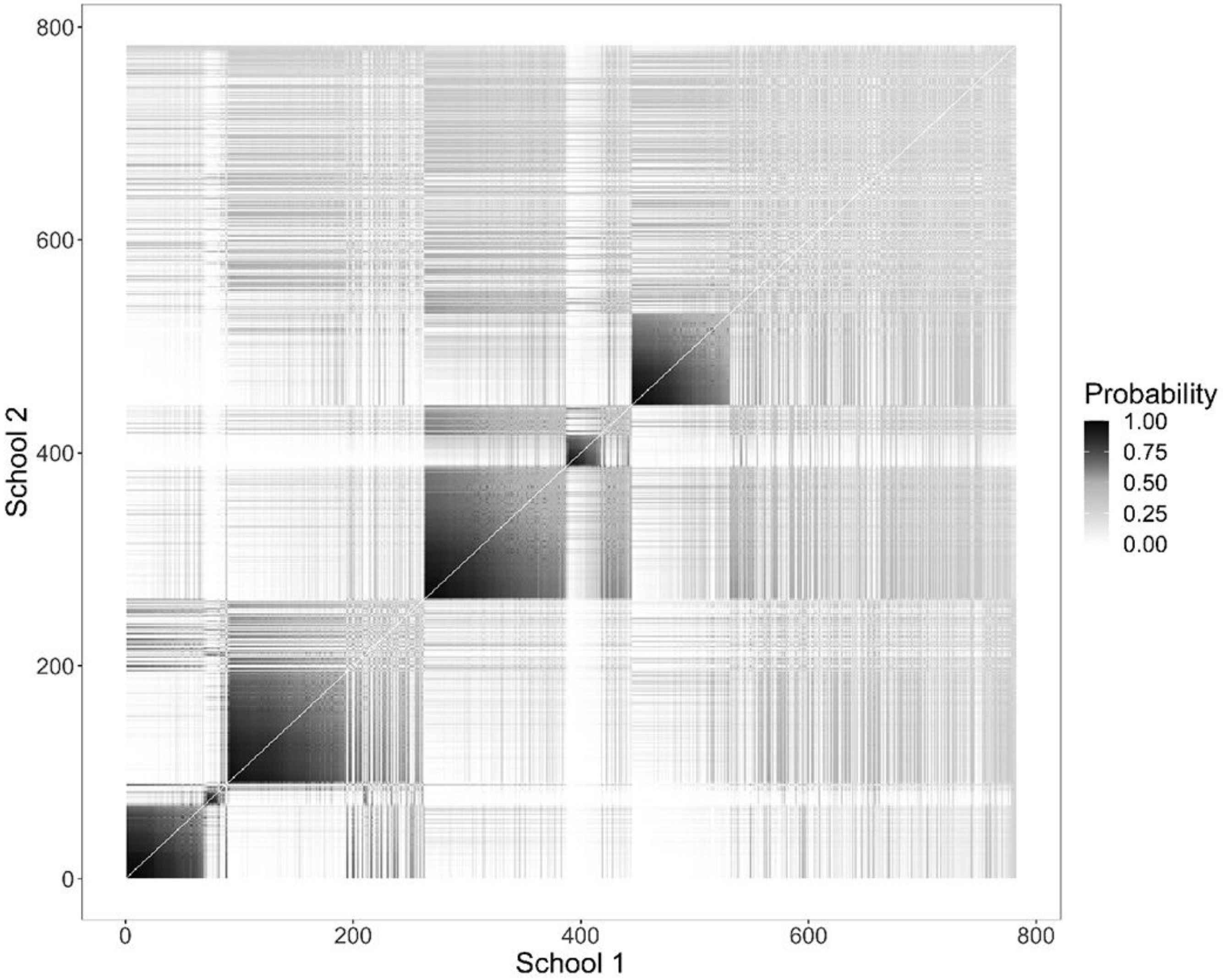

Fig 3.

