

Statistical Methods for Exposure Measurement Error in Built Environment Studies

by

Jung Yeon Won

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2022

Doctoral Committee:

Professor Michael R. Elliott, Co-Chair
Professor Brisa N. Sánchez, Co-Chair
Professor Natalie Colabianchi
Assistant Professor Xu Shi

Jung Yeon Won

jungywon@umich.edu

ORCID iD: [0000-0002-8577-7307](https://orcid.org/0000-0002-8577-7307)

© Jung Yeon Won 2022

ACKNOWLEDGEMENTS

This work would not have been possible without many people over the course of my graduate studies. In particular, I express my deepest gratitude to my co-advisors, Professor Michael R. Elliott and Professor Brisa N. Sánchez, for their extensive support and guidance. I acknowledge their profound impact on my personal and professional development, and I am immensely grateful for their mentorship. Their incisive advice, generosity, and infinite patience have been the backbone of my Ph.D. training and made me move along the way to this final product. I am also deeply indebted to my committee, Professors Natalie Colabianchi and Xu Shi, for their time, advice, and contributions to improve my work.

It has been a great privilege to be a part of Department of Biostatistics at the University of Michigan. I have had many kind, motivated, and brilliant faculty, staff, and friends in the department who enriched my experience and made my time in Michigan so memorable. I have been very fortunate to have had my loving cohort, and the past and present office-mates who shared the peaks and valleys of the doctoral journey. Last but not least, I would like to express my warm and heartfelt appreciation to my husband Byungjun for his love and unwavering support. Thanks to my beloved daughter, Yubin, for being there. Ann Arbor witnessed so much of our precious moments. I also owe a significant debt of gratitude to my parents and parents-in-law for their endless encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	vi
LIST OF TABLES	x
LIST OF APPENDICES	xiii
LIST OF ABBREVIATIONS	xiv
ABSTRACT	xv
CHAPTER	
I. Introduction	1
1.1 Motivation	1
1.2 Sources of measurement error in built-environment studies . .	5
1.2.1 Geocode coarsening	5
1.2.2 Faulty ascertainment of environment features	5
1.3 Objectives	6
II. Split and Combine SIMEX Algorithm To Correct Geocoding Coarsening of Built Environment Exposures	10
2.1 Introduction	10
2.2 Food environment data and measurement error structure . . .	14
2.2.1 Food environment near California public schools . .	14
2.2.2 Exposure measure and measurement error distribution	16
2.2.3 Empirical illustration of measurement error distribu-	
tion	20
2.3 Split-Combine SIMEX	25
2.3.1 Notation and assumptions	25
2.3.2 Brief review of classic SIMEX	25

2.3.3	Proposed split-combine SIMEX method (SC-SIMEX) for geocode coarsened data	26
2.4	Bias analysis and Simulation Study	30
2.4.1	Different patterns of bias by presence of ZIP centroid in the buffer	31
2.4.2	Comparison of methods	33
2.4.3	Impact of the spatial distribution of subjects and out- let locations	35
2.5	Application to Fitnessgram Data	35
2.6	Discussion	38
III. Integrating Multiple Built Environment Data Sources		44
3.1	Introduction	44
3.2	Food exposure in multiple sources	46
3.3	Modeling approach	49
3.3.1	Notation for measurement error model	49
3.3.2	Modeling mixture distribution of the latent exposure counts	51
3.4	Estimation	54
3.5	Simulations and bias analysis	55
3.5.1	Bias in OLS	56
3.5.2	Simulation design	57
3.5.3	Simulation results for naïve bias	59
3.5.4	Evaluation of proposed model	61
3.6	Application to children’s BMI study in California	66
3.7	Discussion	70
IV. Estimating the Effect of Latent Time-Varying Count Expo- sures Using Multiple Lists		75
4.1	Introduction	75
4.2	Exposure measurements using commercial lists	78
4.3	Modeling strategy	80
4.3.1	Measurement error model	81
4.3.2	Dependencies in true exposures	83
4.3.3	Model for time-varying source quality	87
4.4	Simulation study	88
4.4.1	Simulation results	91
4.5	Childhood BMI study in California from 2001 to 2008	93
4.6	Discussion	97
V. Discussion and Future Work		99

APPENDICES	102
A.1 Additional measurement error and bias analysis	103
A.2 Additional data analysis results	108
A.3 Additional simulation results	109
A.4 Derivation of measurement error properties	113
A.4.1 Bias in naïve estimate	116
B.1 Detailed proof for Section 3.3	119
B.2 Detailed estimation steps for Section 3.4	119
B.2.1 Prior and Initial Values	119
B.2.2 Details of Dirichlet Process Mixture model	120
B.2.3 Algorithm outline	121
B.3 Detailed proof for Section 3.5	123
B.4 Relationship between naïve bias and Sensitivity/PPV	128
B.5 Model evaluation for β_0	129
B.6 Results of the model with single Gamma prior for λ_i	130
B.7 Results of the model with $(sen_1 > ppv_1, sen_2 < ppv_2)$	131
B.8 Graphical posterior predictive check of Section 3.6	132
C.1 Proof of marginal distribution of X_{it}	134
C.2 The Markov chain Monte Carlo sampler	134
C.3 Simulation details	138
C.4 Application results	140
BIBLIOGRAPHY	143

LIST OF FIGURES

Figure

2.1	(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 2.4 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=4,348 schools).	13
2.2	(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 2.1c for outlet abbreviations.	23
2.3	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 1,000 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 non-chain (FFN).	32

2.4	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 sub-panel represents the results of separate models, one for each food outlet type, among a sample of Asian children attending California public schools. See Figure 2.1 for outlet abbreviations.	43
3.1	Distributions of exposure to convenience stores and grocery stores from different sources. Exposures are computed by counting the number of stores within 1 mile buffer around California public schools. Means and standard deviations are presented in the legend.	48
3.2	Percent bias of slope (β_x) estimates from naïve regression models (“OLS, source 1”, “OLS, source 2”, “OLS, average”) under the different X distribution, the ratio of sensitivity to PPV, and the magnitude of agreement statistics. Values for <i>realistic</i> (labeled as R in the legend) sensitivity and PPV are (a) $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, $ppv_2 = 0.62$, and (b) $sen_1 = 0.48$, $sen_2 = 0.62$, $ppv_1 = 0.37$, $ppv_2 = 0.5$	60
3.3	Comparison of naïve OLS estimators for β_x and posterior mean of β_x from the proposed model under different misspecification. For naïve approaches, “OLS, InfoUSA” and “OLS, NETS” denote OLS estimators for simple linear regression that use $\mathbf{w}_{\text{InfoUSA}}$ and $\mathbf{w}_{\text{NETS}}$, respectively. “OLS, average” denotes an OLS estimator for a regression using average of $\mathbf{w}_1$ and $\mathbf{w}_2$. For the proposed model, reference values (denoted as “Proposed”) and reference values $\times 1.25$ (denoted as “Proposed $\times 1.25$ ”) are used for working credibility. Plotted are the naïve estimates with 95% confidence interval (black unfilled dots) and the posterior means with 95% credible intervals from our proposed model with reference values (red filled dot), and with $1.25 \times$ reference values (black filled dot).	69
3.4	Comparison of naïve OLS estimators for β_x and posterior mean of β_x from the proposed model without data stratification. See caption of Figure 3.3 for details of each model.	70
4.1	Annual distribution of exposure to convenience stores between the calendar year 2001 to 2008, for N=1,834 urban schools in California, using NETS and InfoUSA.	80

