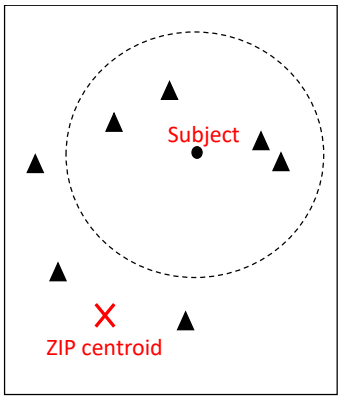
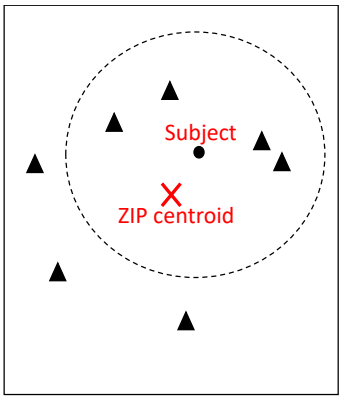
Web-based Supplementary Materials for "Split and combine SIMEX algorithm to correct geocoding coarsening of built environment exposures"

Jung Y. Won, Emma V. Sanchez-Vaznaugh, Yuqi Zhai and Brisa N. Sánchez

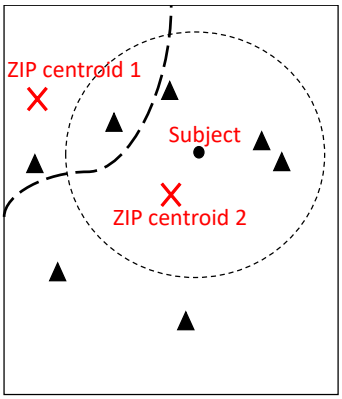
Web Figure 1: Illustration of measurement error mechanism for subjects with and without ZIP centroid within their buffer



(a) Absence of ZIP-code centroid in a buffer zone and the buffer is contained in a ZIP polygon: If all the businesses (triangular dots) are coarsened, a subject (circular dot) loses four businesses.

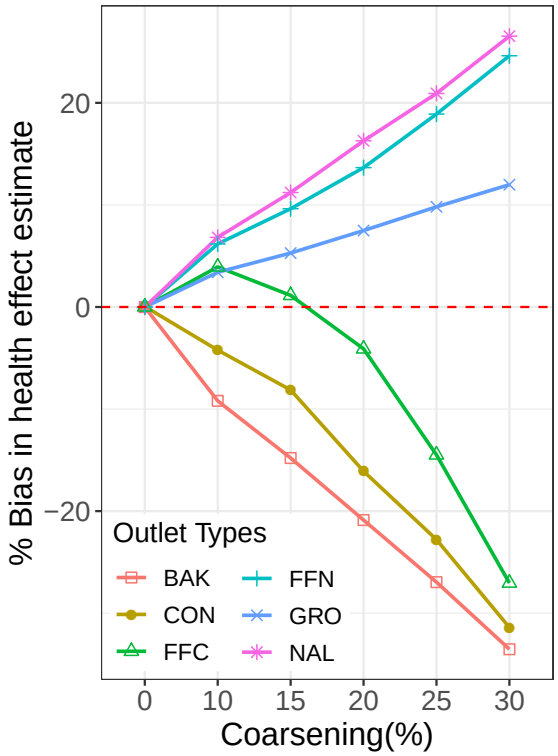


(b) Presence of ZIP-code centroid in a buffer zone and the buffer is contained in ZIP polygon: If all the businesses are coarsened, a subject gains three more businesses.