Heat map of co-clustering probabilities, that is, the probability that any two schools are assigned to the same cluster. The identity line may be interpreted as a school's probability of being clustered with itself. Although this probability is trivially equal to 1, for plotting purposes, in the figure this line is left equal to 0 to more clearly show the plot's line of symmetry.

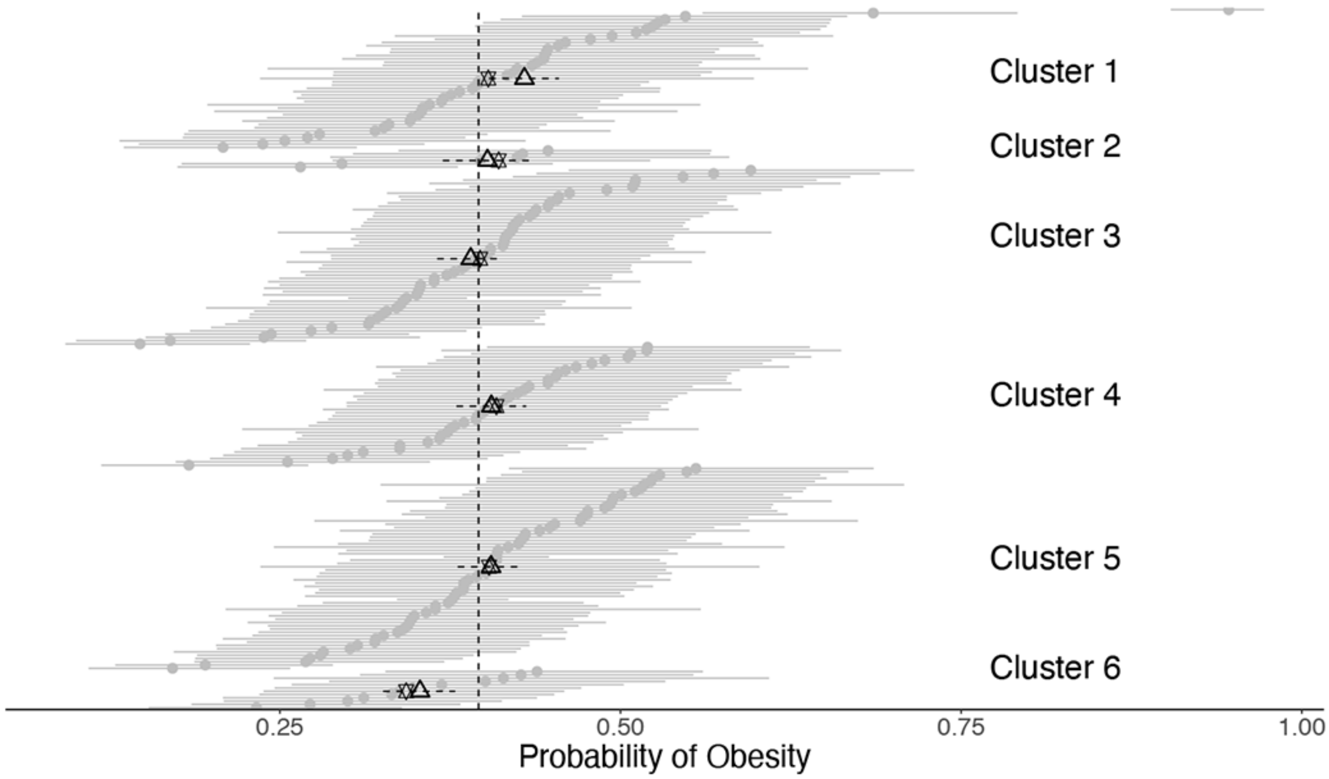

Fig 4.

