Received: 14 November 2020 | Revised: 20 December 2021 | Accepted: 6 January 2022

DOI: 10.1002/sim.9338

RESEARCH ARTICLE

Statistics in Medicine WILEY

Split and combine simulation extrapolation algorithm to correct geocoding coarsening of built environment exposures

Jung Y. Won[1] | **Emma V. Sanchez-Vaznaugh**[2] | **Yuqi Zhai**[1] | **Brisa N. Sánchez**[3]

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

[2]Department of Health Education, San Francisco State University, San Francisco, California

[3]Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, Pennsylvania

Correspondence
Brisa N. Sánchez, Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, PA, USA.
Email: bns48@drexel.edu

Funding information
NTH, Grant/Award Numbers: R01-HL131610, R01-HL136718

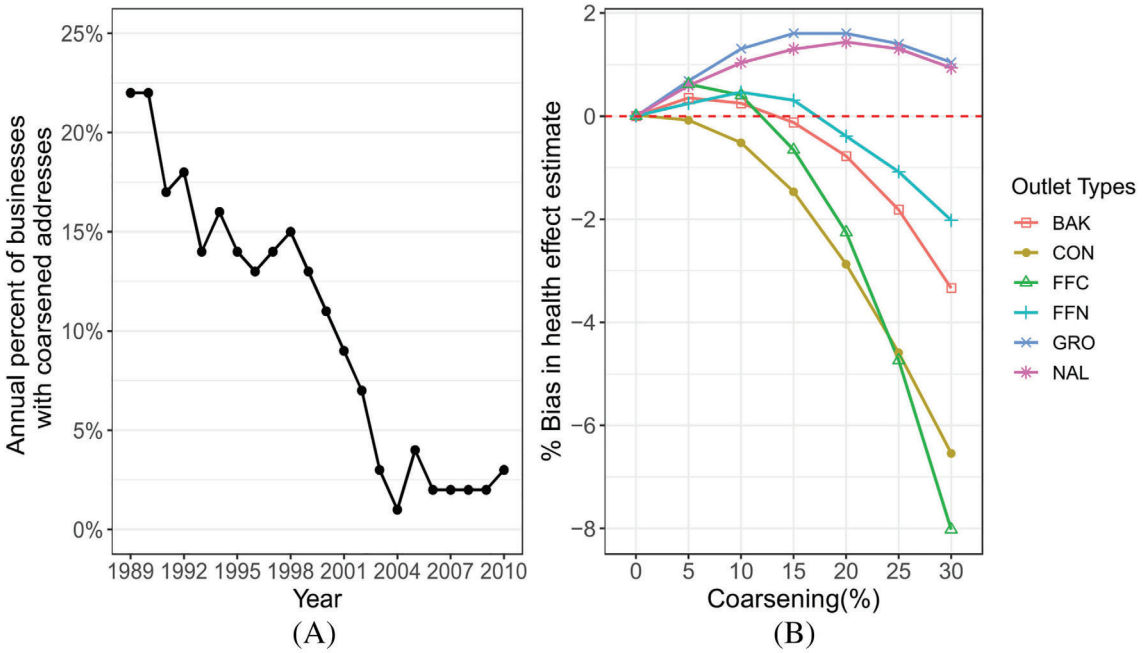
A major challenge in studies relating built environment features to health is measurement error in exposure due to geocoding errors. Faulty geocodes in built environment data introduce errors to exposure assessments that may induce bias in the corresponding health effect estimates. In this study, we examine the distribution of the measurement error in measures constructed from point-referenced exposures, quantify the extent of bias in exposure effect estimates due to geocode coarsening, and extend the simulation extrapolation (SIMEX) method to correct the bias. The motivating example focuses on the association between children's body mass index and exposure to the junk food environment, represented by the number of junk food outlets within a buffer area near their schools. We show, algebraically and through simulation studies, that coarsening of food outlet coordinates results in exposure measurement errors that have heterogeneous variance and nonzero mean, and that the resulting bias in the health effect can be away from the null. The proposed SC-SIMEX procedure accommodates the nonstandard measurement error distribution, without requiring external data, and provides the best bias correction compared to other SIMEX approaches.

KEYWORDS
bias, geocode coarsening, measurement error, simulation-extrapolation

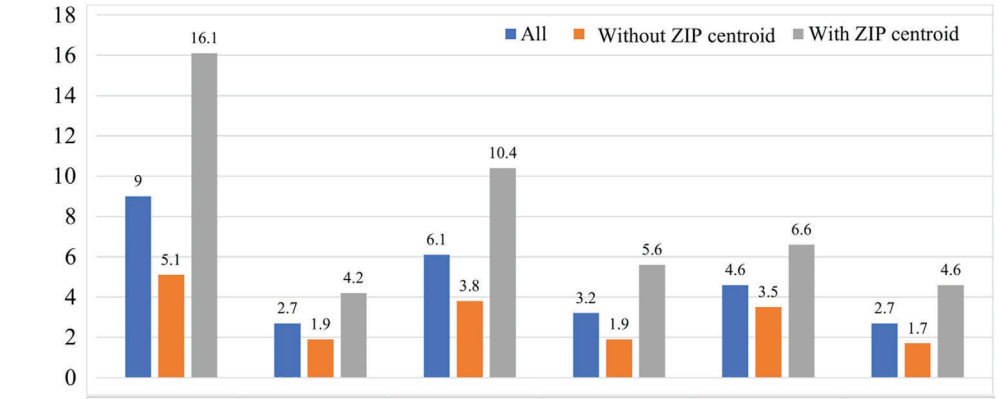
1 | INTRODUCTION

In built-environment health research, health outcomes are examined in relation to exposure to characteristics of the physical environment. Relatively recent research in this area focuses on whether the availability of specific amenities is detrimental or supportive to health. In the project motivating this paper and detailed in Section 2, a question is whether the availability of junk food outlets near schools influences child obesity. Amenities can be thought of as point-referenced built environment features, and accuracy in their recorded location is required to obtain correct exposure measures and estimate unbiased associations of interest. Large-scale databases of environmental features, such as listings of commercial establishments, are widely used to characterize exposure. However, such listings are known to have errors, including inaccurate geocodes.[1]

Research into the impact of geocoding errors is especially important given the prevalence of these errors and because the bias in health effect estimates arising from geocoding errors is not always toward the null. Figure 1A shows the percent of food outlet locations in a database of business establishments where instead of the exact latitude and longitude (lat-lon) of the business, the lat-lon of the centroid of the business ZIP code is recorded (ZIP centroid). Although accuracy has



Average number of food outlets within 1 mile circular buffer around California public schools in 2010, and total outlets state-wide



	Grocery (GRO)	Convenience Store (CON)	Bakery (BAK)	Non-Alcohol drinking place (NAL)	Fast food chain (FFC)	Fast food non-chain (FFN)
State-wide, J	14,297	4,989	11,132	6,206	9,110	4,893

(C)

FIGURE 1 (A) Annual percent of food retail outlets (standard industry codes 54 and 58) in California with coarsened addresses by year of entry into the National Establishment Time-Series (NETS) database; (B) Percent bias in health effect estimates from naïve regression depends on coarsening level and food outlet type. Bias is either positive or negative since measurement error induced by coarsening is not classical measurement error (see Section 4 for details about the simulation used to obtain this figure); (C) For each outlet type, the average number of food outlets is shown for schools overall (N = 12 370 schools), and by whether schools have at least one ZIP centroid within their buffers (N = 4348 schools)

improved over time or other technologies such as GPS could be used to characterize exposure, many longitudinal cohort studies, including MESA,[2,3] REGARDS,[4,5] CARDIA,[6] the Jackson Heart Study,[7] among others,[8] use historical records dating back to 1990 to quantify cumulative exposure. Based on simulations, Figure 1B shows that the bias in estimated health effects estimated via a naïve linear regression that uses the observed exposure in place of true exposure can be either toward or away from zero.

Prior research about the impact of geocoding accuracy on study results has been divided into two broad types according to the type of geocoding error. One type of error is called "positional error," where the latitude and/or longitude may be off

by an unknown distance of often a few meters. Another type is area-level "coarsening," where the centroid of a ZIP code (or other area-level units) is obtained as the output of the geocoding software, instead of the lat-lon of the address. Existing literature has explored the properties and impact of both of these geocoding errors on exposure measures.[9] Positional geocoding errors tend to be spatially autocorrelated and the magnitude of the error is related to urban-rural gradients and local street network characteristics[10,11] and can impact exposure measures.[12,13] For area-level coarsening, Zimmerman[1] developed methods to estimate the spatial intensity of a point process by applying the coarsened-data methodology. While geocoding errors that arise from automated (software-based) geocoding of addresses could potentially be avoided by using global positioning system (GPS) devices,[11,14] use of GPS has several limitations: it would incur additional costs; it is likely to be feasible only in a subset of locations, that is, a validation sample; and it is impossible for both historical datasets and those covering large geographical extents, like the cohorts described previously.

In this article, we focus on the estimation of the health effect of environmental exposures derived from point-referenced built environment features that are subject to area-level coarsening, which has not been previously studied. We first examine the measurement error structure in exposure counts associated with coarsened geocodes, both analytically and via simulation. We demonstrate that the measurement error leads to bias in the health effect estimates that can be away from or toward zero. Second, we propose a simulation-extrapolation (SIMEX) approach to correct bias in health effect estimates due to geocode coarsening. Our implementation of SIMEX relies on readily available indicators about the coarsening level of individual amenities. Although bias in the naive estimate can be positive or negative (Figure 1B) depending on the outlet, the bias of the naive estimator follows a predictable pattern as a function of the percent of businesses with coarsened addresses, making SIMEX feasible. Thus, our proposed method is a practical approach to correct for the spatial imprecision in coordinates because it does not require external validation data, nor does it impose distributional assumptions about the measurement error distribution or assumes a known measurement error variance.