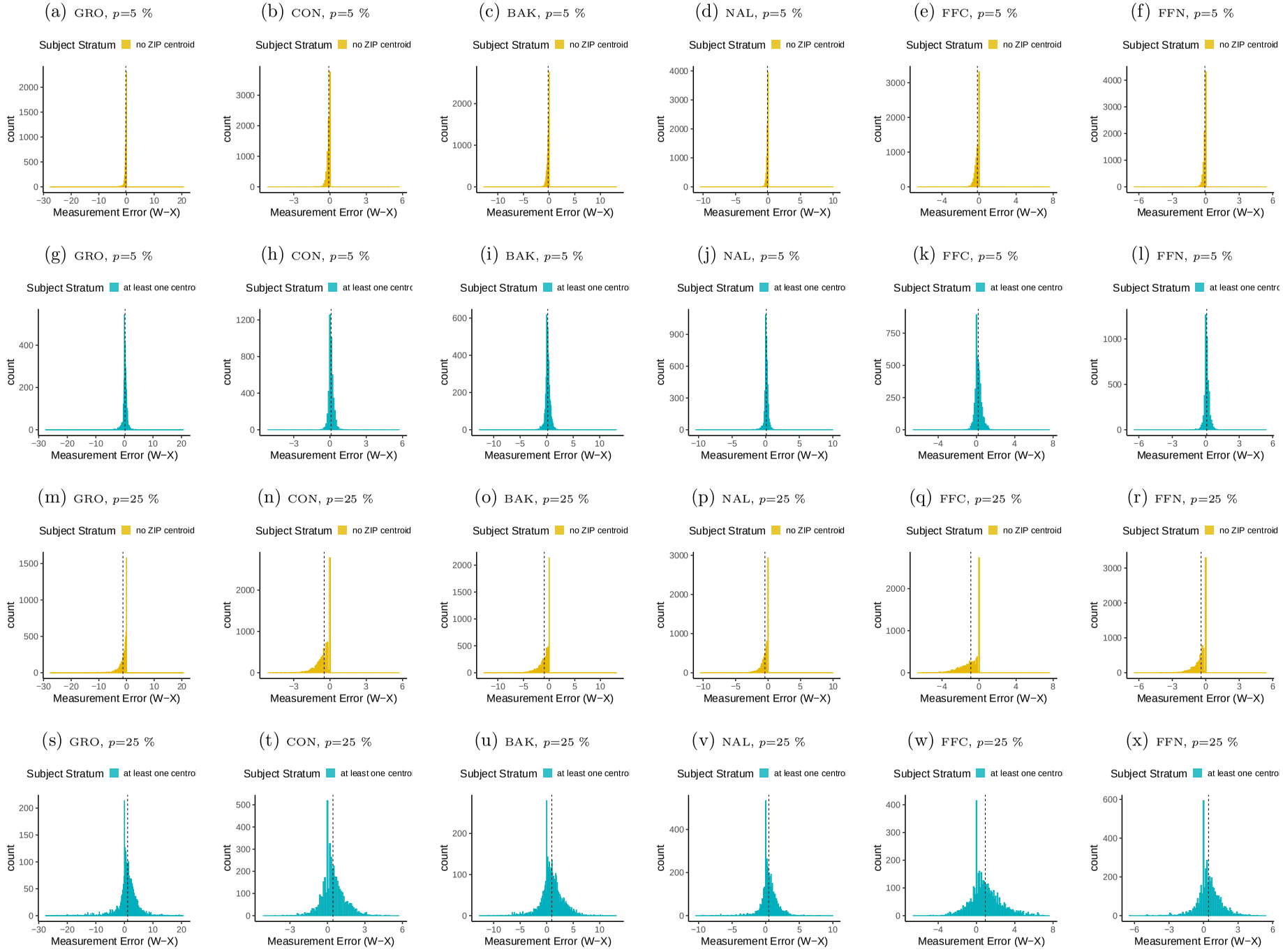


(c) Presence of ZIP-code centroid in a buffer zone and ZIP boundary (long dashed line) crosses the buffer: If all the businesses are coarsened, a subject gains two more business and loses one business.

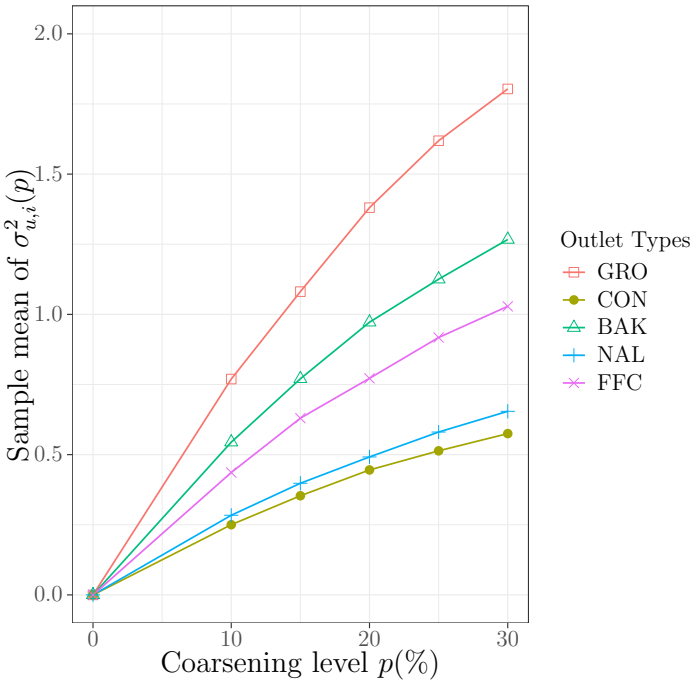
Web Figure 2: Percent bias in health effect estimates from naïve multiple linear regression by coarsening level and food outlet type. Error-prone exposures to six outlet types are regressed in the same regression model. True effect size in the simulation is 3.2 for all outlet types.