4.2	The number of convenience stores within 1-mile buffer by year, for randomly sampled five schools. Each school is represented by a colored line. The convenience store is defined as establishments with SIC code 541102 (convenience stores), 55410000 (gasoline service stations), 55419901 (filling stations, gasoline), and 55419903 (truck stops) are treated as convenience stores in the NETS database. For the InfoUSA database, convenience stores are defined as establishments that have NAICS codes 541103 (convenience stores), 554101 (service stations—gasoline and oil), 554102 (gas—diesel), and 554103 (truck stops and plazas).	80
4.3	Comparison of estimated change in mean standardized BMI per doubling convenience stores around 1-mile buffer from a school between 2001 and 2008 after fitting the proposed model, naïve model with single-sourced exposure (NETS, InfoUSA), and naïve model with averaged exposure (Average). Stratified analyses are conducted by grade (5th, 7th) and sex. Plotted are posterior means (filled circle dots) with 95% credible intervals for the proposed model, and generalized least square estimates (open triangle dots) with 95% confidence intervals for naïve models.	96
4.4	Estimated difference in mean standardized BMI per doubling convenience stores around schools between 2001 and 2008, using unstratified analysis. See caption of Figure 4.3 for the description of methods.	97
A.1	Illustration of measurement error mechanism for subjects with and without ZIP centroid within their buffer	103
A.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.	104
A.3	Strata-specific histogram of measurement error mean, among 12,370 schools when coarsening probability is $p = 5\%$ and $p = 25\%$ (subfigures (a)-(x)). 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.	105
A.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.	107

B.1	Percent bias of slope (β_x) estimates from single-source naïve regression models $OLS_s, s = 1$ by the ratio of sensitivity to PPV, and the magnitude of agreement statistics. Grayscale lines represent different X distributions that are defined in Section 3.5.2 For each X distribution, source is created with different values of sen_s , holding ppv_s at 0.8, 0.5, or 0.2. Note that when sensitivity is greater than PPV (i.e., $\frac{sen_s}{ppv_s} > 1$), naïve estimate for the slope is always attenuated as described in Section 3.5.4 and displayed in Figure 3.2(b), regardless of values of agreement statistics and X distribution. When X is equidispersed (Mixture 1), naïve bias only depends on the value of PPV (Section 3.5).	128
B.2	Distribution of observed exposures W_1 (NETS) and W_2 (InfoUSA), and density of 500 replicated W s from a fitted model (lightblue lines). Subfigures (a-d) show exposures to convenience stores for 5th and 7th-grade schools; (e-h) show exposures to grocery stores. Subfigures (a, b, e, f) are from our proposed model with the reference values and (c, d, g, h) are from our proposed model with reference values $\times 1.25$. See Section 3.6 and Figure 3.3 for reference values of agreement statistics.	132
C.1	Simulated true exposure $\{X_{it}\}$ for $t = 1, \dots, 8$ and $N = 300$ with different scenarios of p . The value of individual expected baseline exposure is either $\lambda_{i1} = 5$ or $\lambda_{i1} = 15$ with the probability 0.9 and 0.1, respectively. Exposures of subjects with $\lambda_{i1} = 5$ are light gray curves and exposures of subjects with $\lambda_{i1} = 15$ are colored in dark gray. Arrival rate μ_i takes a value of 0.5 or 1.5, with equal probabilities. Magenta line shows the curve of annual sample mean of X_{it} by the combination of arrival rate and expected baseline exposure.	138
C.2	True sensitivities and PPVs (red vertical lines) when $T = 8$. Prior with correct information ($sen_T^A, ppv_T^A, sen_T^B, ppv_T^B$) matches the data generating distribution (black curve). Misspecified prior (blue curve) shows incorrect prior beliefs about source qualities.	139
C.3	Annual distribution of <i>observed</i> exposures to convenience stores from 2001 to 2008 using NETS and InfoUSA, for 5th grade female students who attend urban schools in California. Black curve is an empirical density of <i>observed</i> exposures, which approximates the true density function using <code>density()</code> in R. Red curves are empirical density of posterior predictive W 's the proposed model that we estimate sen_t^s and ppv_t^s for $t = 1, \dots, T, s = A, B$ as described in Section 4.3. Blue curves represent the empirical density of posterior predictive W 's from the model that fixed sen_t^s and ppv_t^s to the reference values without any updates so that $(sen_t^{NETS}, ppv_t^{NETS}) = (0.3769, 0.4812)$ and $(sen_t^{InfoUSA}, ppv_t^{InfoUSA}) = (0.5047, 0.6182)$ for all t . This figure shows the need of estimating source credibility for historical data. . .	141

LIST OF TABLES

Table

2.1	Average of variances of measurement errors U_{ik} , observed exposures, W_i , and true exposures, X_{ik} , for $k = \text{GRO, CON, BAK, NAL, FFC, FFN}$, and correlations them, in the presence of 15% of geocode coarsening for each business type. Values are based on 1,000 Monte Carlo simulations, among $N=1,000$ randomly selected schools.	24
2.2	Simulation results comparing naïve, traditional SIMEX, adjusted SIMEX, and split-combine SIMEX; Weighted cubic regression is used for extrapolation $f(\cdot)$. We used randomly sampled $N=1,000$ schools. The results are from 1,000 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 non-chain (FFN), Grocery (GRO), Fast food chain (FFC), Non-Alcohol drinking place (NAL).	34
2.3	Multiple outlet model. 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 2.1 for outlet abbreviations. . .	39
3.1	Contingency tables comparing exposure measures between truth and claims from two databases.	49
3.2	Different distributions of X with the same expectation for the simulation study.	58
3.3	($sen_1 < ppv_1, sen_2 < ppv_2$) Posterior summaries of the proposed model for β_x when true source credibility is $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. True $\beta_x = 0.5$. The proposed model is evaluated with and without estimating X distribution. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications	63

3.4	($sen_1 > ppv_1, sen_2 > ppv_2$) Posterior summaries of the proposed model for β_x when true source credibility is $sen_1 = 0.48$, $sen_2 = 0.62$, $ppv_1 = 0.37$, and $ppv_2 = 0.5$. True $\beta_x = 0.5$	65
4.1	Contingency tables comparing truth and observed exposures from two databases ($s = A, B$) for subject i at time t . Each table has TP_{it}^s (true positive of subject i at time t from source s), FN_{it}^s (false negative), FP_{it}^s (false positive), and corresponding cell probabilities ($\theta_{11t}^s, \theta_{12t}^s, \theta_{21t}^s$).	81
4.2	Summaries of the sampling distribution of β_x based on 50 simulated data sets. Root mean square error (RMSE), percent bias, and coverage probability (CP) are presented for each scenario. $N=300$ sites, $T=\{4,8\}$, and business dynamics $p = \{0.99, 0.8, 0.5\}$, $\mu_i = (0.5, 1.5)$, and $\lambda_{i1} = (5, 15)$. Regression parameters are $\beta_0 = 0.85$, $\beta_x = \{0.05, 0.5\}$ $\sigma = 0.35$, and $\rho = 0.5$	91
A.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.	108
A.2	Extension of Table 2.2, 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 $S=30$	109
A.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 2.2, 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 $S=30$	110
A.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 2.2, 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 $S=30$	111