Originally developed to correct the bias in the regression coefficient of a continuous regressor contaminated by zero-mean Gaussian measurement error,[15] the SIMEX method has now been extended in several directions. For example, MC-SIMEX can be applied to misclassification in a categorical response or categorical discrete regressors, or both. Before applying MC-SIMEX, misclassification probabilities need to be estimated from independent validation data.[16,17] Parveen and others relax the assumptions of the measurement error distribution to include a covariate-dependent measurement error with a nonzero mean.[18,19] However, they require that the measurement error variance be constant and known or that it be reliably estimated. Heteroskedasticity of the measurement error variance has also been studied for SIMEX, and obtaining a consistent SIMEX-based estimate was found to be difficult unless simulated errors in the simulation-step matched the true errors closely in distribution.[20] The SIMEX method has also been used in a spatial context to correct bias in the health effect of air pollution exposure.[21] The authors assumed the measurement error followed a multivariate Gaussian distribution with a zero mean and a spatial covariance approximated by external validation data.

In Section 2, we use simulations, derivations, and data related to the location of food outlets and public schools in the state of California to illustrate the structure of the measurement error induced by geocode coarsening. In Section 3, we introduce the notation and the newly proposed method. The results of simulations used to evaluate the proposed method are included in Section 4, followed by the application of the methodology to the motivating study data in Section 5. Final remarks are given in Section 6.

2 | FOOD ENVIRONMENT DATA AND MEASUREMENT ERROR STRUCTURE

2.1 | Food environment near California public schools

Several available data sources contain point-referenced data on commercial establishments.[22,23] One of these sources is the widely-used national establishment time-series (NETS) database[22] which contains information on over 44.2 million unique US establishments updated annually between 1990 and the present. The NETS database is a compilation of annual data sets originally collected by Dun and Bradstreet (Short Hills, NJ).[24] The information collected includes company names, addresses, sales, and Standard Industrial Classification Codes;[25] this information is used to categorize each entity (eg, restaurant vs convenience store).

Importantly, NETS also has the latitude and longitude of each establishment and a categorical variable that indicates the accuracy level of the coordinates (eg, whether the lat-lon is coarsened to the ZIP centroid or not). This

indicator variable is part of the default output of geocoding software and is thus available in any geocoded dataset. Changes in the geocoding practices of Dun and Bradstreet, as well as the historical nature of the data, has resulted in varying degrees of geocoding coarsening across time (Figure 1A). In this article, we use two years of NETS data for different purposes. For simulation experiments where the true geocodes are needed to produce error-free exposures (Sections 2-4), we use 2010 NETS data, which has very high geocoding quality. In analyses for our motivating study (Section 5), however, we use exposure measurements from 2000 NETS data, the year where child obesity data is also available.

In our motivating study, specific food retail outlets are of interest: small grocery stores (GRO), convenience stores (CON), bakeries (BAK), non-alcohol drinking places (NAL), fast food restaurant chains (FFC), and nonchain fast food restaurants (FFN). These food outlets are considered as a source of junk food more accessible for children, and their availability near schools can influence children's junk food consumption directly through purchasing or indirectly through advertising, and eventually influence childhood obesity. In 2010 in California, the most prevalent outlet type was small grocery stores, 14 297 out of 50 627 food outlets, followed by bakeries and fast food chain outlets (Figure 1C).

Data on school locations were downloaded from the California Department of Education (CDE) website. Per these data, there were 12 370 public schools open in California in 2010. As done in many built environment studies (eg, Athens et al[26]), we defined the neighborhood around a school as a one-mile radius circular buffer centered on each school location and obtained the number of each of the types of food outlets within the buffer of each school. Using this circular buffer around each school, Figure 1C shows the availability of food stores by outlet type in the school neighborhood. The relative prevalence of each of the food outlet types near schools follows the total distribution of the businesses in California.

2.2 | Exposure measure and measurement error distribution

2.2.1 | Exposure measure

For simplicity, first assume there is only one type of food outlet of interest, and coordinates for both subjects and businesses are fully accurate. Let $\mathcal{X}$ be an $N \times J$ matrix where N is the total number of subjects (eg, schools) and J is the number of food outlets. Define each element x_{ij} of $\mathcal{X}$ as a binary indicator for whether business j is inside the buffer around subject i. Then, the true exposure of the ith subject, denoted by X_i, is the sum of indicators x_{ij}: $X_i = \sum_{j=1}^{J} x_{ij}$.

2.2.2 | Geocode coarsening

The geocoding accuracy level for each establishment is given as part of the automated geocoding process. In the NETS data, the coarsest geocode is given as the lat-lon of the ZIP centroid for the business, instead of the lat-lon of its street address. Let A_j be an accuracy indicator for the jth establishment, $j = 1, \ldots, J$, which takes a value 1 if the jth business's geocode is accurate to the street address level and 0 if the geocode is coarsened. Denote the observed proportion of coarsened addresses as $p_0 = \sum_{j=1}^{J} I(A_j = 0)/J$. Values of p_0 for NETS are given in Figure 1A. Naturally, coarsened geocodes may introduce exposure measurement error.

2.2.3 | Measurement error

To describe how the observed exposure is measured and define the measurement error, let $\mathcal{W}(p)$ be an observed version of $\mathcal{X}$ when $100p\%$ of businesses have coarsened addresses. $\mathcal{W}(p)$ has entries w_{ij} that indicate if the observed (possibly coarsened) business geocode is within the buffer for subject i. The observed exposure for subject i is then computed as $W_i = \sum_{j=1}^{J} w_{ij}$, and the measurement error is $U_i = W_i - X_i$. The measurement error is a count that, as shown below and further detailed in the Appendix Section A.1, follows a mixture distribution and violates all classical measurement error assumptions such as a zero mean, constant variance, and independence from the true exposure.

To more rigorously describe the properties of the measurement error U_i and enable us to define an approach to correct the resulting bias, for the jth business in the neighborhood of the ith subject, we define indicators of whether its corresponding w_{ij} matches x_{ij}:

$$I_{ij}^{11} = \mathbb{1}\{x_{ij} = 1, w_{ij} = 1\}, \qquad I_{ij}^{00} = \mathbb{1}\{x_{ij} = 0, w_{ij} = 0\},$$
$$I_{ij}^{10} = \mathbb{1}\{x_{ij} = 1, w_{ij} = 0\}, \qquad I_{ij}^{01} = \mathbb{1}\{x_{ij} = 0, w_{ij} = 1\}.$$

When the jth business's address has been accurately geocoded, that is, $A_j = 1$, then the observed w_{ij} will match x_{ij}, and thus either I_{ij}^{11} or I_{ij}^{00} will be 1, depending on the location of the school and business. However, when the business j has been coarsened ($A_j = 0$), then the *possibility* arises that $I_{ij}^{01} = 1$ (when without coarsening $I_{ij}^{00} = 1$), leading to the jth outlet being incorrectly included in W_i. Similarly, if without coarsening $I_{ij}^{11} = 1$, then the possibility arises that $I_{ij}^{10} = 1$ with coarsening. Whether or not I_{ij}^{10} or I_{ij}^{01} is 1 when the address is coarsened depends on the location of the ZIP centroid for the ZIP Code corresponding to the jth business. The indicator I_{ij}^{00} may still be 1 if business j and its corresponding ZIP centroid both lie outside the buffer of subject i. Likewise, I_{ij}^{11} may still be 1 if the ZIP centroid corresponding to business j is located inside the ith subject's buffer.

Utilizing these indicators, we can write the true and observed exposure as $X_i = \sum_{j=1}^{J} x_{ij} = \sum_{j=1}^{J}(I_{ij}^{10} + I_{ij}^{11})$ and $W_i = \sum_{j=1}^{J} w_{ij} = \sum_{j=1}^{J}(I_{ij}^{01} + I_{ij}^{11})$. Hence $U_i = W_i - X_i = \sum_{j=1}^{J}(I_{ij}^{01} - I_{ij}^{10})$ is a mixture of binary indicators. As discussed in the Appendix Section A.1, upon conditioning on the locations of schools, true addresses of businesses, the locations of ZIP centroids, and under the assumption of coarsening at random (ie, $P(A_j = 0) = p$), the measurement error has $E(U_i) = (|S_i^{01}| - |S_i^{10}|)p$ and $Var(U_i) = (|S_i^{01}| + |S_i^{10}|)p(1-p)$, where S_i^{01} is a set of businesses that, when coarsened, will be incorrectly included in the exposure count for subject i, and S_i^{10} is a set of businesses that will be incorrectly excluded from the exposure count for subject i when coarsened, and $|\cdot|$ denotes set cardinality. In general, the cardinality of these sets is unknown since the true business locations are not known for all businesses, thus the mean and variance of the measurement error are unknown.

2.2.4 | Measurement error distribution depends on the location of ZIP centroids near the subject

The measurement error U_i can also be seen as the difference between the total number of businesses that newly "enter" the buffer, $\sum_{j=1}^{J} I_{ij}^{01}$, and the number of businesses that are "excluded", $\sum_{j=1}^{J} I_{ij}^{10}$, due to the coarsening. This difference depends on whether or not the subject's buffer area contains any ZIP centroids. When there are no ZIP centroids within the subject's buffer, businesses cannot enter the buffer due to coarsening, that is, $\sum_{j=1}^{J}(I_{ij}^{01}) = 0$, as depicted in Appendix Web Figure 1a. Thus, the measurement error for subjects without ZIP centroids within their buffers will be nonpositive. Otherwise, when ≥ 1 ZIP centroids are located inside the buffer, there are more possibilities, depending on whether ZIP Code boundaries are cut across the school buffer. The count can increase, decrease or stay the same (Appendix Web Figure 1b,c). Note also that multiple ZIP centroids can lie in the school buffer when the school is located near ZIP code boundaries, or when some ZIP Code polygons are smaller than the buffer as may occur in dense urban areas.