Probability of obesity in relation to fast food restaurant (FFR) proximity. Estimates from the Bayesian Kernel Machine Regression (BKMR) are shown for each school (gray dot), along with 95 % credible intervals (gray line), and are sorted according to the cluster mode assignment; Cluster 1 is at the top of the figure, and Cluster 6 (FFRs furthest away) is at the bottom. The black stars represent the overall median probability of obesity for children attending schools in the given cluster. Triangles (and horizontal black dashed lines) denote the median posterior probability of obesity for children attending schools in each cluster estimated from the consensus GLM (CGLM) along with the 95% credible interval interval. The dotted vertical line is the posterior median probability of obesity when all adjusting covariates are 0 (that is, a majority White sub-urban high school with at least one FFR within a mile of the school’s location, and with the average percent of college educated adults and median census tract household income). BKMR and CGLM results are estimated using the consensus data set.

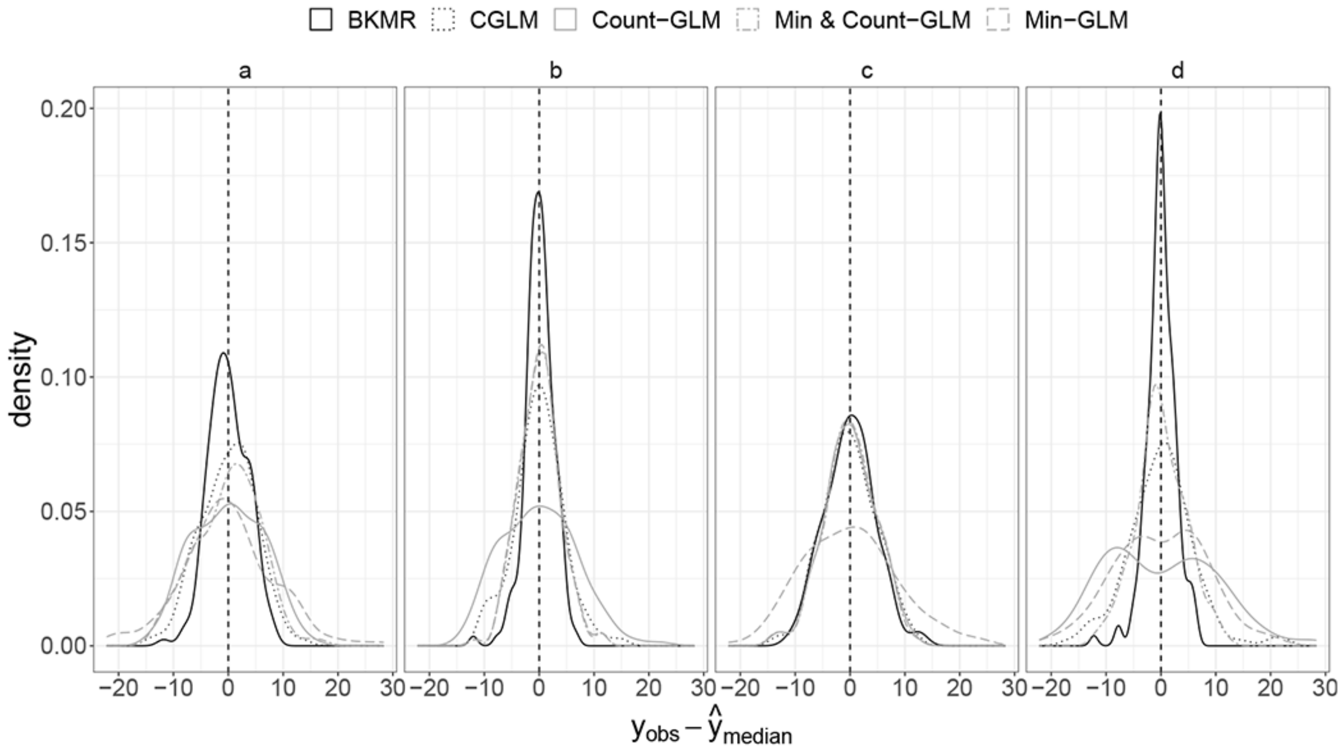

Fig 5.

Density estimate of residuals paneled by simulation scenario as described in the text. Black lines correspond to proposed models, gray lines correspond to competitor models. Line type and color defines the exact model used.