A.5	Sample standard deviation of the estimates (SE) and the average of bootstrapped standard error estimates (ESE) of SIMEX methods in Table 2.2. 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.	112
B.1	($sen_1 < ppv_1, sen_2 < ppv_2$) Posterior summaries of the proposed model for β_0 when true source reliability is $sen_1 = 0.37$, $sen_2 = 0.4$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. True $\beta_0 = 1.5$. The proposed model is evaluated with and without estimating X distribution. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications	129
B.2	(Same data as Table 3.3) Posterior summaries of the proposed model for β_x when Gamma(1,1) prior is assumed for all $\lambda_i = \lambda$ instead of Dirichlet process prior. True $\beta_x = 0.5$ and the true source credibility is $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications	130
B.3	($sen_1 > ppv_1, sen_2 < ppv_2$) Posterior summaries of the proposed model for β_x when true source credibility is $sen_1 = 0.48$, $sen_2 = 0.5$, $ppv_1 = 0.37$, and $ppv_2 = 0.62$. True $\beta_x = 0.5$. The proposed model is evaluated with unknown X distribution. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications	131
C.1	Posterior summaries of variables from Childhood BMI study in Section 4.5.	140

LIST OF APPENDICES

Appendix

A.	Chapter II Supplementary Materials	103
B.	Chapter III Supplementary Materials	119
C.	Chapter IV Supplementary Materials	134

LIST OF ABBREVIATIONS

GIS	Geographic Information System
SIMEX	Simulation Extrapolation
SC-SIMEX	Split-Combine SIMEX
NETS	National Establishment Time-series
GRO	Grocery store
CON	Convenience store
BAK	Bakery
NAL	Non-Alcohol drinking place
FFC	Fast food chain
FFN	Fast food non-chain
SIC	Standard Industrial Classification
NAICS	North American Industry Classification System
OLS	Ordinary Least Squares
DP	Dirichlet Process
CRP	Chinese Restaurant Process

ABSTRACT

Collecting information about the built environment is a challenging process and often leads to measures of environmental exposures that are error-prone and thus result in biased inference. While the use of GIS (Geographic Information System) databases is growing exponentially in health research, there is little guidance on strategies to handle erroneous exposure measures specific to built environment data. Our motivating examples focus on the environmental exposure that is expressed as the number of food outlets (e.g., fast food restaurants) within a circular buffer around the subjects' locations. Measurement error in the exposure can mislead inferences about the impact of the availability of environmental features on health, which is the key factor in place-based strategies or policies to improve public health. This dissertation focuses on developing methods to address measurement challenges in studies of built environment health effects, namely measurement error and data integration.

In Chapter II, we propose a split-combine simulation extrapolation (SC-SIMEX) method that accommodates non-classical measurement error due to incorrect geocodes without requiring external data. The standard SIMEX (Cook and Stefanski, 1994) is widely used to correct the effect of measurement error on regression problems under distributional assumptions on errors. However, the exposure measurement errors that arise due to geocode coarsening have a novel distribution compared to the commonly used distributions in existing measurement error literature. The proposed SC-SIMEX relaxes the measurement error assumptions required for the standard SIMEX. The utility of SC-SIMEX is demonstrated in a study of children's obesity in California in

relation to the junk food environment.

In Chapter III, we develop a multi-source measurement error model to effectively integrate different data sources using external knowledge of source credibility. Secondary commercial lists often provide conflicting claims about the presence and location of businesses. Consequently, disagreement among exposures collected from different sources gives different inferences. Our work shows that bias in the naïve regression model is dependent upon the dispersion of true exposure and the source credibility. Knowing the reliability of databases from field validation studies, the proposed measurement error model can derive complete information about latent true exposures using incomplete data tables with partially known margins. Our method uses a Bayesian nonparametric model that makes few distributional assumptions about the counts of businesses in a region and handles the overdispersion in exposure counts. The application assesses the association between children’s BMI in urban schools and exposure to nearby convenience and grocery stores.

Chapter IV extends the data integration methodology in Chapter III to the time-varying setting. For example, commercial business lists often provide annual business listings, which enables us to compute time-varying exposures. We extend the method described in Chapter III to flexibly model the latent time-dependences in true exposures using integer-valued autoregressive process. We aim to estimate the dynamic latent exposures and reduce bias in longitudinal analyses of the exposure health effect.

The dissertation concludes with a discussion of future work and the wider implications of the proposed methods. Our methods help future built-environment studies where measurement errors are inevitable and their impacts are little known. This dissertation contributes to improved understanding of measurement error properties that need to be addressed when using built-environment databases and provides novel methods to overcome the bias in epidemiologic analyses.

CHAPTER I

Introduction

1.1 Motivation

The advance of Geographic Information Systems (GIS) technology has accelerated public health research that evaluates the role of the built environment in human health. GIS technology has been acknowledged as one of six innovations on the frontier of social science research for its contribution to quantifying the built environment (Butz and Torrey, 2006). GIS data are especially important in health sciences that study the contextual attributes of neighborhoods in relation to physical and mental health (Pearce et al., 2006; Thornton et al., 2011; D’Angelo et al., 2014; Fletcher-Lartey and Caprarelli, 2016).

While primary data about built-environment features that are gathered by the researcher offer the most accurate information about environmental characteristics of the study area, the data collecting process is often expensive. Alternatively, the majority of epidemiologic studies (Fleischhacker et al., 2013; Wong et al., 2018; Zhou et al., 2021) rely on secondary GIS databases about built-environment features from private providers such as InfoUSA (InfoUSA, 2021) and Dun and Bradstreet (Dun and Bradstreet, 2021). These commercial databases collect nation-wide establishments and their characteristics such as geocodes, addresses, outlet types, sales volume, number of employees, and years of opening and closing, and are updated on an

annual basis. Typically, business coordinates and business type classifications from these databases are used to measure the density and proximity to built-environment features around study subjects (Zhou et al., 2021). However, these commercial business databases are primarily published on the purpose of economic analyses or marketing tools, and are not designed for health research. Therefore, accuracy of the data is a key factor to correctly quantify environmental exposures in order to use such commercial databases in health studies. Prior research suggests that the secondary data have sufficient error to mislead inferences of analyses (Cummins and Macintyre, 2009; Thornton et al., 2011; Jacquez, 2012). Despite these concerns, little attention has been given to statistical measurement error corrections, in the context of exposures to the built environment.