Web Figure 3: Strata-specific histogram of measurement error mean, among 12,370 schools when coarsening probability is $p = 5\%$ and $p = 25\%$. Measurement error means are obtained by averaging across 50 simulated values of U_i for each school. Dotted vertical line represents the overall mean measurement error within the stratum. Note that subjects without ZIP centroid in their buffer (8,022 schools) always have non-positive error, while the average error for subjects with at least one ZIP centroid in the buffer (4,348 schools) can be either positive or negative.



Web Figure 4: Sample mean of individual level measurement error variance by coarsening level, for a single randomly selected school. Error variance on each subject was estimated from 50 replications.



Web Table 1: Number of schools by urbanicity and number of ZIP centroids within 1 mile of the school. The majority of the schools do not have any ZIP centroids within their 1 mile circular buffer. Presence of at least one ZIP centroid within the buffer around school implies that school is likely to be located in urbanized area. However, absence of ZIP centroid does not necessarily imply rurality. Therefore, urbanization level and number of ZIP centroids within 1 mile buffer of school are not interchangeable characteristics of school location, although they are moderately correlated.

	Number of ZIP centroids within 1 mile buffer of school		
Urbanization Level	0	1	2 or more
Rural	2,571	270	0
Second city	1,836	372	1
Sub-Urban	1,691	645	13
Urban	1,924	2,562	485
Total	8,022	3,849	499

Web Table 2: Extension of Table 2 in the main paper, showing the performance of different extrapolation functions: Quadratic (Quad), weighted quadratic (Quad (wt)), and cubic regressions are used for extrapolation $f(\cdot)$ for the SIMEX-based approaches. N=1,000 schools are randomly sampled. Results are from 1,000 Monte Carlo simulations with B=20, K=100, and S=30.

Criteria/ Method	$f(\cdot)$	BAK	CON	FFN	GRO	FFC	NAL
Percent Bias							
Naïve		-12.6%	-12.7%	2.6%	4.6%	5.7%	10.3%
Naïve + centroid		0.3%	-4.7%	3.2%	6.2%	23.2%	2.6%
Traditional SIMEX	Quad	6.4%	14.9%	3.3%	-4.1%	8.2%	-4.5%
	Quad (wt)	4.0%	9.0%	2.5%	-2.9%	3.7%	-3.2%
	Cubic	-0.4%	-1.1%	2.1%	0.5%	-6.8%	-0.7%
Traditional SIMEX	Quad	-2.0%	1.4%	4.2%	-1.6%	-13.3%	-0.5%
+centroid	Quad (wt)	-1.5%	0.3%	2.7%	-1.0%	-9.4%	-0.3%
	Cubic	-0.9%	-2.5%	0.2%	0.7%	-2.9%	0.5%
Split-combine SIMEX	Quad	4.8%	4.7%	2.6%	2.7%	8.3%	0.6%
	Quad (wt)	2.8%	2.8%	1.6%	1.8%	4.7%	0.2%
	Cubic	-1.7%	-1.3%	-0.4%	-1.2%	-3.4%	-1.1%
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	Quad	8.9	37.4	12.5	2.8	12.3	7.0
	Quad (wt)	4.9	18.6	8.6	1.7	4.9	4.8
	Cubic	4.8	16.9	11.7	1.2	5.2	5.1
Traditional SIMEX	Quad	4.6	12.5	11.8	1.2	22.5	4.6
+centroid	Quad (wt)	3.3	9.2	8.1	0.9	12.1	3.6
	Cubic	4.5	15.5	11.2	1.1	4.8	4.9
Split-combine SIMEX	Quad	3.8	6.7	4.6	1.2	8.4	1.8
	Quad (wt)	2.2	5.0	4.1	0.8	3.6	1.8
	Cubic	1.8	5.3	4.1	0.6	2.7	2.0
Coverage							
Naïve		25%	73%	98%	52%	88%	51%
Naïve + centroid		96%	89%	93%	9%	0%	92%
Traditional SIMEX	Quad	86%	67%	98%	72%	83%	94%
	Quad (wt)	84%	73%	94%	75%	91%	92%
	Cubic	100%	100%	100%	100%	100%	100%
Traditional SIMEX	Quad	98%	99%	96%	94%	25%	99%
+centroid	Quad (wt)	94%	93%	93%	93%	32%	97%
	Cubic	100%	100%	100%	100%	100%	100%
Split-combine SIMEX	Quad	83%	91%	96%	81%	55%	98%
	Quad (wt)	85%	89%	92%	82%	73%	94%
	Cubic	100%	100%	100%	100%	100%	100%