Therefore, the measurement error can be formulated according to the presence of ZIP centroids in the subject's buffer:

$$U_i = W_i - X_i = \begin{cases} \sum_{j=1}^{J}(-I_{ij}^{10}) \leq 0, & \text{no ZIP centroid within subject's buffer} \\ \sum_{j=1}^{J}(I_{ij}^{01} - I_{ij}^{10}), & \text{ZIP centroid within the subject's buffer.} \end{cases} \tag{1}$$

From (1), regardless of a subject's location, U_i violates the classical measurement error assumption that the measurement error is independent of the true value X_i. This is shown analytically in Appendix Section A.1 and illustrated empirically in Table 1. In addition, $Corr(U_i, W_i) \neq 0$ thus the error is not Berkson error either. Moreover, the measurement errors U_i and U_r for $i \neq r$ are also correlated when subjects are geographically close to each other (see technical arguments in Appendix Section A.1).

In summary, the measurement error due to coarsening is a mixture of binomial components; errors are spatially correlated across subjects and are correlated with both the observed and true exposures. The measurement error

TABLE 1 Average of variances of measurement errors U_{ik}, observed exposures, W_i, and true exposures, X_{ik}, for k = GRO, CON, BAK, NAL, FFC, FFN, and correlations them, in the presence of 15% of geocode coarsening for each business type

		Corr(U_k, $U_{k'}$)					
Food outlet type	**Var(U_k)**	**GRO**	**CON**	**BAK**	**NAL**	**FFC**	**FFN**
Grocery (GRO)	3.66	1	0.41	0.50	0.38	0.40	0.38
Convenience store (CON)	0.61	0.41	1	0.37	0.27	0.36	0.29
Bakery (BAK)	1.90	0.50	0.37	1	0.42	0.46	0.40
Non-alcohol drinking place (NAL)	0.78	0.38	0.27	0.42	1	0.37	0.35
Fast food chain (FFC)	1.45	0.40	0.36	0.46	0.37	1	0.36
Fast food nonchain (FFN)	0.59	0.38	0.29	0.40	0.35	0.36	1
		Corr(W_k, $U_{k'}$)					
	Var(W_k)	**GRO**	**CON**	**BAK**	**NAL**	**FFC**	**FFN**
Grocery (GRO)	209.76	−0.12	−0.03	−0.05	−0.08	0.00	−0.05
Convenience store (CON)	10.21	−0.08	0.06	−0.01	−0.05	0.01	−0.03
Bakery (BAK)	60.98	−0.08	0.01	0.00	−0.07	0.01	−0.03
Non-alcohol drinking place (NAL)	37.59	−0.14	−0.03	−0.07	−0.09	−0.02	−0.07
Fast food chain (FFC)	20.56	−0.05	0.01	−0.02	−0.04	0.02	−0.01
Fast food nonchain (FFN)	16.62	−0.13	−0.02	−0.06	−0.10	−0.01	−0.01
		Corr(W_k, $U_{k'}$)					
	Var(W_k)	**GRO**	**CON**	**BAK**	**NAL**	**FFC**	**FFN**
Grocery (GRO)	220.14	−0.25	−0.08	−0.11	−0.13	−0.05	−0.10
Convenience store (CON)	10.52	−0.18	−0.18	−0.10	−0.12	−0.08	−0.10
Bakery (BAK)	62.71	−0.17	−0.05	−0.17	−0.15	−0.07	−0.10
Non-alcohol drinking place (NAL)	39.36	−0.19	−0.06	−0.13	−0.23	−0.07	−0.12
Fast food chain (FFC)	21.76	−0.15	−0.08	−0.14	−0.14	−0.24	−0.11
Fast food nonchain (FFN)	17.31	−0.20	−0.07	−0.14	−0.16	−0.07	−0.19

Note: Values are based on 1000 Monte Carlo simulations, among N = 1000 randomly selected schools.

distribution depends on whether the subject's buffer has a ZIP centroid within it and this complex measurement error structure violates both the classical and Berkson measurement error model assumptions.

2.2.5 | Multivariate measurement error distribution: multiple outlet types

The extension to multiple types of food outlets is straightforward. Following the notation above, suppose there are Q different types of food outlets (eg, fast food restaurants, convenience stores). Assuming each business type has m_q establishments for q=1, … , Q, the grand total number of outlets is $J = \sum_{q=1}^{Q} m_q$. Then, the ith subject's true exposure to the qth food outlet type is $X_i^q = \sum_{j=1}^{J}(I_{ij}^{10} + I_{ij}^{11}) * I(\text{business } j \text{ is of type } q)$, while $W_{iq} = \sum_{j=1}^{J}(I_{ij}^{01} + I_{ij}^{11}) * I(\text{business } j \text{ is of type } q)$. Therefore, the measurement error for the count of the qth business type for subject i is $U_{iq} = W_{iq} - X_{iq} = \sum_{j=1}^{J}(I_{ij}^{10} - I_{ij}^{10}) * I(\text{business } j \text{ is of type } q)$.

It is possible that a correlation exists between measurement errors of different business types (eg, measurement error of exposure to fast food chains is correlated with a measurement error of exposure to groceries). This relationship is due to the potential similarity of business locations, as food outlets of different types are sometimes clustered. Hence, since true exposures are correlated, measurement errors will be as well (Table 1). The correlations among the errors in multivariate exposures result in either negative or positive bias in regression coefficients, even in the case of classical measurement error.[27] Similarly, the pattern of bias may also be different in our case when a single outlet type is considered

in a regression model compared to when multivariate exposures are considered (compare Figure 1B and Appendix Web Figure 2).

Carroll et al[27] emphasize that the error distribution is a key factor that determines the effects of measurement error on coefficient estimation. In our case, the measurement error due to geocode coarsening can have different distributions not only between subjects, as shown in Equation (1), but also across business categories. The measurement error distribution depends on more complex factors such as the buffer radius and the spatial distribution of schools and/or businesses. The following subsection empirically shows the distribution of the measurement error through simulations.

2.3 | Empirical illustration of measurement error distribution

We illustrate the distribution of the measurement error through simulations. In order to make the simulations as realistic as possible, we base the simulations on the actual spatial locations of schools and food outlets in California in 2010. While using the actual spatial locations of schools and food outlets has the significant advantage of making the simulation results more realistic, a small disadvantage is that there is a small percentage of businesses that have coarsened geocodes. For California in 2010, 3% of all food-related businesses have coarsened geocodes (Figure 1A), whereas there were 1.5% (745 businesses) with coarsened geocodes among the six business types of interest. Hence, in order to be able to calculate the "true" exposure for all schools in the simulations, we remove the businesses (1.5%) with coarsened geocodes (ie, with $A_j = 0$). The true geocodes of the remaining 49 882 outlets are used to calculate $\mathcal{X}$ and the true exposure X_i around each of the 12 370 schools. Given (1) and that Figure 1C shows that schools that have one or more ZIP centroids within their one-mile buffers have more food outlets near them, we examine measurement errors for schools with and without ZIP centroids in their buffers.

To create samples of the measurement errors U_i, we first create $\mathcal{W}$ and then calculate the observed exposure W_i, by replacing a proportion p_0 of accurate business coordinates with their ZIP centroid coordinates. For each subject, the error is then computed as the difference $W_i - X_i$. To investigate how the measurement error distribution and resultant bias in the health outcome regression depend on the initial coarsening proportion, p_0, we vary it from $p_0 = 0\%$ to 30%, in 5% increments. For each proportion p_0, businesses were coarsened-at-random. This process is repeated 50 times to stabilize the Monte Carlo error.

Figure 2A,B illustrates the overall mean of the measurement errors from the simulations, showing that it differs according to whether the school has at least one ZIP centroid nearby. As the proportion of coarsened businesses increases, the overall mean of the measurement error for schools with at least one ZIP centroid in their neighborhoods also increases, reflecting an "overcount" compared to the truth (ie, averaging $E(U_i)$ among these schools). On the other hand, among schools with no ZIP centroid nearby, the overall mean of the measurement error is negative, with an increase in magnitude as more geocodes are contaminated–that is, an under-count. Appendix Web Figure 3 shows histograms of $E(U_i)$ for schools with and without ZIP centroids within their buffers.

Figure 2C,D shows the measurement error variance averaged across schools. As expected, the variance of the measurement error also increases as the proportion of coarsened businesses increases. This is an important finding because the SIMEX procedure requires having an approach to replicate measurement error with an increasing magnitude of measurement error variance. Since schools that have a ZIP centroid in their neighborhood also have a higher number of businesses nearby (Figure 1C), the variance of the measurement error is also higher among that subset of schools (Figure 2D).

Measurement errors also have different distributions by food store type, and the errors are correlated among different food store types. Figure 2 and Table 1 show that the absolute magnitude of the measurement error means and variances follow the prevalence of food store types. Groceries have the largest error variance, bakeries have the second largest, and the least variance is for nonchain fast food restaurants. Table 1 also illustrates the previously described positive correlation among measurement errors across different food store types, as well as nonzero covariance among the measurement errors and the observed counts.