The effect of measurement errors in the predictors in regression problems is a long-standing problem that has been studied by a variety of measurement error models. Using mismeasured predictors is known to cause bias in the estimated regression coefficients and raise a question about validity of inferences (Armstrong, 1985; Stefanski, 1985; Carroll et al., 2006). The identification and estimation of measurement error models are generally based on distributional assumptions about the measurement error or availability of instrumental or repeatedly measured data (Schennach, 2016; Yi, 2017). A large number of measurement error models have been developed around the methods with distributional assumptions on the measurement error, known as classical additive measurement errors (Rosner et al., 1989; Carroll and Stefanski, 1990). The measurement error, denoted as U , is *classical* when the error is independent of the latent true variable X and *additive* when the error is assumed to be a difference between the observed variable W and the true X (i.e., $W = X + U$) (Carroll et al., 2006; Carroll, 1998). In general, this error term is assumed to be continuous and normally distributed with zero mean and variance σ_u^2 . Under classical additive measurement error assumption, the ordinary least squares regression of Y on W for the

slope coefficient β_x in the simple linear regression $Y = \beta_0 + \beta_x X + \epsilon$ yields an attenuated estimate. Using observed W in place of X produces the ordinary least squares estimator $\hat{\beta}_x = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \beta_x$. This estimator underestimates the true coefficient value by a factor (also known as *reliability ratio*) which is always less than 1 since $\sigma_u^2 > 0$. There are also discrete types of errors, often named as misclassification errors, in dichotomous or categorical data (Greenland, 1980; Spiegelman et al., 2000; Buonaccorsi et al., 2005). General strategies for misclassifications are based on specifying *misclassification probabilities*, namely $Pr(W = w|X = x)$. For binary data, logistic regression models are used to characterize misclassifications with covariates. When a discrete variable is measured with discrete error, the errors are often nonclassical and have complex distributions. Zero-mean error is not possible due to the constrained support of the error and the independence assumption between the errors and true values is violated (Greenland, 2000). Therefore, conventional measurement error models are not suitable for discrete measurement errors. Schennach (2016) reviewed existing studies to handle an effect of having misclassification errors in regressors or response data. Among discrete type of measurement error, we focus on errors in count covariates because built-environment health studies often use counting measures to quantify exposures (e.g., number of junk food stores in the neighborhood).

For count type errors, dispersion in counts should be carefully considered when making assumptions on conditional distribution of observed data given true values. Li et al. (2004) provided corrected score equations for the regression coefficients when a count surrogate W_i during constant time L_i is Poisson distributed with mean $X_i L_i$. In this setting, $E(W_i|X_i) = Var(W_i|X_i) = X_i L_i$. Then, consistent estimators of parameters θ given observed W and a covariate Z can be obtained by maximizing the “corrected” log-likelihood $L^*(\theta; Y, W, Z)$ such that $E(L^*(\theta; Y, W, Z)|Y, X, Z) = L(\theta; Y, X, Z)$. The authors show that this unbiased log-likelihood includes a correction term to adjust for the naïve bias. This method only considers one surrogate and

requires Poisson assumption of W given X .

To allow smaller or larger variations in observed count than the true count, Osthus et al. (2021) used the latent log normal distribution. The observed W_i is defined as discretized version of $\widetilde{W}_i$ that has a log normal distribution with mean X_i and variance αX_i , where α is a factor that deflates or inflates a standard deviation. Authors applied this idea to repeated measurements by multiple observers and incorporated j th observer’s bias γ_j as a multiplicative error so that $E(\widetilde{W}_{ij}|X_i) = X_i\gamma_j$. All parameters are estimated under Bayesian method and the deflation/inflation factor α was assumed to be constant across all observed data.

Zero-inflated distributions are also widely used in ecological data to account for the excess absence in observations (Martin et al., 2005; Wenger and Freeman, 2008; Denes et al., 2015). In ecological studies, the detection of animal may fail and a false absence is reported. In this case, modeling process involves the probability of presence of species at i th site, denoted as p_i , and the detection probability at a visit, denoted as q . When the true occupancy at the i th site is X_i , observed count abundance for this site at the v th visit is expressed as $W_{iv} \sim \text{Binom}(N_i, q)$, $N_i = I(\text{presence}_i)X_i$, $I(\text{presence}_i) \sim \text{Bern}(p_i)$ with $X_i \sim \text{Pois}(\lambda_i)$. This is an alternative representation of zero-inflated Poisson model. Standard logistic regression or Bayesian approach produces an estimate of occupancy probability p and the true site occupancy.

Despite these efforts to deal with mismeasured count covariates, many epidemiologic analyses of the built environment using count exposure measures still rely on resource-intensive data cleaning or field observations to directly reduce the measurement error, or simply use oversimplified heuristics that measurement error attenuates the regression coefficient estimates (van Smeden et al., 2020).

To address measurement errors in built-environment studies, this dissertation focuses on two sources of errors that are embedded in the secondary databases: locational errors (Liese et al., 2010; Boone et al., 2008) and inaccurate collection of

establishments, i.e., ascertainment errors (Powell et al., 2011; Jones et al., 2017; Liese et al., 2010). Both types of errors in the database hamper exposure measurement precision and affect the validity of analyses.

1.2 Sources of measurement error in built-environment studies

1.2.1 Geocode coarsening

When establishment data are collected by the commercial data companies, the geocode of each business is obtained along with its accuracy level as part of an automated geocoding process. Coordinates of businesses with lower geocode accuracy level are assigned to area-level coordinates such as centroids of census-tracts or ZIP Codes (InfoUSA, 2021; Dun and Bradstreet, 2021). For example, businesses with the highest geocode accuracy will have latitude and longitude at a street level whereas businesses with the lowest geocode accuracy are matched to the centroid of their ZIP Codes. Although most businesses have recently had accurate geocodes at the street level, businesses that existed in the past are more subjective to the geocoding errors (Kaufman et al., 2015a). This error-prone locational information produces errors in exposure measurements. However, existing methods designed for data with geocoding inaccuracies have not examined bias in health effect estimates of the built-environment (Zimmerman, 2008; Zimmerman et al., 2010) or suggested significant re-geocoding efforts (Sonderman et al., 2012) to reduce the prevalence of inaccurate geocodes.

1.2.2 Faulty ascertainment of environment features

Ascertainment errors arise due to incomplete listings with mistaken or omitted features. For example, commercial databases may list closed or non-existing businesses

or falsely omit truly existing businesses. Incorrect listings of businesses leads us to undercount or overcount the prevalence of businesses in a given area. To tackle this problem, researchers can verify the commercial data by validating the business list using gold-standard data such as administrative data or ground observations (Powell et al., 2011; Liese et al., 2013). Alternatively, manual data cleaning to remove faulty listings or merging different databases are used to achieve a high agreement with primary data (Jones et al., 2017; Caspi and Friebur, 2016). While these suggested methods may produce better quality data, they are labor-intensive and not feasible in retrospective studies that use historical databases.