Web Table 3: Simulation results comparing naïve, traditional SIMEX, adjusted SIMEX, and split-combine SIMEX; weighted cubic regression is used for extrapolation $f(\cdot)$. Compared to Table 2 in the main paper, this table includes a large sample size consisting of all schools (N=12,370); results from 1,000 Monte Carlo simulations with B=20, K=30, and S=30.

Criteria / Method	BAK	CON	FFN	GRO	FFC	NAL
Percent Bias						
Naïve	-15.1%	-9.2%	9.9%	5.5%	1.7%	11.0%
Naïve + centroid	1.0%	3.1%	14.4%	5.7%	14.4%	1.8%
Traditional SIMEX	0.3%	-2.1%	1.0%	0.4%	-5.4%	-0.2%
Traditional SIMEX+centroid	-1.1%	-2.3%	0.2%	0.9%	-1.4%	0.6%
Split-combine SIMEX	-1.2%	-0.9%	-1.2%	-1.0%	-2.0%	-0.2%
Mean Squared Error x 100						
Naïve	23.9	10.6	11.9	3.3	0.8	12.8
Naïve + centroid	0.7	2.8	22.7	3.5	21.8	0.8
Traditional SIMEX	1.5	4.7	4.2	0.5	4.3	1.3
Traditional SIMEX+centroid	1.6	4.3	3.7	0.5	1.4	1.3
Split-combine SIMEX	1.0	3.1	2.7	0.2	1.7	0.7
Coverage Probability						
Naïve	0%	15%	9%	0%	68%	0%
Naïve + centroid	63%	49%	1%	0%	0%	55%
Traditional SIMEX	99%	97%	98%	99%	64%	98%
Traditional SIMEX+centroid	98%	95%	98%	97%	98%	98%
Split-combine SIMEX	95%	93%	95%	95%	90%	98%

Web Table 4: Simulation results comparing naïve, traditional SIMEX, adjusted SIMEX, and split-combine SIMEX; weighted cubic regression is used for extrapolation $f(\cdot)$. Compared to Table 2 in the main paper, the schools in this simulation are a random sample (N=1,000) of 7th grade schools. The comparison of Table 2 and this table illustrates that the direction and magnitude of bias in health effect estimate depends on the spatial distribution of study locations. Results are from 1,000 Monte Carlo simulations with B=20, K=30, and S=30.

Criteria / Method	BAK	CON	FFN	GRO	FFC	NAL
Percent Bias						
Naïve	-0.2%	-0.1%	-10.7%	-2.2%	-0.8%	7.3%
Naïve + centroid	11.3%	11.1%	0.4%	0.4%	9.6%	6.2%
Traditional SIMEX	-0.3%	-2.8%	1.5%	0.5%	-5.0%	-0.8%
Traditional SIMEX+centroid	-0.4%	-1.6%	1.0%	0.3%	-5.0%	-0.2%
Split-combine SIMEX	-1.6%	-1.1%	-1.2%	-0.6%	-2.7%	-0.4%
Mean Squared Error x 100						
Naïve	1.9	5.1	17.4	0.9	1.4	7.4
Naïve + centroid	30.7	10.2	7.0	0.4	2.5	51.7
Traditional SIMEX	1.2	4.4	3.7	0.3	3.6	1.7
Traditional SIMEX+centroid	1.4	4.5	3.9	0.4	3.7	1.8
Split-combine SIMEX	1.4	3.4	3.6	0.4	1.2	1.5
Coverage Probability						
Naïve	99%	99%	83%	97%	99%	79%
Naïve + centroid	0%	79%	88%	95%	87%	0%
Traditional SIMEX	100%	100%	100%	100%	98%	100%
Traditional SIMEX+centroid	100%	100%	100%	100%	97%	99%
Split-combine SIMEX	100%	100%	100%	100%	100%	100%