The question might arise as to whether the relationship between the presence or absence of ZIP centroid and the measurement error distribution could be explained by factors that increase the likelihood that a ZIP centroid is in the subject's buffer and also related to the overall true exposure, however, this is not the case. For instance, urbanicity is arguably one of the most potent predictors of both of these. There is both a greater availability of food outlets in urban areas and a greater number of urban schools are likely to have ZIP centroid(s) in their one-mile buffer due to the smaller size of

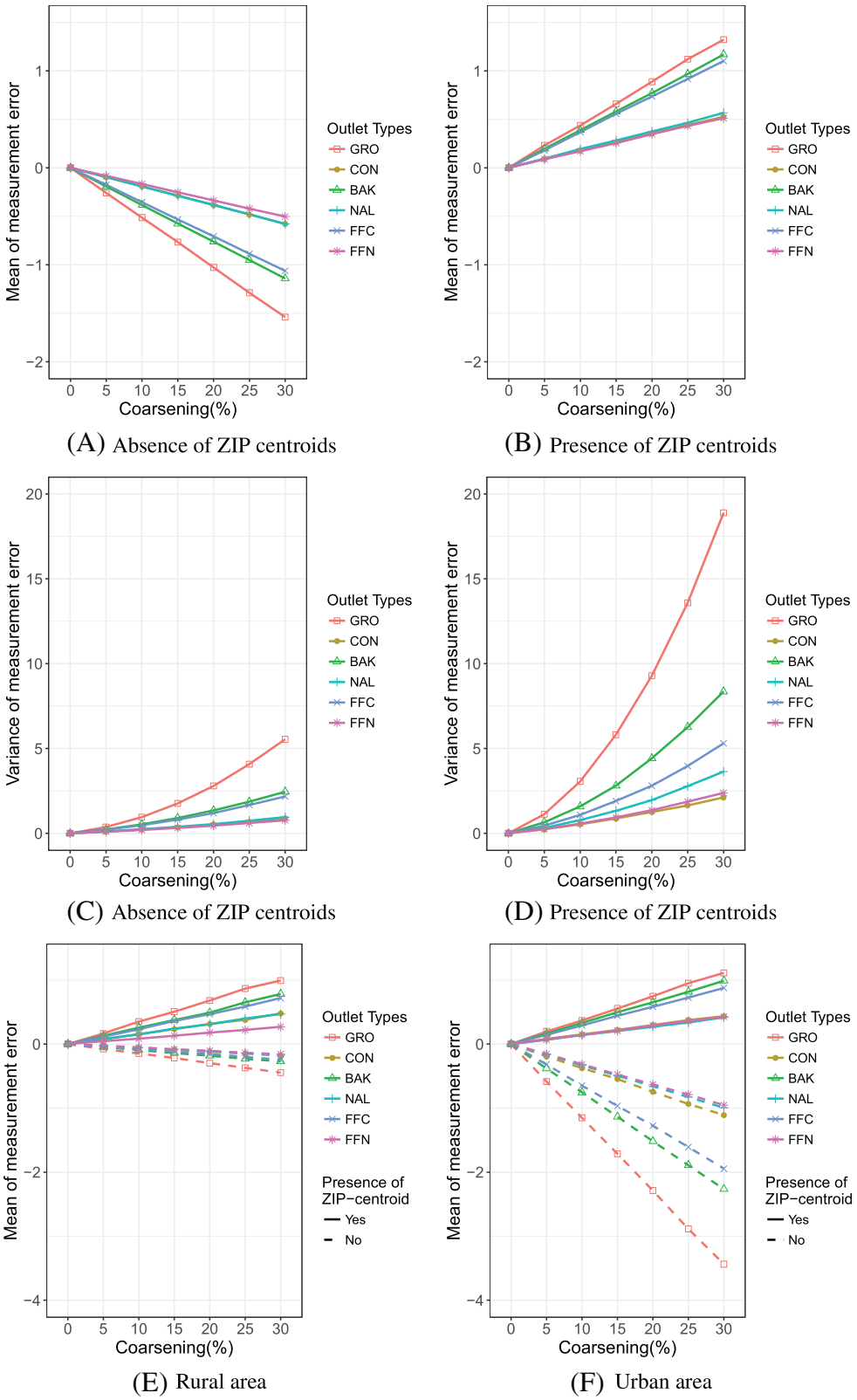


FIGURE 2 (A-D) Overall mean and variance of measurement errors by presence or absence of ZIP centroid within 1 mile of school, and (E,F) mean of measurement errors stratified by urbanization level. See Figure 1C for outlet abbreviations

urban ZIP Codes. However, the diverging pattern of the measurement error distribution between absence and presence of the ZIP centroid in the school buffer does not vanish even if we stratify the schools by whether they are in an urban area or not (Figure 2E,F). Thus, although there may be other factors that impact the absolute magnitude of the mean and variance of the measurement errors, such as urbanicity, the direction of the mean and variance of the measurement error are driven by the presence or absence of ZIP centroid in the buffer.

3 | SPLIT-COMBINE SIMEX

3.1 | Notation and assumptions

Let $\boldsymbol{Y}$ be a vector of continuous outcomes for N subjects, which can be modeled with linear regression: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{z}} + \boldsymbol{\epsilon}$. The $N \times Q$ exposure matrix $\boldsymbol{X}$ is a matrix of true unmeasured exposure, where the qth column of $\boldsymbol{X}$ denotes subject's exposures to the qth environment feature, and X_{iq} is as described in Section 2.2. The $\boldsymbol{Z}$ is a matrix of error-free covariates including an intercept and a corresponding parameter is $\boldsymbol{\beta}_{\boldsymbol{z}}$. The residual $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ is a vector of random errors with mean zero and variance σ^2_ϵ. We define $\boldsymbol{\beta} = (\boldsymbol{\beta}^T_{\boldsymbol{x}}, \boldsymbol{\beta}^T_{\boldsymbol{z}})^T$ and our primary interest is estimating $\boldsymbol{\beta}_{\boldsymbol{x}}$, the association between a continuous health outcome $\boldsymbol{Y}$ and unobserved count exposure $\boldsymbol{X}$.

In the presence of geocoding coarsening, we observe $\mathcal{W}$ instead of $\mathcal{X}$, and thus the analysis dataset will have observe data $(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{A})$ instead of true data $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$. Here $\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{U}$ with measurement error $\boldsymbol{U}$, and $\boldsymbol{A} = (A_1, \ldots, A_J)^T$, where A_j denotes whether the geocode for food outlet j has been coarsened, as described in Section 2. We propose a method to correct a bias in the estimation of $\boldsymbol{\beta}_{\boldsymbol{x}}$ due to measurement error $\boldsymbol{U}$ incurred by geocode coarsening. We briefly describe the classic SIMEX method to enable us to draw analogies in the description of our proposed method.

3.2 | Brief review of classic SIMEX

The simulation-extrapolation (SIMEX) procedure consists of two steps:[15] First, in the simulation step, a computer-simulated error that mimics the measurement error is added to the observed $\{W_i\}^n_{i=1}$, as follows. A grid of constants λ_s, $s = 1, \ldots, S$ is chosen (typically $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_s = 2$,[15]) and used to progressively magnify the variance of the computer-simulated error. Let U_{ib} represent the simulated error, then the simulated exposure is $W_{i,b}(\lambda_s) = W_i + \sqrt{\lambda_s} U_{ib}$. U_{ib} is assumed to follow the same distribution as the original measurement error distribution $N(0, \sigma^2_u)$, where σ^2_u is assumed to be a known and finite variance. For a given value of λ_s, $s = 1, \ldots, S$, the simulation is repeated $b = 1, \ldots, B$ times to reduce sampling variability. For each λ_s and each b, the model of interest is estimated using $W_{i,b}$ as the exposure. For each λ_s, the estimated regression coefficients are averaged across $b = 1, \ldots, B$. Second, in the extrapolation step, a trend between the averaged estimated regression coefficients and the values of λ_s is fitted. The fitted trend is used to predict the estimator for the case with no measurement error ($\lambda = -1$).

In the classic SIMEX method, the classical measurement error assumption enables the expectation of simulated covariate $W_{i,b}$ to be the same as the expectation of W_i, and the assumption of known σ^2_u enables the simulation of $W_{i,b}$. Theoretically, however, the SIMEX method is not limited to a known measurement error distributions. If we can replicate the measurement error process, we can still apply SIMEX to correct bias in coefficient estimates.[27]

3.3 | Proposed split-combine SIMEX method (SC-SIMEX) for geocode coarsened data

Unlike assumptions used in the traditional SIMEX procedure, the measurement error induced by geocode coarsening is a count that does not adhere to either classical or Berkson measurement error assumptions. However, given available indicators A_j, we can replicate the geocoding coarsening process in order to generate exposure measures that have progressively larger measurement error. In contrast to typical SIMEX that simulates pseudo-errors from the assumed measurement error distribution, we instead directly generate $\mathcal{W}(p)$ by simulating additional coarsening of individual business addresses and use it to obtain more contaminated $\boldsymbol{W}$. After this simulation step, we stratify the subjects into two groups, defined by the presence of one or more ZIP centroids in the subjects buffers, and run separate estimation steps. This is because, as was shown in the prior section, the measurement error distribution depends on the presence of ZIP centroids within the buffer of the subject, and in turn it has a different impact on the bias of the estimated coefficients (Appendix

Section A.2 and Section 4). Then, separate extrapolation steps produce two SIMEX estimates, $\widehat{\boldsymbol{\beta}}_{SIMEX,1}$ and $\widehat{\boldsymbol{\beta}}_{SIMEX,2}$, one for each strata, that are then combined to obtain the final estimates. Although stratification does not make measurement errors satisfy classical or Berkson error assumptions, within strata the naïve estimates have a less complex and more predictable pattern when plotted against the coarsening proportion facilitating the extrapolation step.