1.3 Objectives

This dissertation aims to develop statistical methods to obtain valid inferences in regression models that use count exposures that are measured with errors in the context of built-environment studies. We examine the distributions of the integer-valued measurement errors that are not standard in the existing measurement error literature. To our knowledge, there is no statistical and epidemiological studies that studied the extent of bias in effect estimation when we have count measurement errors in regressors. This dissertation quantifies how much bias we have if we ignore the measurement errors and develops methods to correct the effect of measurement error. We focus on measurement errors from the two aforementioned sources: geocoding errors and inaccurate ascertainment of features in large-scale built-environment databases. In Chapter II, inspired by standard simulation-extrapolation method (Cook and Stefanski, 1994), we propose a split-combine simulation extrapolation (SC-SIMEX) method to solve this problem without requiring validation data or re-geocoding. With motivating examples that focus on childhood obesity and exposure to the nearby food environment, we illustrate the pattern of naïve coefficient estimates by degree of geocoding errors in the database and how measurement error distribution can be de-

composed by the geographic location of the subject. Given our ability to replicate the measurement error mechanism by the computer simulation, the SC-SIMEX method addresses non-classical measurement errors due to incorrect geocodes and produces a corrected inference. We show that the SC-SIMEX does not require parametric distribution of latent true exposure or the distribution of measurement error.

In Chapter III and IV, we suggest a new approach to alleviate the ascertainment error problem by combining multiple databases. In Chapter III, we develop a joint model using multiple commercial databases to accomodate non-classical integer-valued errors in the count exposures. Although true number of environment features in the subject’s neighborhood is unknown, we have error-prone measures from multiple commercial databases. The key of this joint modeling is how we can integrate disagreeing measures from different sources using external knowledge of their source credibility. A trinomial distribution is used to construct a measurement error model so that true positives, false negatives, and false positives can be simultaneously modeled and the source credibility can be used to compute each cell probability. Under a Poisson sampling framework, the trinomial distribution has a natural connection to the Poisson distributed count exposure at a local geographic level. We then consider varying expected value of true exposures by different subjects due to their unique neighborhood characteristics. To model this heterogeneity of exposures without supplemental information, a Bayesian nonparametric approach is introduced to model the mixture of Poisson distributions across the geographic areas. Bayesian nonparametric modeling for latent count variable has been frequently used in ecological applications where animal abundance entails considerable heterogeneity (Dorazio et al., 2008; Diana et al., 2018, 2020). The most popularly adopted Bayesian nonparametric model is the Dirichlet Process mixture model that gives infinite number of mixture components (Ferguson, 1973). Because the Dirichlet process prior learns information about number, size, and parameters of mixture components from the data, we can

flexibly model the unobserved true exposures and account for the overdispersions of count exposures. To the best of our knowledge, although many ecological studies have showed the successful use of Bayesian nonparametric priors to produce correct estimates of true exposure, there is no study that investigated its utility in bias reduction in regression analyses when the true exposure is used as a regressor. Therefore, in Chapter III, we present a bias analysis to show how naïve use of the secondary commercial databases leads to biased estimates of regression coefficients and how the proposed method can improve the results. The new model will be demonstrated with a study that examines the association between school-level obesity and the prevalence of junk food stores around schools in 2008. In this example, we used exposures computed from two commercial business listings that have different levels of source credibility.

Chapter IV builds upon Chapter III to correct the effect of time-varying measurement errors in longitudinal exposures. This extension can bring a considerable benefits to retrospective studies where ground observation is not feasible. The proposed model handles the time-dependencies in count exposures using an integer-valued autoregressive process (Alzaid and Al-Osh, 1988). Integer-valued autoregressive process has been widely used to model count time series data (Aldor-Noiman et al., 2016; Dail and Madsen, 2011) and this model fits our data well as the parameters have intuitive interpretations of business dynamics such as business openings and closings over the years. Bayesian modeling also allows us to estimate historical source qualities by using priors that are constructed with external information. Extending the model from Chapter IV, Dirichlet process priors are used to flexibly model both heterogeneous expected baseline exposures and exposure trajectories that can be expressed with the business dynamics around the subject’s neighborhood. Our method is shown to reduce bias in the estimate of the association between time-varying latent true exposures and time-varying health outcomes given the availability of historical

business databases. Our method is applied to an extended example of Chapter III with a study time period between 2001 and 2008.

CHAPTER II

Split and Combine SIMEX Algorithm To Correct Geocoding Coarsening of Built Environment Exposures

2.1 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 2.2, 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 (Zimmerman, 2008).

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 2.1a 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 improved over time or other technologies such as GPS could be used to characterize exposure, many longitudinal cohort studies, including MESA (Auchincloss et al., 2012; Diez Roux et al., 2016), REGARDS (Kaufman et al., 2015a; Lovasi, 2015), CARDIA (Hirsch et al., 2016), the Jackson Heart Study (Hickson et al., 2011), among others (Berger et al., 2019), use historical records dating back to 1990 to quantify cumulative exposure. Based on simulations, Figure 2.1b 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 (Jacquez, 2012). 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 (Zimmerman and Li, 2010; Zandbergen et al., 2012) and can impact exposure measures (Healy and Gilliland, 2012; Zandbergen, 2007). For area-level coarsening, Zimmerman (2008) 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 (Goldberg, 2008; Zandbergen et al., 2012), use of GPS has several limitations: it would

incur additional costs; it is likely to be feasible only in a subset of locations, i.e., a validation sample; and it is impossible for both historical datasets and those covering large geographical extents, like the cohorts described previously.

In this paper, 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 2.1b) 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 (Cook and Stefanski, 1994), 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 (Küchenhoff et al., 2006, 2007). Parveen et al. (2015) and Parveen et al. (2017) relax the assumptions of the measurement error distribution to include a covariate-dependent measure-