Web Table 5: Sample standard deviation of the estimates (SE) and the average of bootstrapped standard error estimates (ESE) of SIMEX methods in Table 2 in the main paper. Considering that all SIMEX methods that are presented in our paper generate pseudo-data directly without generating random measurement errors from any specified distribution, it is expected to observe ESE larger than SE.

SE	BAK	CON	FFN	GRO	FFC	NAL
Traditional SIMEX	0.193	0.318	0.286	0.104	0.189	0.181
Traditional SIMEX + centroid	0.176	0.282	0.259	0.096	0.173	0.170
SC-SIMEX	0.121	0.213	0.196	0.066	0.116	0.133
ESE	BAK	CON	FFN	GRO	FFC	NAL
Traditional SIMEX	0.211	0.368	0.333	0.106	0.207	0.211
Traditional SIMEX + centroid	0.198	0.334	0.307	0.099	0.192	0.201
SC-SIMEX	0.157	0.261	0.249	0.088	0.159	0.172

A.1. Derivation of measurement error properties

For each school i, define four mutually exclusive sets of businesses according to whether or not businesses are included in subject i's buffer when their addresses are coarsened: $\mathcal{S}_i^{11}$ (always correctly included), $\mathcal{S}_i^{00}$ (correctly excluded), $\mathcal{S}_i^{10}$ (incorrectly excluded), $\mathcal{S}_i^{01}$ (incorrectly included).

Conditioning on whether a business address is coarsened or not and taking into account that businesses belong to only one of the above sets, the number of outlets that are incorrectly excluded from the exposure count of school i is

$$\begin{aligned} U_i^{10} &= \sum_{j=1}^{J} I_{ij}^{10} = \sum_{j=1}^{J} \{I(x_{ij}=1, w_{ij}=0) \\ &= \sum_{j=1}^{J} \{I(x_{ij}=1, w_{ij}=0 | A_j=1) I(A_j=1) \\ &\quad + I(x_{ij}=1, w_{ij}=0 | A_j=0) I(A_j=0)\} \\ &= \sum_{j \in \mathcal{S}_i^{10}} I(A_j=0) \end{aligned}$$

Similarly, $U_i^{01} = \sum_{j \in \mathcal{S}_i^{01}} I(A_j=0)$. Since coarsening is assumed to occur at random, the indicators $I(A_j=0)$ are *iid* Bernoulli random variables with probability p equal to the probability of coarsening. Thus, U_i^{10} and U_i^{01} follow Binomial distributions: $U_i^{10} \sim Bin(|S_i^{10}|, p)$ and similarly $U_i^{01} \sim Bin(|\mathcal{S}_i^{01}|, p)$. Moreover, these two binomial random variables are independent, given the coarsening at random assumption and that the sets are mutually exclusive. If coarsening is not equally likely for all businesses, U_i^{01} and U_i^{10} will have mixture of Bernoulli distributions, instead of Binomial distributions, making the derivations that follow more complex, but the overall concepts apply to this more general scenario as well.

Thus, we have $\mu_{U_i} = E(U_i) = E(U_i^{01} - U_i^{10}) = p(|\mathcal{S}_i^{01}| + |\mathcal{S}_i^{10}|)$ and $\sigma_{u_i}^2 = Var(U_i) = Var(U_i^{01} - U_i^{10}) = Var(U_i^{01}) + Var(U_i^{10}) - 2Cov(U_i^{01}, U_i^{10}) = p(1-p)\boldsymbol{C}_{ii}$ where $\boldsymbol{C}_{ii} = (|\mathcal{S}_i^{01}| + |\mathcal{S}_i^{10}|)$. Let $\boldsymbol{\mu}_{\mathrm{U}}$ denote the vector of measurement error means for all subjects, and $\boldsymbol{\Sigma}_{\mathrm{U}}$ denote the variance-covariance matrix of measurement errors, with diagonal entries equal to $\sigma_{u_i}^2$.