3.3.1 | Simulation step

We use the proportion (p_s) of the coarsened businesses in place of λ_s in the traditional SIMEX approach. We assume $p_0 \times 100\%$ of each business type have coordinates coarsened to their ZIP centroids, where $p_0 = \sum_{j=1}^{J} I(A_j = 0)/J$ is the proportion of coarsened geocodes in the observed data. Observed exposures for the ith school $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iQ})$ are computed with these data, using the observed matrix of indicators $\mathcal{W}(p_0)$ as described in Section 2.2. In the simulation step, an independent coarsening procedure replaces accurate geocodes with coarsened geocodes for an additional $(p_s - p_0)\%$ businesses, for a total of $p_s\%$ coarsened business addresses. The resulting simulated matrix of indicators $\mathcal{W}(p_s)$, is used to obtain the pseudo-covariates $\boldsymbol{W}_{i,b}(p_s) = (W_{i1}(p_s), \ldots, W_{iQ}(p_s))$ at coarsening level p_s. This preserves the spatial dependence of measurement errors that may arise, for example, when a coarsened business drops out of the buffer for one school and spuriously enters that of another. This process is repeated for $s = 1, \ldots, S; 0 < p_0 < p_1 < \cdots < p_S$, where S denotes the number of coarsening levels to be simulated, and for $b = 1, \ldots, B$ replications within each level s to reduce Monte Carlo error. By increasing the number of coarsened geocodes, $\boldsymbol{W}_{i,b}(p_s)$ will accordingly have greater amount of measurement errors. After obtaining $\boldsymbol{W}_{i,b}(p_s)$, the observations (eg, schools) are stratified according to presence or absence of a ZIP centroid in the subject's buffer for strata-specific estimation and extrapolation steps. For simplicity of notation, we drop the index for stratum. For each b and s, we fit the model $\boldsymbol{Y} = \boldsymbol{W}_b(p_s)\boldsymbol{\beta}_x + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{z}} + \boldsymbol{\epsilon}$ within each stratum. The estimate of $\boldsymbol{\beta}_x$ using contaminated exposure $\boldsymbol{W}_b(p_s)$ is denoted by $\widehat{\boldsymbol{\beta}}_{xb}(p_s)$. In each stratum, we compute the average of the estimates from the B simulated pseudo-data for each proportion p_s, $\widehat{\boldsymbol{\beta}}(p_s) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}_b(p_s)$, where $\widehat{\boldsymbol{\beta}}_b(p_s) = (\widehat{\boldsymbol{\beta}}_{xb}^T(p_s), \widehat{\boldsymbol{\beta}}_{zb}^T(p_s))$.

3.3.2 | Weighted extrapolation step

For each stratum, we carry out the extrapolation step separately. First we plot each element q of the vector $\widehat{\boldsymbol{\beta}}(p_s)$, $s = 1, \ldots, S$ against p_s and fit an extrapolation function $f(p)$ using regression methods. The SIMEX estimator for the exposure effect of the qth food outlet is obtained as the intercept of the extrapolation function, $\widehat{\beta}_{SIMEX_q} = \hat{f}_q(0)$, which provides an estimate for when the ideal case is available, that is, no geocode coarsening $p = 0\%$. The extrapolation function can be any continuous curve, although quadratic or cubic polynomials have been shown to work well in practice.[28] In our simulations we explore both of these polynomials as extrapolation functions. The same extrapolation procedure is repeated for each element of the vector of coefficients for the error-free covariates $\boldsymbol{Z}$, since they also need bias correction when they are correlated with the true exposure.

To take the variability of estimates into account, we propose using weighted quadratic or cubic regressions where the weight is the inverse of the empirical variance of estimates at each coarsening level, that is, $1/V_{s,q}$ where $V_{s,q} = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\beta_{q,b}}(p_s) - \widehat{\beta_q}(p_s))^2$. Naturally, an unweighted version of the extrapolation function can also be fitted. However, as shown in the simulation section, the performance of the weighted extrapolation function is better. This weighting takes into account the fact that as p_s increases, $V_{s,q}$ tends to increase as well.

For each stratum, the covariance matrix of $\widehat{\boldsymbol{\beta}}_{SIMEX} = (\widehat{\boldsymbol{\beta}_{\boldsymbol{x}SIMEX}}^T, \widehat{\boldsymbol{\beta}_{\boldsymbol{z}SIMEX}}^T)^T$ can be obtained by a standard bootstrap approach. We resample the subjects with replacement and repeat entire SIMEX procedure to obtain $\widehat{\boldsymbol{\beta}}_{SIMEX}^{(k)}$ for $k = 1, \ldots, K$ boostrapped samples. Thus, for each stratum, the boostrap variance estimator is $\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX}) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{\beta}}_{SIMEX}^{(k)} - \overline{\widehat{\boldsymbol{\beta}}_{SIMEX}}\right)\left(\widehat{\boldsymbol{\beta}}_{SIMEX}^{(k)} - \overline{\widehat{\boldsymbol{\beta}}_{SIMEX}}\right)^T$.

3.3.3 | Inverse-variance weighted combination of estimates

Since the strata defined by presence or absence of a ZIP centroid in the subject's neighborhood are independent, we propose using an inverse-variance weighted average to combine $\widehat{\boldsymbol{\beta}}_{SIMEX,1}$ and $\widehat{\boldsymbol{\beta}}_{SIMEX,2}$:

$$\widehat{\boldsymbol{\beta}}_{SIMEX,\text{combined}} = \boldsymbol{\gamma_1} \odot \widehat{\boldsymbol{\beta}}_{SIMEX,1} + \boldsymbol{\gamma_2} \odot \widehat{\boldsymbol{\beta}}_{SIMEX,2}. \tag{2}$$

where $\odot$ denotes element-wise product operation, and $\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$ are Q-dimensional weight vectors. Each element of the weight vector is proportional to the inverse variance of the SIMEX estimator for each element of the regression coefficient vector within each stratum. That is, the weights for the qth coefficient are

$$\gamma_{1_q} = \frac{\frac{1}{Var(\widehat{\beta}_{SIMEX,1_q})}}{\frac{1}{Var(\widehat{\beta}_{SIMEX,1_q})} + \frac{1}{Var(\widehat{\beta}_{SIMEX,2_q})}}, \gamma_{2_q} = \frac{\frac{1}{Var(\widehat{\beta}_{SIMEX,2_q})}}{\frac{1}{Var(\widehat{\beta}_{SIMEX,1_q})} + \frac{1}{Var(\widehat{\beta}_{SIMEX,2_q})}}. \tag{3}$$

If $\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$ are known, then

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX,\text{combined}}) = diag(\boldsymbol{\gamma_1}^{\odot 2})\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX,1}) + diag(\boldsymbol{\gamma_2}^{\odot 2})\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX,2}). \tag{4}$$

The operation $^{\odot 2}$ denotes that each element of a vector is raised to the second power, and $diag()$ turns a vector into a diagonal matrix. However, the variance of the SIMEX estimator in each stratum is estimated empirically, using the standard bootstrap method described in the previous section. Thus, weights $\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$ can be also estimated by the empirical bootstrap variance of SIMEX estimates. Then, the Q-dimensional vector of variances of combined SIMEX estimators can be estimated by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX,\text{combined}}) = diag(\widehat{\boldsymbol{\gamma}}_{\mathbf{1}}^{\odot 2})\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX,1}) + diag(\widehat{\boldsymbol{\gamma}}_{\mathbf{2}}^{\odot 2})\widehat{Var}(\widehat{\boldsymbol{\beta}}_{SIMEX,2}). \tag{5}$$

We show in simulations that (5) is an adequate variance estimator.

4 | BIAS ANALYSIS AND SIMULATION STUDY

In this section, we examine the bias of the naïve estimator and evaluate the performance of our proposed method through simulation, and compare it with alternative methods. For both of these goals, we use a subset of data that has "true" street-level geocodes as described in Section 2.3 and then set 15% of the businesses to have ZIP centroid geocodes. We considered a linear regression to generate the response variable Y following assumptions in Section 3.1. We generated a subject-specific covariate Z_i using a Bernoulli distribution with probability $\frac{1}{2}$, we assume Q=6 distinct outlet types within 1 mile from subject i, and generated a standard normal residual error ϵ_i. We let $\beta_0 = 20$, $\beta_z = 2$, and we use the same value of the coefficient, $\beta_x = 3.2$ for all business types. In the regression model with multiple error-prone covariates, having equal true coefficients for $\boldsymbol{X}$ enables us to identify the impact of the (potentially) different measurement error distributions across outlet types. With a constant coefficient, any differences in the observed percent bias can be attributed to differences in the measurement error distribution, not different coefficient values.

We generated a total of 1000 datasets of size $N = 1000$, by sampling the locations of 1000 schools. Other simulation parameters for each dataset are $B = 20$, $K = 100$, and $S = 30$ levels of coarsening from $p_0 = 15\%$ to $p_S = 50\%$.

4.1 | Different patterns of bias by presence of ZIP centroid in the buffer

Figure 3 displays the different pattern of average estimates $\widehat{\boldsymbol{\beta}_{\boldsymbol{x}}}(p_s)$ against each coarsening level p_s between subjects with and without ZIP centroids within one-mile circular regions, illustrating the benefit of two independent SIMEX procedures instead of single simulation and extrapolation steps. Compared with Figure 1B which is an all-subject combined analysis, Figure 3 shows qualitatively different patterns of bias after stratifying subjects. In the case of no ZIP centroid, estimates are inflated given the systematic undercounting, whereas in the case of at least one ZIP centroid in the buffer, the bias can be either positive or negative for a particular business. However, in both strata, the bias remains in a consistent direction for each business type.