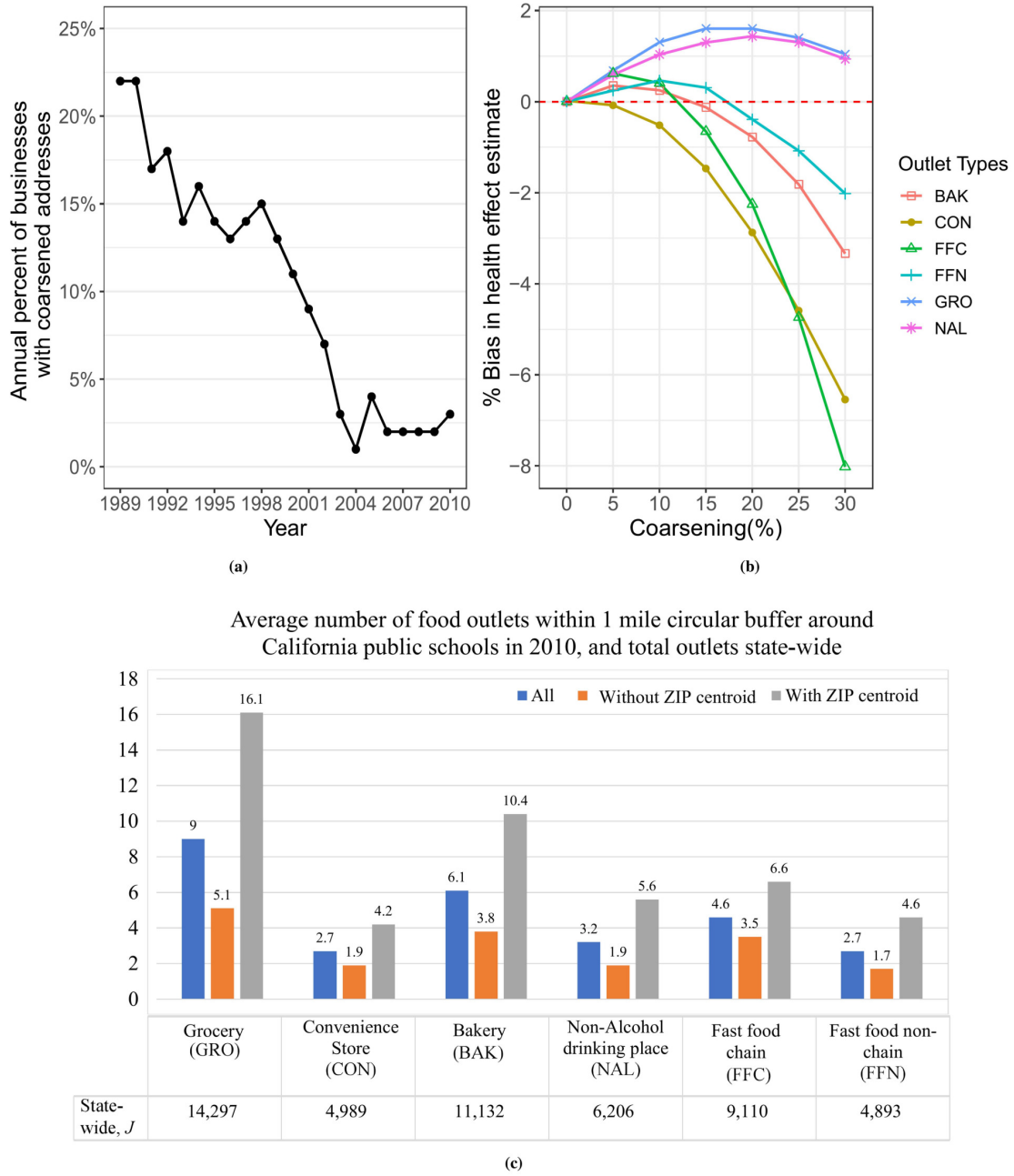


Figure 2.1: (a) Annual percent of food retail outlets (standard industry codes 54 and 58) in California with coarsened addresses by year of entry into the National Establishment Time-Series (NETS) database; (b) Percent bias in health effect estimates from naïve regression depends on coarsening level and food outlet type. Bias is either positive or negative since measurement error induced by coarsening is not classical measurement error (see Section 2.4 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 ($N=4,348$ schools).

ment error with a non-zero mean. 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 (Lockwood and McCaffrey, 2017). The SIMEX method has also been used in a spatial context to correct bias in the health effect of air pollution exposure (Alexeeff et al., 2016). 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 2.2, 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 2.3, we introduce the notation and the newly proposed method. The results of simulations used to evaluate the proposed method are included in Section 2.4, followed by the application of the methodology to the motivating study data in Section 2.5. Final remarks are given in Section 2.6.

2.2 Food environment data and measurement error structure

2.2.1 Food environment near California public schools

Several available data sources contain point-referenced data on commercial establishments (NETS, 2021; InfoUSA, 2021). One of these sources is the widely-used National Establishment Time-Series (NETS) database (NETS, 2021) 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) (Dun and Bradstreet, 2021). The information collected includes company names, addresses, sales, and Standard

Industrial Classification Codes (NAICS, 2019); this information is used to categorize each entity (e.g., 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 (e.g., 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 2.1a). In this paper, we use two years of NETS data for different purposes. For simulation experiments where the true geocodes are needed to produce error-free exposures (Section 2.2-2.4), we use 2010 NETS data, which has very high geocoding quality. In analyses for our motivating study (Section 2.5), 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 non-chain 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 2.1c).

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 (e.g., Athens et al. (2016)), 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 2.1c 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.

2.2.2 Exposure measure and measurement error distribution

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 (e.g. 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 i th subject, denoted by X_i , is the sum of indicators x_{ij} : $X_i = \sum_{j=1}^J x_{ij}$.

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 j th establishment, $j = 1, \dots, J$, which takes a value 1 if the j th 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 2.1a. Naturally, coarsened geocodes may introduce exposure measurement error.

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 A.4, 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 j th business in the neighborhood of the i th subject, we define indicators of whether its corresponding w_{ij} matches x_{ij} :

$$\begin{aligned} I_{ij}^{11} &= \mathbb{1}\{x_{ij} = 1, w_{ij} = 1\}, & I_{ij}^{00} &= \mathbb{1}\{x_{ij} = 0, w_{ij} = 0\} \\ I_{ij}^{10} &= \mathbb{1}\{x_{ij} = 1, w_{ij} = 0\}, & I_{ij}^{01} &= \mathbb{1}\{x_{ij} = 0, w_{ij} = 1\}. \end{aligned}$$

When the j th business's address has been accurately geocoded, i.e., $A_j = 1$, 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 j th 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 j th 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 i th 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 A.4, upon conditioning on the locations of schools, true addresses of businesses, the locations of ZIP centroids, and under the assumption of coarsening at random (i.e., $P(A_j = 0) = p$), the measurement error has $E(U_i) = (|\mathcal{S}_i^{01}| - |\mathcal{S}_i^{10}|)p$ and $Var(U_i) = (|\mathcal{S}_i^{01}| + |\mathcal{S}_i^{10}|)p(1 - p)$, where $\mathcal{S}_i^{01}$ is a set of businesses that, when coarsened, will be incorrectly included in the exposure count for subject i , and $\mathcal{S}_i^{10}$ 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.

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, i.e., $\sum_{j=1}^J (I_{ij}^{01}) = 0$, as depicted in Figure A.1a. Thus, the measurement error for subjects without ZIP centroids within their buffers will be non-positive. 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 (Figure A.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:

$$U_i = W_i - X_i = \begin{cases} \sum_{j=1}^J (-I_{ij}^{10}) \leq 0, & \text{no ZIP centroid within subject's buffer} \\ \sum_{j=1}^J (I_{ij}^{01} - I_{ij}^{10}), & \text{ZIP centroid within the subject's buffer.} \end{cases} \quad (2.1)$$

From (2.1), 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 A.4 and illustrated empirically in Table 2.1. In addition, $\text{Corr}(U_i, W_i) \neq 0$ 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 A.4).

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 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.

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 (e.g., fast food restaurants, convenience stores). Assuming each business type has m_q establishments for $q=1, \dots, Q$, the grand total number of outlets is $J = \sum_{q=1}^Q m_q$. Then, the i th subject's true exposure to the q th 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 q th business type for subject i is $U_{iq} = W_{iq} - X_{iq} = \sum_{j=1}^J (I_{ij}^{10} - I_{ij}^{11}) * I(\text{business } j \text{ is of type } q)$.

It is possible that a correlation exists between measurement errors of different business types (e.g., 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 2.1). 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 (Carroll et al., 2006). 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 2.1b and Figure A.2).

Carroll et al. (2006) 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 (2.1), 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.

2.2.3 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 2.1a), 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 (i.e., 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 2.1c 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.

Figures 2.2a and 2.2b illustrate 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 (i.e., 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—i.e., an under-count. Figure A.4 shows histograms of $E(U_i)$ for schools with and without ZIP centroids within their buffers.

Figures 2.2c and 2.2(d) show 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 2.1c), the variance of the measurement error is also higher among that subset of schools (Figure 2.2d).