The remaining entries of $\boldsymbol{\Sigma}_{\mathrm{U}}$ can be derived by considering the intersections of the union of sets of businesses that can introduce errors U_i and U_r in the exposure measures for two different subjects, i and r:

$$\begin{aligned} (\mathcal{S}_i^{01} \cup \mathcal{S}_i^{10}) \cap (\mathcal{S}_r^{01} \cup \mathcal{S}_r^{10}) &= (\mathcal{S}_i^{01} \cap \mathcal{S}_r^{01}) \cup (\mathcal{S}_i^{01} \cap \mathcal{S}_r^{10}) \cup (\mathcal{S}_i^{10} \cap \mathcal{S}_r^{01}) \cup (\mathcal{S}_i^{10} \cap \mathcal{S}_r^{10}) \\ &= \mathcal{S}_{ir}^{01,01} \cup \mathcal{S}_{ir}^{01,10} \cup \mathcal{S}_{ir}^{10,01} \cup \mathcal{S}_{ir}^{10,10}, \end{aligned}$$

where $\mathcal{S}_{ir}^{01,01}$ is the set of businesses that can incorrectly enter the buffers for subject i and subject r simultaneously, when coarsened; $\mathcal{S}_{ir}^{01,10}$ is the set of businesses that can incorrectly enter the ith subject's buffer, but incorrectly leave the rth subject's buffer; and analogously for $\mathcal{S}_{ir}^{10,01}$ and $\mathcal{S}_{ij}^{01,01}$. Let $U_{ik}^{m(1-m),l(1-l)}$ represents a random variable connected to $\{\mathcal{S}_{ij}^{m(1-m),l(1-l)}\}, m, l \in \{0,1\}$, i.e., $U_{ik}^{m(1-m),l(1-l)} \sim Binom(|\mathcal{S}_{ij}^{m(1-m),l(1-l)}|, p)$. Businesses that belong to any of these sets could give correlation among U_i and U_r, as these businesses

introduce errors for both subjects. Therefore,

$$\begin{aligned} Cov(U_i, U_r) &= Cov(U_i^{01} - U_i^{10}, U_j^{01} - U_j^{10}) \\ &= Cov(U_i^{01}, U_j^{01}) - Cov(U_i^{01}, U_j^{10}) - Cov(U_i^{10}, U_j^{01}) + Cov(U_i^{10}, U_j^{10}) \\ &= Var(U_{ir}^{01,01}) - Var(U_{ir}^{01,10}) - Var(U_{ir}^{10,01}) + Var(U_{ir}^{10,10}) \\ &= p(1-p)\{|\mathcal{S}_{ij}^{01,01}| - |\mathcal{S}_{ij}^{01,10}| - |\mathcal{S}_{ij}^{10,01}| + |\mathcal{S}_{ij}^{10,10}|\} \\ &= p(1-p)\boldsymbol{C}_{ir}, \end{aligned} \tag{1}$$

where $\boldsymbol{C}_{ir} = \{|\mathcal{S}_{ir}^{01,01}| - |\mathcal{S}_{ir}^{01,10}| - |\mathcal{S}_{ir}^{10,01}| + |\mathcal{S}_{ir}^{10,10}|\}$.

Naturally, the sizes of these sets, and therefore the correlation among measurement errors, depend on the overlap of their respective buffers which is related to their spatial proximity. In particular, the sets are empty for subjects whose buffers do not intersect with each other, nor with the same ZIP polygons; correspondingly the correlation is zero for those pairs of subjects. Thus, in large geographical extents, such as the state-wide analysis in the motivating data, $\boldsymbol{\Sigma}_{\text{U}}$ will be a sparse matrix.

In summary, $\boldsymbol{\Sigma}_{\text{U}} = p(1-p)\boldsymbol{C}$, with the entries of $\boldsymbol{C}$ given above. Given $\boldsymbol{\Sigma}_{\text{U}}$'s dependence on p, it is clear that replicating the coarsening process with increasing p (within the interval $(0, 0.5)$) simulates measurement error with *inflated* variance, even though we do not know the value of the measurement error variance for any one subject, $\sigma_{u,i}^2$. Also, even when manipulating only the coarsening proportion p, the heteroscedasticity of the measurement error variance will be preserved, since $\boldsymbol{C}$ is a constant. Clearly $p \to 0$ implies $\boldsymbol{\Sigma}_{\text{U}} \to 0$. Web Figure 4 shows the measurement error variance $\hat{\sigma}_{u,i}^2$ for each store type, for a single randomly selected school, whereas the figures in the main paper showed the total variance of measurement error ($\sigma_u^2(p)$, defined below).