To see the how the differences in the direction of bias arise, consider the ordinary least square (OLS)-estimators. Omitting $\boldsymbol{Z}$ for simplicity, the estimate of $\boldsymbol{\beta}_x$ can be derived as $\widehat{\boldsymbol{\beta}_x} = \boldsymbol{\Sigma}_W^{-1} cov(\boldsymbol{W}, \boldsymbol{Y}) \approx \boldsymbol{\Sigma}_W^{-1} cov(\boldsymbol{W}, \boldsymbol{X}\beta_x) = \boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_{XU})\boldsymbol{\beta_x} = \left[\boldsymbol{I} - \boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_{WU}\right]\boldsymbol{\beta_x} \equiv [\boldsymbol{I} - \boldsymbol{\Omega}]\,\boldsymbol{\beta_x}$. Attenuation or inflation in estimates is determined by $[\boldsymbol{I} - \boldsymbol{\Omega}]$. As shown in Table 1,

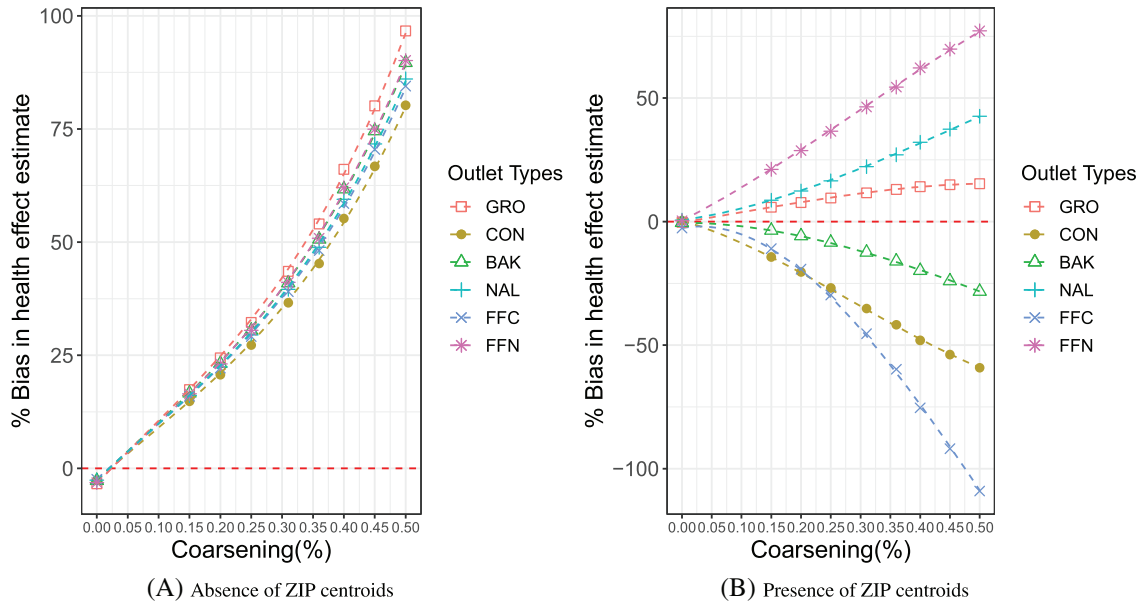


FIGURE 3 Percent bias in health estimate against each coarsening level p_s, stratified by whether schools do not have a ZIP centroid within their buffers (A) or have at least one (B). The naïve estimates occur at $p_0 = 15\%$ in both strata. Additional coarsening to original observed data magnifies bias, as expected. SIMEX estimators at $p = 0\%$ are obtained by fitting the extrapolation curve (dashed curve) using weighted cubic regression. Other simulation parameters for this experiment are S = 30, B = 20, and 1000 Monte Carlo samples. Presented outlet types are: grocery (GRO), convenience store (CON), bakery (BAK), non-alcohol drinking place (NAL), fast food chain (FFC), fast food nonchain (FFN)

generally $\mathbf{\Omega}$ cannot be null because $\mathbf{\Sigma}_{WU}(= \mathbf{\Sigma}_{XU} + \mathbf{\Sigma}_U)$ is not zero. Furthermore, as Table 1 shows, there are nonzero positive or negative covariances. The sign and magnitude of covariance matrices are determined by spatial locations of subjects and businesses. Appendix Section A.2 shows algebraic justifications for the varying directions of bias shown in Figure 3, for the case of regressions with a single exposure.

4.2 | Comparison of methods

We compared our SC-SIMEX method with an adapted SIMEX that uses our newly proposed simulation strategy as defined above but does not stratify the data (labeled "Traditional"), and a covariate-adjusted SIMEX that includes a binary indicator of the presence of ZIP centroid within each subject's buffer as a covariate in the outcome model ('Traditional + centroid') but does not stratify. For the SIMEX extrapolation functions, we use quadratic, weighed quadratic, cubic, and weighted cubic regression due to the curvature in the pattern of the estimated health effect by coarsening level.

Table 2 presents the percent bias, coverage probabilities of their 95% Wald confidence intervals (CI), and mean squared error×100 (MSE) for the naïve estimates, and the SIMEX estimates that use a weighted cubic extrapolation function. The naïve estimates have a mix of attenuated and inflated estimates depending on the outlet type. The estimates for bakeries and convenience stores are underestimated, while those for fast food chains, groceries, fast food nonchains, and non-alcohol drink places are overestimated. Compared with the naïve estimate, corrected estimates using traditional, adjusted SIMEX, and split-combine SIMEX all substantially reduced the bias. However, in addition, the split-combine SIMEX has the lowest mean squared error.

As shown in the literature,[27] the properties of the SIMEX estimator depend on the extrapolation functions. As in prior work, we also note that cubic regression works well in all cases, correcting for more bias than quadratic regression (Appendix Web Table 2). In Table 2, we have a wide confidence interval due to overly conservative bootstrap standard errors of SIMEX methods. However, similar amount of difference between the sample standard deviation of the estimates and the average of bootstrapped standard error estimates has been generally shown in other SIMEX literature (see eg, References 21 and bib29). Despite this difference, SC-SIMEX produces smaller standard error estimates than other SIMEX methods (Appendix Web Table 5). To increase the Monte Carlo precision, increasing B could be

TABLE 2 Simulation results comparing naïve, traditional SIMEX, adjusted SIMEX, and split-combine SIMEX; weighted cubic regression is used for extrapolation $f(\cdot)$

Criteria/method	BAK	CON	FFN	GRO	FFC	NAL
Estimate (percent bias)						
Naïve	2.80 (−12.6%)	2.79 (−12.7%)	3.28 (2.6%)	3.35 (4.6%)	3.38 (5.7%)	3.53 (10.3%)
Naïve + centroid	3.21 (0.3%)	3.05 (−4.7%)	3.30 (3.2%)	3.40 (6.2%)	3.94 (23.2%)	3.28 (2.6%)
Traditional SIMEX	3.19 (−0.5%)	3.15 (−1.4%)	3.22 (1.1%)	3.23 (0.5%)	3.04 (−5.0%)	3.19 (−0.4%)
Traditional SIMEX+centroid	3.18 (−0.5%)	3.15 (−1.7%)	3.22 (−0.1%)	3.20 (0.5%)	3.15 (−1.5%)	3.21 (0.3%)
Split-combine SIMEX	3.16 (−1.1%)	3.17 (−0.9%)	3.19 (−0.4%)	3.17 (−0.9%)	3.13 (−2.3%)	3.17 (−0.8%)
Mean squared error x 100						
Naïve	18.1	22.4	5.1	2.6	4.9	12.5
Naïve + centroid	1.6	7.1	4.9	4.3	56.4	2.2
Traditional SIMEX	3.7	10.3	8.3	1.1	6.1	3.3
Traditional SIMEX+centroid	3.1	8.3	6.7	0.9	3.2	2.9
Split-combine SIMEX	1.6	4.6	3.9	0.5	1.9	1.8
Coverage probability						
Naïve	25%	73%	98%	52%	88%	51%
Naïve + centroid	96%	89%	93%	9%	0%	92%
Traditional SIMEX	100%	100%	100%	100%	97%	100%
Traditional SIMEX+centroid	100%	100%	100%	100%	100%	100%
Split-combine SIMEX	99%	99%	100%	99%	98%	99%

Note: We used randomly sampled N = 1000 schools. The results are from 1000 Monte Carlo simulations with B = 20, K = 100, and S = 30. True coefficient value is 3.2 for all store types. Results are presented in the order of bakery (BAK), convenience store (CON), fast food nonchain (FFN), grocery (GRO), fast food chain (FFC), non-alcohol drinking place (NAL).

helpful, as recommended by Carroll,[27] although it comes at a price of the computational burden. In our simulations, we kept B relatively small ($B = 20$) to alleviate this computational burden, however, in the following data analysis, we use B = 100.

4.3 | Impact of the spatial distribution of subjects and outlet locations

We examined the impact of the spatial distribution of subjects and businesses on the performance of the split-combine SIMEX and other methods, in two ways. First, we conducted simulations using 400 subjects uniformly distributed on a 10×10 grid, and 400 environment features. The ZIP centroids were assumed to be the centroid of each 1 unit square on the grid. Even in this simple scenario, the split-combine SIMEX has better MSE and lower bias compared to the other SIMEX-based approaches. However, the differences among the methods are small as the other methods also perform well in this simple scenario (results not shown).

We also used a more realistic alternate scenario for the distribution of study locations. Specifically, we considered only the location of 1000 schools for seventh-grade children. Children in seventh grade typically attend middle school, which typically consists of grades 6 to 8. In contrast to elementary schools (grades 1-5) which often serve a more localized residential neighborhood, middle schools draw from a larger geographical area and thus tend to have a different spatial distribution. The results of this simulation (Appendix Web Tables 3 and 4) show that the extent of the bias of the naïve approach depends on the spatial location of subjects. For instance, while the bias for the naïve estimate was positive for nonchain fast food (FFN) when considering all schools, the bias is negative when considering only the location of schools for 7th-grade children.

5 | APPLICATION TO FITNESSGRAM DATA

5.1 | Data

We implement our proposed method to investigate the association between the availability of six types of food stores near schools (40 476 outlets) and the body mass index of children who attended California public schools during 2001. On average, 7.2% of the food outlets were coarsened to their ZIP centroid. Grocery stores had the highest coarsening percentage (8.8%), followed by bakeries (8.1%), nonchain fast food restaurants (7.4%), non-alcohol drinking places (6.9%), convenience stores (5.4%), and fast food restaurant chains (5.1%).