Measurement errors also have different distributions by food store type, and the errors are correlated among different food store types. Figure 2.2 and Table 2.1 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 non-chain fast food restaurants. Table 2.1 also illustrates the previously described positive correlation among measurement errors across different food store types, as well as non-zero 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 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 2.2e and 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.

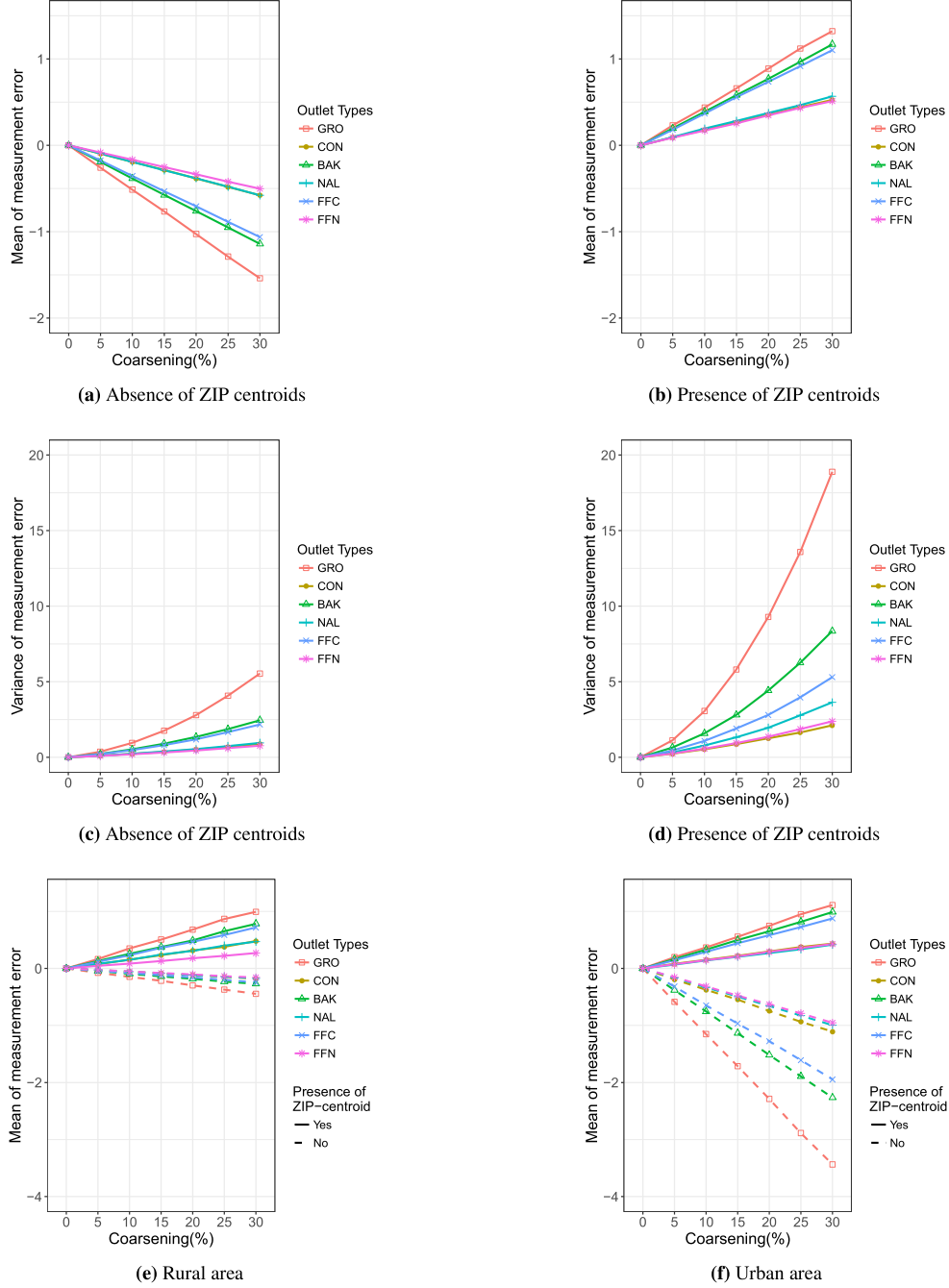


Figure 2.2: (a) - (d) Overall mean and variance of measurement errors by presence or absence of ZIP centroid within 1 mile of school, and (e)-(f) mean of measurement errors stratified by urbanization level. See Figure 2.1c for outlet abbreviations.

Table 2.1: Average of variances of measurement errors U_{ik} , observed exposures, W_i , and true exposures, X_{ik} , for $k = \text{GRO, CON, BAK, NAL, FFC, FFN}$, and correlations them, in the presence of 15% of geocode coarsening for each business type. Values are based on 1,000 Monte Carlo simulations, among $N=1,000$ randomly selected schools.

Food outlet type	$\text{Var}(U_k)$	$\text{Corr}(U_k, U_{k'})$					
		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 non-chain (FFN)	0.59	0.38	0.29	0.40	0.35	0.36	1

	$\text{Var}(W_k)$	$\text{Corr}(W_k, U_{k'})$					
		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 non-chain (FFN)	16.62	-0.13	-0.02	-0.06	-0.10	-0.01	-0.01

	$\text{Var}(X_k)$	$\text{Corr}(X_k, U_{k'})$					
		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 non-chain (FFN)	17.31	-0.20	-0.07	-0.14	-0.16	-0.07	-0.19

2.3 Split-Combine SIMEX

2.3.1 Notation and assumptions

Let $\mathbf{Y}$ be a vector of continuous outcomes for N subjects, which can be modeled with linear regression: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_x + \mathbf{Z}\boldsymbol{\beta}_z + \boldsymbol{\epsilon}$. The $N \times Q$ exposure matrix $\mathbf{X}$ is a matrix of true unmeasured exposure, where the q th column of $\mathbf{X}$ denotes subject's exposures to the q th environment feature, and X_{iq} is as described in Section 2.2.2. The $\mathbf{Z}$ is a matrix of error-free covariates including an intercept and a corresponding parameter is $\boldsymbol{\beta}_z$. The residual $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)^T$ is a vector of random errors with mean zero and variance σ_ϵ^2 . We define $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^T, \boldsymbol{\beta}_z^T)^T$ and our primary interest is estimating $\boldsymbol{\beta}_x$, the association between a continuous health outcome $\mathbf{Y}$ and unobserved count exposure $\mathbf{X}$.

In the presence of geocoding coarsening, we observe $\mathcal{W}$ instead of $\mathcal{X}$, and thus the analysis dataset will have observe data $(\mathbf{Y}, \mathbf{W}, \mathbf{Z}, \mathbf{A})$ instead of true data $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$. Here $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with measurement error $\mathbf{U}$, and $\mathbf{A} = (A_1, \dots, A_J)^T$, where A_j denotes whether the geocode for food outlet j has been coarsened, as described in Section 2.2. We propose a method to correct a bias in the estimation of $\boldsymbol{\beta}_x$ due to measurement error $\mathbf{U}$ incurred by geocode coarsening. We briefly describe the classic SIMEX method to enable us to draw analogies in the description of our proposed method.