Following analogous computations of the intersections of sets, the entries of the covariance matrix between the measurement errors and the true exposure, $\boldsymbol{\Sigma}_{\text{UX}}$, can be shown to be:

$$\begin{aligned} Cov(U_i, X_i) &= -p(1-p)\boldsymbol{E}_{ii} \\ Cov(U_i, X_r) &= -p(1-p)\boldsymbol{E}_{ir} \end{aligned}$$

Hence, $\boldsymbol{\Sigma}_{\text{UX}} = -p(1-p)\boldsymbol{E}$ where $\boldsymbol{E}$ is a square matrix with $\boldsymbol{E}_{ii} = |\mathcal{S}_i^{10}|$ in diagonal and $\boldsymbol{E}_{ir} = \{|\mathcal{S}_{ir}^{10,10}| + |\mathcal{S}_{ir}^{10,11}| - |\mathcal{S}_{ir}^{01,10}| - |\mathcal{S}_{ir}^{01,11}|\}$ in off-diagonal.

Because the measurement errors are heteroskedastic, it is also useful to define the following total variances and covariances, $\sigma_u^2(p)$ and $\sigma_{ux}^2(p)$. Let $\boldsymbol{U}$ be the vector of measurement

errors for all subjects, and $\overline{u} = \frac{1}{N}\mathbf{1}_N^T \boldsymbol{U}$ be the mean of the error across subjects. Then:

$$\begin{aligned}(N-1)\sigma_u^2(p) &= E[(\boldsymbol{U} - \mathbf{1}_N\overline{u})^{\mathrm{T}}(\boldsymbol{U} - \mathbf{1}_N\overline{u})] \\ &= E[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}], \quad where\ \boldsymbol{A} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathrm{T}} \\ &= tr\{\boldsymbol{A}\boldsymbol{\Sigma}_{\mathrm{U}}\} + \boldsymbol{\mu}_{\mathrm{U}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}_{\mathrm{U}} \\ &= p(1-p)tr\{\boldsymbol{A}\boldsymbol{C}\} + p^2\boldsymbol{d}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{d} \\ &\text{where } \boldsymbol{d} \text{ is a N-dimensional vector with } i\text{th entry } |\mathcal{S}_i^{01}| - |\mathcal{S}_i^{10}| \\ &= p(1-p)c_1 + p^2c_2.\end{aligned} \tag{2}$$

Similarly,

$$\begin{aligned}(N-1)\sigma_{ux}(p) &= E[(\boldsymbol{X} - \mathbf{1}_N\overline{x})^{\mathrm{T}}(\boldsymbol{U} - \mathbf{1}_N\overline{u})] = E[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}] \\ &= tr\{\boldsymbol{A}\boldsymbol{\Sigma}_{\mathrm{UX}}\} + \boldsymbol{\mu}_{\mathrm{X}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}_{\mathrm{U}} \\ &= -p(1-p)tr\{\boldsymbol{A}\boldsymbol{E}\} + p\boldsymbol{\mu}_{\mathrm{X}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{d} \\ &\text{where } \boldsymbol{\mu}_{\mathrm{X}} \text{ is a N-dimensional vector with } i\text{th entry } |\mathcal{S}_i^{11}| + |\mathcal{S}_i^{10}| \\ &= -p(1-p)c_3 + pc_4.\end{aligned} \tag{3}$$

The quantity $(N-1)\sigma_{uw}(p) = E[(\boldsymbol{U} - \mathbf{1}_N\overline{u})^{\mathrm{T}}(\boldsymbol{W} - \mathbf{1}_N\overline{w})]$ can similarly be shown to equal $(N-1)[\sigma_u^2(p) + \sigma_{ux}(p)]$.