Our outcome measure is child-level body mass index z-scores (BMIz) that were collected using the Fitnessgram® assessment in California in 2001. Fitnessgram® is a widely used youth fitness test; its use is state-mandated in California for grades 5, 7, and 9 in public schools as part of the state's broader assessment and reporting programs.[30] In contrast to adults where adiposity is measured and BMI computed as weight in kilograms divided by height in meters squared, age- and sex-specific BMIz scores are used to report given that they are still growing. To obtain the BMIz, each child's BMI is standardized according to an age- and sex-specific reference distribution published by the US center for disease control (CDC).[31] In all analyses, we adjust for child and school-level confounders. Specifically, in addition to child-level control variables such as sex and age, we also adjusted for school characteristics, including the level of urbanization near the school,[32] the proportion of adults with a college degree or higher within the school's census tract, median household income of residents in the school census tract, the school's enrollment by race/ethnicity, and percent of students eligible for free and reduced-price meals.

We focus our analysis on BMIz scores for N = 53 239 Asian children in fifth, seventh, or ninth grade, who attended one of 1166 schools. We focus the analysis on Asian children because previous research had documented a counter-intuitive finding among this population subgroup (Sanchez et al[33]). More specifically, prior research showed a significant negative association between exposure to fast food restaurants near schools and obesity among Asian children. We hypothesize that this counterintuitive result is at least partly due to errors in measuring exposure due to geocode coarsening.

Given that children are nested in schools, we use a mixed model with a random intercept of schools to examine the association between food outlets and child-level BMIz. The SIMEX procedure is known to also perform well in linear mixed measurement error models.[34] Exposure to each outlet type was measured by the count of businesses within a one-mile buffer around the school. Because initial exploratory analysis revealed a lack of linearity in the association between exposure to food outlets and BMIz, we use the log_2 transformation in our analysis (adding one to all outlet counts to ensure inclusion of schools with 0 outlets nearby). Note that, the use of the log_2 transformation does not create challenges for the SC-SIMEX approach, for several reasons. First, given its monotonicity, the log_2 transformation does not fundamentally change the measurement error properties that motivates the SC-SIMEX, such as nonpositive mean and negative covariance between the measurement error and the true exposure among subjects without ZIP centroids in their buffers. Second, the SIMEX algorithm, more broadly, retains its bias-correction properties for monotone transformations.[35] Third, given that our simulation procedure directly generates $\mathcal{W}$ instead of measurement errors U in a conventional SIMEX approach, changes to the distribution of U due to the log_2 transformation do not impede the simulation step. Moreover, the transformation takes place at the estimation step once the error-prone count exposure $\boldsymbol{W}(p)$ has been simulated. Hence, we take advantage of direct simulation of $\mathcal{W}$ to generate $\log_2(\boldsymbol{W}(p) + 1)$.

Our primary analysis runs separate "single outlet" models for each food outlet type since one of our goals was to examine if measurement error was a reason for counterintuitive findings in previously published single outlet models. However, we also present results from a model that simultaneously includes exposure to all outlet types (ie, six separate predictors) to highlight efficiency gains of SC-SIMEX when multiple outlets are included, as demonstrated from the simulation study in Section 4.

For each model, we compare the naïve results with corrected estimates from traditional SIMEX, a traditional SIMEX adjusted for the presence of ZIP centroid within the school buffer, and our proposed split-combine SIMEX. We use $B = 100, K = 100$, and $S = 10$, and p_0 is specific to the outlet type (as shown above) and $p_S = 50\%$. Final inferences were derived by trimming the top and bottom 2% of the K bootstrap samples to remove outliers. Given the simulation results, in this analysis, we only used weighted cubic extrapolation functions.

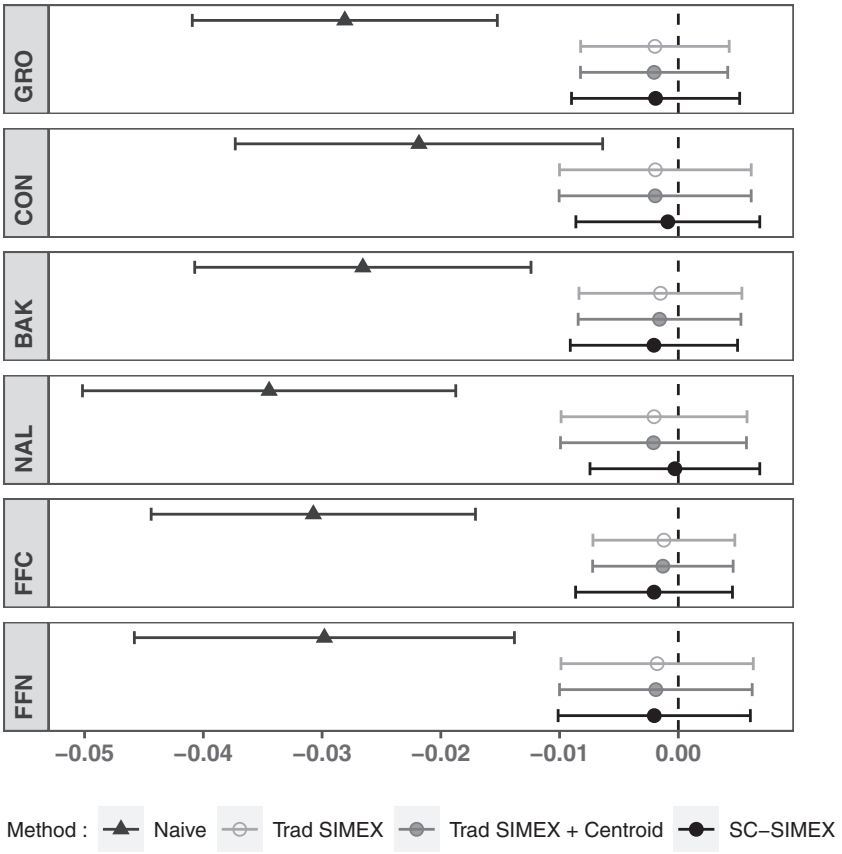


FIGURE 4 Single outlet models. Coefficient estimates and 95% confidence intervals from the naïve, traditional, covariate-adjusted, and SC-SIMEX approaches, illustrating the impact of geocode coarsening on the naïve regression. The estimates represent the difference in BMIz per each two-fold higher exposure to a given food outlet type. Each subpanel represents the results of separate models, one for each food outlet type, among a sample of Asian children attending California public schools. See Figure 1 for outlet abbreviations

5.2 | Results

Figure 4 shows coefficient estimates and confidence intervals of the association between exposure to different food outlet types and children's BMIz score, estimated from models that included one store type at a time. The naïve regression analysis shows a negative association between BMIz and exposure to any one type of food outlet, including replication of the prior finding of fast food restaurants. However, the corrected estimates obtained from the SIMEX methods do not support the naïve results. The corrected estimates for all outlet types show an overall attenuation compared to the naïve coefficients, and all confidence intervals include 0. The figure thus demonstrates the usefulness of correction for geocode coarsening and points to measurement error as a possible explanation for the counter-intuitive effects shown in prior research.

In addition to the primary analysis, Table 3 highlights that SC-SIMEX offers an efficiency advantage for the multiple-store model, compared to other SIMEX approaches. The table shows that SC-SIMEX estimates have the lowest standard errors for all outlet types.

6 | DISCUSSION

In this article, we focused on two major objectives. First, we examined the measurement error distribution and bias in estimating the health impact of point-referenced built environment exposure that results from the presence of geocode coarsening. We showed analytically and via simulation that measurement error that occurs from inaccurate geocodes does not follow classical measurement assumptions nor known parametric distributions. We also showed that bias in the

TABLE 3 Multiple outlet model

	Naïve	Traditional	Trad + centroid	SC-SIMEX
Outlet Type	$\widehat{\beta}(SE)$	$\widehat{\beta}(SE)$	$\widehat{\beta}(SE)$	$\widehat{\beta}(SE)$
GRO	−0.96 (1.17)	−0.20 (0.77)	−0.20 (0.77)	−0.14 (**0.71**)
CON	0.59 (1.05)	−0.05 (0.71)	−0.04 (0.71)	0.06 (**0.61**)
BAK	0.18 (1.17)	0.06 (0.73)	0.06 (0.73)	−0.07 (**0.70**)
NAL	−1.92 (1.06)	−0.14 (0.67)	−0.14 (0.67)	0.21 (**0.52**)
FFC	−1.50 (1.09)	0.10 (0.69)	0.09 (0.69)	−0.14 (**0.58**)
FFN	−0.60 (1.16)	−0.04 (0.79)	−0.05 (0.79)	−0.11 (**0.66**)

Note: Coefficient estimates and standard error (SE) of the estimates from naïve, traditional, covariate-adjusted, and SC-SIMEX approaches, showing lower standard errors for the SC-SIMEX for all outlet types. The estimates represent the difference in BMIz per each two-fold higher exposure to a given food outlet type. Estimates and standard errors were multiplied by 100 to improve readability. The results are from a single model that simultaneously includes all outlet types, among a sample of Asian children attending California public schools. See Figure 1 for outlet abbreviations. In each row, bold values identify the smallest standard error for the coefficient of each outlet type.

estimation of the health effect is not always toward the null. The second objective was to correct the bias incurred by coarsened geocodes. Based on our finding that the measurement error distribution is a mixture that can be decomposed by whether or not the location of the subject had a ZIP centroid within a circular buffer, we proposed a new split-combine SIMEX approach which takes the mixture of the measurement error distributions into account. Our simulation for evaluating the method shows that the SC-SIMEX reduces the bias considerably and is more efficient than traditional and covariate-adjusted SIMEX methods. After we addressed our two objectives, we were able to apply our new approach to Fitnessgram® data to estimate the association between food environment around schools and childhood BMI, and demonstrated that qualitative differences in study results could occur when we account for the measurement error due to geocode coarsening.