Table 1

Descriptive statistics of school characteristics by mode cluster assignment. For categorical predictors, the value shown is the percent of schools in the column that fall in a given category. The column designated as “Cluster 0” reports summary statistics for high schools without any fast food restaurants within one mile of their location. The income and education variables refer to characteristics of the population living in the census tract in which schools are located.

	Mode Cluster							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
N of Schools	**426**	**103**	**28**	**231**	**105**	**252**	**31**	**1,176**
FFR Quantity within 1 mile								
[1,4]	0	42	39	48	34	47	55	29
≥5	0	58	61	52	66	53	45	35
Zero	100	0	0	0	0	0	0	36
Urbanicity								
Rural	39	10	14	13	10	6	19	21
Sub-Urban	34	40	39	44	44	46	42	40
Urban	27	50	46	43	47	48	39	39
Majority Race/ethnicity among enrolled students								
African American	2	2	4	1	0	0	3	1
Asian	2	6	4	4	4	5	3	4
Hispanic	27	38	21	27	29	29	16	28
No Majority	9	11	14	13	18	13	19	12
White	59	44	57	55	50	52	58	55
Median Household Income (1,000 USD)								
Median	53.9	55.2	55.7	61.0	69.7	61.1	67.4	58.6
IQR	34.1	33.3	30.2	39.1	41.7	30.1	43.0	35.1
Proportion of adults with ≥ 16 years of education								
Median	24.9	25.0	25.4	25.4	25.6	25.3	25.5	25.2
IQR	2.1	2.7	3.4	2.6	2.9	2.6	2.5	2.4

Table 2

Descriptive statistics for schools included in the Consensus GLM vs. not. IQR = Inter-quartile range; FFR = Fast Food Restaurant. For categorical variables, the entries represent the percentage of schools in each column that fall in a given category. The income and education variables pertain to the population living in the census tract where schools are located.

	In Consensus	Not in Consensus	All[*]
Proportion Obese			
Median (Q1-Q3)	40.9 (33.3-47.4)	41.3 (34.1-48.2)	41.3 (33.9-48)
IQR	14	14.1	14.2
FFR Quantity within 1 mile			
[1,4]	21	55	45
≥5	79	45	55
Urbanicity			
Rural	8	11	10
Sub-Urban	35	47	44
Urban	57	42	46
Majority Race			
African American	1	1	1
Asian	5	4	4
Hispanic	30	28	29
No Majority	16	13	14
White	49	53	52
Median Income (1,000 USD)			
Median (Q1-Q3)	60.4 (43.2-78.4)	61.4 (46.2-82.9)	61.2 (45.3-81.5)
IQR	35.2	36.7	36.2
Proportion of residents with ≥ 16 years of education			
Median (Q1-Q3)	25.3 (24.3-26.7)	25.4 (24.2-27)	25.4 (24.2-26.9)
IQR	2.5	2.7	2.7

[*] N=17 schools were omitted from the health outcome models due to missing data on obesity.

Table 3

Widely Applicable Information Criterion (WAIC) and (out-of-sample Mean Square Error) for Bayesian Kernel Machine Regression (BKMR), and Consensus and Mode GLM (CGLM, MGLM), and Traditional models 1-3, for both Consensus and Full datasets corresponding to "In Consensus" and "All" columns from Table 3, respectively. Each model contains the same adjusting covariates but different measures of FFR exposure in a logistic regression modeling 9th grader obesity. "Count GLM" includes the number of FFR within 1 mile of the school. "Min-GLM" includes the distance to the closest FFR and "Min & Count-GLM" includes both measures.

	Dataset	
Models	**Full**	**Consensus**
Proposed		
BKMR	11,126.52 (0.014)	6,922.43 (0.011)
CGLM	-	9,883.43 (0.011)
MGLM	17,612.0 (0.009)	-
Traditional		
Min & Count-GLM	26,096.42 (0.043)	12,169.42 (0.040)
Min-GLM	33,972.94 (0.043)	17,040.08 (0.038)
Count-GLM	30,566.95 (0.035)	12,217.12 (0.039)