A.2. Bias in naïve estimate

We use a simple linear regression model to analytically illustrate that bias can be positive or negative. Suppose the model of interest is

$$Y_i = \beta_0 + \beta_x X_i + \epsilon_i, i = 1, \cdots, N \tag{4}$$

where ϵ_i is independent of X_i and the measurement error U_i is independent of Y_i given X_i. Instead of unobservable true exposure X_i, we have $W_i = X_i + U_i$. Then,

$$plim\widehat{\beta}_w(p) = \frac{\sigma_x^2 + \sigma_{ux}(p)}{\sigma_x^2 + \sigma_u^2(p) + 2\sigma_{ux}(p)}\beta_x, \tag{5}$$

with the quantities $\sigma_u^2(p)$ and $\sigma_{ux}(p)$ as defined above. The total variance of X is similarly written as $(N-1)\sigma_x^2 = E[(\mathbf{x} - \mathbf{1}_N\overline{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{1}_N\overline{x})]$. Given the quadratic nature of the dependence of $\sigma_u^2(p)$ and $\sigma_{ux}(p)$ on p, it is obvious that $\sigma_u^2(p) \to 0$ and $\sigma_{ux}(p) \to 0$ as $p \to 0$, and that $plim\widehat{\beta}_w(p)$ is a continuous, differentiable function of p. Therefore, the naïve estimates will follow a predictable pattern that can be fitted with a smooth extrapolation curve. When evaluated at $p = 0$, the extrapolation will yield an unbiased estimate of the β_x.

Equation (5) shows that the direction of bias in the naïve coefficient estimate can be positive or negative. Whereas in the classical error assumption the independence between the measurement error and the true covariate makes the reliability ratio <1, in our case the reliability ratio, $\frac{\sigma_x^2+\sigma_{ux}(p)}{\sigma_x^2+\sigma_u^2(p)+2\sigma_{ux}(p)}$, can be greater or less than one due to non-zero $\sigma_{ux}(p)$. The reliability ratio is greater than 1 (i.e., bias away from the null) when $\sigma_{uw}(p) = \sigma_u^2(p) + \sigma_{ux}(p) < 0$, and less than 1 otherwise (attenuation bias). Whether the inequality $\sigma_u^2(p) + \sigma_{ux}(p) < 0$ is satisfied depends on p (which can vary), and the constants c_1,..., c_4, which are fixed. Thus, in general, the direction of bias could be different for different values of p, as shown for three of the outlets in Figure 1(b) of the main paper where the bias changes direction from first being positive and subsequently negative. However, when stratifying subjects according to the presence of ZIP centroids in the subject's buffers, the reliability

ratio in (5) has a more predictable pattern as a function of p, as shown empirically in Figure 3 of the main paper.

For the strata without ZIP centroids in their buffer, the bias is always away from the null independent of p. This is because when subject i and r do not have ZIP centroids in the buffer, $\mathcal{S}_i^{01}, \mathcal{S}_i^{11}$, $\mathcal{S}_r^{01}$, and $\mathcal{S}_r^{11}$ are empty sets. Thus,

$$\begin{aligned} \boldsymbol{C}_{ii} &= |\mathcal{S}_i^{10}|, \quad & \boldsymbol{C}_{ir} &= |\mathcal{S}_{ir}^{10,10}|, \\ d_i &= -|\mathcal{S}_i^{10}|, \quad & \boldsymbol{\mu}_{\mathrm{x}i} &= |\mathcal{S}_i^{10}|, \\ \boldsymbol{E}_{ii} &= |\mathcal{S}_i^{10}|, \quad & \boldsymbol{E}_{ir} &= |\mathcal{S}_{ir}^{10,10}|. \end{aligned}$$

From this, it can be shown that, $c_1 = c_3$, $c_2 > 0$ and $c_4 = -c_2$. Thus,

$$\begin{aligned} (N-1)(\sigma_u^2(p) + \sigma_{ux}(p)) &= (c_1 - c_3)p(1-p) + c_2 p^2 + c_4 p \\ &= p(p-1)c_2 < 0 \qquad \text{for } 0 < p < 1. \end{aligned}$$

Therefore, the naïve coefficient estimate is inflated when subjects do not have any ZIP centroids in their buffers. Moreover, it can be shown that the reliability ratio is monotonic increasing in p. Specifically, the first derivative of the reliability ratio evaluated at $p = 0$ is positive, thus the reliability ratio is increasing for at least small values of p. In addition, it can be shown that the derivative of the reliability ratio is a quotient with a quadratic polynomial in the numerator, which has roots at $p = \pm 1$. Thus, the derivative of the reliability ratio is positive for $0 \leq p < 1$.