One advantage of SIMEX is that it is a general methodology that can be adapted to cases where the measurement error biases cannot be derived in closed-form.[27] However, its applications were until now limited to parametric measurement error distributions where estimated parameter values were available (eg, measurement error variances). The split-combine SIMEX fits in the broader SIMEX literature by extending the SIMEX to cases where the measurement errors have a more complex, heteroscedastic distribution but are still replicable by computer simulations. In our case, replication of the measurement error distribution is achieved by using an indicator of the accuracy of the geocodes, which is a readily available by-product of the geocoding process.

While a large amount of measurement error literature has focused on continuous measurement error and misclasification error, there has been much less attention for mismeasurement of a count covariate and its effect on bias in regression coefficient estimates. Error-prone count predictors are typically transformed so that their distributions are normalized (or at least more symmetric),[27] or a normal approximation is used when the count is large such as in population estimation studies.[36] There are other works that use Poisson regression models for count *outcomes* with continuous erroneous predictors.[37,38] Other discrete error distributions such as misclassification errors have been studied for categorical predictors.[39-41] Our extension of SIMEX will provide a method to tackle measurement errors in count exposure.

Our work points to several areas for future research of SC-SIMEX applications. First, we suggest further investigation of different coarsening probabilities according to covariates. Our work here considers that businesses are coarsened-at-random (CAR) so that we randomly sampled businesses to be coarsened during pseudo-error generation. However, it may be possible to assign a different probability of being coarsened according to the characteristics of business locations. For example, NETS data for California shows that rural businesses have a 2 to 4 times higher proportion of coarsening to ZIP centroids than urban businesses, depending on the year in which the data are recorded. Another potential improvement can be achieved by partitioning samples into more strata so that we can define homogeneous measurement error distribution within the strata. If a subject's buffer is fully contained by one large ZIP polygon and its centroid is within the buffer, then the exposure of the subject will always be equal to or overcounted compared to the true exposure. This type of location will only have nonnegative measurement errors. Splitting subjects into more strata will improve the performance, but it is also complicated because each stratum should have enough sample size for consistent estimation. Additional research is needed to extend the implementation of SC-SIMEX to different measures of exposures, such as distance to establishments or density-based clustering.

Our proposed method of dividing the sample in this article is motivated by studies that use counts in a buffer which is one of the widely used exposure methods. There are a large number of public health studies that use count exposures to investigate the health effects of environmental features, for which the split-combine SIMEX provides one possible measurement error correction strategy which may be beneficial to correct bias induced by geocoding coarsening and result in qualitatively different findings.

ACKNOWLEDGEMENT

The authors would like to thank Michael R. Elliott for feedback on the paper. The work was funded by NIH R01-HL131610; data collection used in the applications was partially funded by NIH R01-HL136718.

DATA AVAILABILITY STATEMENT

The data that support this study are not available from the authors due to privacy concerns (Fitnessgram data) and data use agreements for proprietary data (NETS). Fitnessgram data may be obtained directly from the California Department of Education. NETS data may be obtained from Walls and Associates.

ORCID

Jung Y. Won https://orcid.org/0000-0002-8577-7307
Brisa N. Sánchez https://orcid.org/0000-0002-4824-7200

REFERENCES

1. Zimmerman DL. Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*. 2008;64(1):262-270.
2. Auchincloss AH, Moore KA, Moore LV, Diez RAV. Improving retrospective characterization of the food environment for a large region in the United States during a historic time period. *Health Place*. 2012;18(6):1341-1347.
3. Roux AV, Mujahid MS, Hirsch JA, Moore K, Moore LV. The impact of neighborhoods on cardiovascular risk: the MESA neighborhood study. *Glob Heart*. 2016;11(3):353-363.
4. Kaufman TK, Sheehan DM, Rundle A, et al. Measuring health-relevant businesses over 21 years: refining the National Establishment Time-Series (NETS), a dynamic longitudinal data set. *BMC Res Notes*. 2015;8:507.
5. Lovasi GS. Communities designed to support cardiovascular health for older adults; 2015. https://grantome.com/grant/NIH/R01-AG049970-01A1.
6. Hirsch JA, Meyer KA, Peterson M, et al. Obtaining longitudinal built environment data retrospectively across 25 years in four US Cities. *Front Public Health*. 2016;4:65.
7. Hickson DA, Diez Roux AV, Smith AE, et al. Associations of fast food restaurant availability with dietary intake and weight among African Americans in the Jackson heart study, 2000-2004. *Am J Public Health*. 2011;101(Suppl. 1):S301-S309.
8. Berger N, Kaufman TK, Bader MDM, et al. Disparities in trajectories of changes in the unhealthy food environment in New York City: a latent class growth analysis, 1990-2010. *Soc Sci Med*. 2019;234:112362.
9. Jacquez GM. A research agenda: does geocoding positional error matter in health GIS studies? *Spat Spatio-temporal Epidemiol*. 2012;3:7-16.
10. Zimmerman DL, Li J. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *Int J Health Geogr*. 2010;9:10.
11. Zandbergen PA, Hart TC, Lenzer KE, Camponovo ME. Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets. 2012;3:69-82.
12. Healy MA, Gilliland JA. Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research. 2012;3:55-67.
13. Zandbergen PA. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health*. 2007;7:37.
14. Goldberg DW. *A Geocoding Best Practices Guide*. Springfield, IL: North American Association of Central Cancer Registries; 2008.
15. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *J Am Stat Assoc*. 1994;89:1314-1328.
16. Küchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*. 2006;62:85-96.
17. Küchenhoff H, Lederer W, Lesaffre E. Asymptotic variance estimation for the misclassification SIMEX. *Comput Stat Data Anal*. 2007;51(12):6197-6211.
18. Parveen N, Moodie E, Brenner B. The non-zero mean SIMEX: improving estimation in the face of measurement error. *Observ Stud*. 2015;1:91-123.
19. Parveen N, Moodie E, Brenner B. Correcting covariate-dependent measurement error with non-zero mean. *Stat Med*. 2017;36:2786-2800.
20. Lockwood JR, McCaffrey DF. Simulation-extrapolation with latent heteroskedastic error variance. *Psychometrika*. 2017;82:717-736.
21. Alexeeff SE, Carroll RJ, Coull B. Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics*. 2016;17:377-389.

22. Business Dynamics Research Consortium. National Establishment Time-Series (NETS) Database, Denver, CO http://exceptionalgrowth.org. Accessed June 5, 2019.
23. Reference USA. http://www.referenceusa.com. Accessed December 17, 2021.
24. Dun and bradstreet;2021. https://www.dnb.com/duns-number.html. Accessed December 17, 2021.
25. NAICS Association. SIC codes and counts by division. https://www.naics.com/sic-codes-counts-division/?div=G. Accessed January 21, 2019.
26. Athens JK, Duncan DT, Elbel B. Proximity to fast-food outlets and supermarkets as predictors of fast-food dining frequency. *J Acad Nutr Diet*. 2016;116(8):1266-1275.
27. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press; 2006.
28. Apanasovich TV, Carroll RJ, Maity A. SIMEX and standard error estimation in semiparametric measurement error models. *Electron J Stat*. 2009;3:318-348.
29. Mao G, Wei Y, Liu Y. SIMEX method for censored quantile regression with measurement error. *Commun Stat Simul Comput*. 2017;46(10):7552-7560.
30. California Department of Education. Physical Fitness Testing (PFT). http://www.cde.ca.gov/ta/tg/pf/. Accessed June 05, 2019.
31. Pietrobelli A, Faith MS, Allison DB, Gallagher D, Chiumello G, Heymsfield SB. Body mass index as a measure of adiposity among children and adolescents: a validation study. *J Pediatr*. 1998;132(2):204-210.
32. Ingram DD, Franco SJ. 2013 NCHS urban-rural classification scheme for counties. *Vital Health Stat*. 2014;2(166):1-73.
33. Sánchez BN, Sanchez-Vaznaugh EV, Uscilka A, Baek J, Zhang L. Differential associations between the food environment near schools and childhood overweight across race/ethnicity, gender, and grade. *Am J Epidemiol*. 2012;175(12):1284-1293.
34. Wang N, Carroll RJ, Lin X, Gutierrez RG. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *J Am Stat Assoc*. 1998;93(441):249-261.
35. Eckert RS, Carroll RJ, Wang N. Transformations to additivity in measurement error models. *Biometrics*. 1997;53(1):262-272.
36. Cheng Y, Chakraborty A, Datta G. Hierarchical Bayesian methods for combining surveys. Proceedings of the Survey Research Methods Section, American Statistics Association; 2015:4099-4111.
37. Markus T. Modelling count data with heteroscedastic measurement error in the covariates Discussion Paper 58, SFB 386. Ludwig-Maximilians-Universität München; 1997.
38. Kukush A, Schneeweis H, Wolf R. Three estimators for the Poisson regression model with measurement errors. *Stat Pap*. 2004;45(3):351-368.
39. Gustafson P, Le Nhu D. Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics*. 2002;58(4):878-887.
40. Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC Press; 2010.
41. Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1–basic theory and simple methods of adjustment. *Stat Med*. 2020;39(16):2197-2231.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Won JY, Sanchez-Vaznaugh EV, Zhai Y, Sánchez BN. Split and combine simulation extrapolation algorithm to correct geocoding coarsening of built environment exposures. *Statistics in Medicine*. 2022;41(11):1932-1949. doi: 10.1002/sim.9338