2.3.2 Brief review of classic SIMEX

The simulation-extrapolation (SIMEX) procedure consists of two steps (Cook and Stefanski, 1994): First, in the simulation step, a computer-simulated error that mimics the measurement error is added to the observed $\{W_i\}_{i=1}^n$, as follows. A grid of

constants λ_s , $s = 1, \dots, S$ is chosen (typically $0 = \lambda_0 < \lambda_1 < \dots < \lambda_S = 2$, Cook and Stefanski (1994)) 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 $N(0, \sigma_u^2)$, where σ_u^2 is assumed to be a known and finite variance. For a given value of λ_s , $s = 1, \dots, S$, the simulation is repeated $b = 1, \dots, B$ 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 $b = 1, \dots, B$. 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_u^2 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 (Carroll et al., 2006).

2.3.3 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 $\mathbf{W}$. 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 A.4.1 and Section 2.4). Then, separate extrapolation steps produce two SIMEX estimates, $\hat{\beta}_{\text{SIMEX},1}$ and $\hat{\beta}_{\text{SIMEX},2}$, 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.

2.3.3.1 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 i th school $\mathbf{W}_i = (W_{i1}, \dots, W_{iQ})$ are computed with these data, using the observed matrix of indicators $\mathcal{W}(p_0)$ as described in Section 2.2.2. 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 $\mathbf{W}_{i,b}(p_s) = (W_{i1}(p_s), \dots, W_{iQ}(p_s))$ 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 $s = 1, \dots, S$; $0 < p_0 < p_1 < \dots < p_S$, where S denotes the number of coarsening levels to be simulated, and for $b = 1, \dots, B$ replications within each level s to reduce Monte Carlo error. By increasing the number of coarsened geocodes, $\mathbf{W}_{i,b}(p_s)$ will accordingly have greater amount of measurement errors. After obtaining $\mathbf{W}_{i,b}(p_s)$, the observations (e.g., 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 $\mathbf{Y} = \mathbf{W}_b(p_s)\boldsymbol{\beta}_x + \mathbf{Z}\boldsymbol{\beta}_z + \boldsymbol{\epsilon}$ within each stratum. The estimate of $\boldsymbol{\beta}_x$ using contaminated exposure $\mathbf{W}_b(p_s)$ is denoted by $\hat{\boldsymbol{\beta}}_{xb}(p_s)$. In each stratum, we compute the average of the estimates from the B simulated pseudo-data for each proportion p_s , $\hat{\boldsymbol{\beta}}(p_s) = \frac{1}{B} \sum_{b=1}^B \hat{\boldsymbol{\beta}}_b(p_s)$, where $\hat{\boldsymbol{\beta}}_b(p_s) = (\hat{\boldsymbol{\beta}}_{xb}^T(p_s), \hat{\boldsymbol{\beta}}_{zb}^T(p_s))$.

2.3.3.2 Weighted extrapolation step

For each stratum, we carry out the extrapolation step separately. First we plot each element q of the vector $\hat{\boldsymbol{\beta}}(p_s)$, $s = 1, \dots, S$ against p_s and fit an extrapolation function $f(p)$ using regression methods. The SIMEX estimator for the exposure effect of the q th food outlet is obtained as the intercept of the extrapolation function, $\hat{\beta}_{\text{SIMEX}_q} = \hat{f}_q(0)$, which provides an estimate for when the ideal case is available, i.e., 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, as in Apanasovich et al. (2009). 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 $\mathbf{Z}$, 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 vari-

ance of estimates at each coarsening level, i.e. $1/V_{s,q}$ 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 $\widehat{\beta}_{\text{SIMEX}} = (\widehat{\beta}_{\mathbf{x}_{\text{SIMEX}}}^T, \widehat{\beta}_{\mathbf{z}_{\text{SIMEX}}}^T)^T$ can be obtained by a standard bootstrap approach. We resample the subjects with replacement and repeat entire SIMEX procedure to obtain $\widehat{\beta}_{\text{SIMEX}}^{(k)}$ for $k = 1, \dots, K$ bootstrapped samples. Thus, for each stratum, the bootstrap variance estimator is $\widehat{Var}(\widehat{\beta}_{\text{SIMEX}}) = \frac{1}{K-1} \sum_{k=1}^K \left(\widehat{\beta}_{\text{SIMEX}}^{(k)} - \overline{\widehat{\beta}_{\text{SIMEX}}} \right) \left(\widehat{\beta}_{\text{SIMEX}}^{(k)} - \overline{\widehat{\beta}_{\text{SIMEX}}} \right)^T$.

2.3.3.3 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 $\widehat{\beta}_{\text{SIMEX},1}$ and $\widehat{\beta}_{\text{SIMEX},2}$:

$$\widehat{\beta}_{\text{SIMEX,combined}} = \gamma_1 \odot \widehat{\beta}_{\text{SIMEX},1} + \gamma_2 \odot \widehat{\beta}_{\text{SIMEX},2}. \quad (2.2)$$

where $\odot$ denotes element-wise product operation, and γ_1 and γ_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 q th coefficient are

$$\gamma_{1q} = \frac{\frac{1}{Var(\widehat{\beta}_{\text{SIMEX},1q})}}{\frac{1}{Var(\widehat{\beta}_{\text{SIMEX},1q})} + \frac{1}{Var(\widehat{\beta}_{\text{SIMEX},2q})}}, \gamma_{2q} = \frac{\frac{1}{Var(\widehat{\beta}_{\text{SIMEX},2q})}}{\frac{1}{Var(\widehat{\beta}_{\text{SIMEX},1q})} + \frac{1}{Var(\widehat{\beta}_{\text{SIMEX},2q})}}. \quad (2.3)$$

If γ_1 and γ_2 are known, then

$$\widehat{Var}(\widehat{\beta}_{\text{SIMEX,combined}}) = \text{diag}(\gamma_1^{\odot 2})\widehat{Var}(\widehat{\beta}_{\text{SIMEX,1}}) + \text{diag}(\gamma_2^{\odot 2})\widehat{Var}(\widehat{\beta}_{\text{SIMEX,2}}). \quad (2.4)$$

The operation $\odot^2$ denotes that each element of a vector is raised to the 2nd power, and $\text{diag}()$ 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 γ_1 and γ_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

$$\widehat{Var}(\widehat{\beta}_{\text{SIMEX,combined}}) = \text{diag}(\widehat{\gamma}_1^{\odot 2})\widehat{Var}(\widehat{\beta}_{\text{SIMEX,1}}) + \text{diag}(\widehat{\gamma}_2^{\odot 2})\widehat{Var}(\widehat{\beta}_{\text{SIMEX,2}}). \quad (2.5)$$

We show in simulations that (2.5) is an adequate variance estimator.

2.4 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 2.2.3 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 2.3.1. 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 $\mathbf{X}$ 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 1,000 datasets of size $N=1,000$, by sampling the locations of 1,000 schools. Other simulation parameters for each dataset are $B=20$, $K=100$, and $S=30$ levels of coarsening from $p_0 = 15\%$ to $p_S = 50\%$.

2.4.1 Different patterns of bias by presence of ZIP centroid in the buffer

Figure 2.3 displays the different pattern of average estimates $\widehat{\beta}_{\mathbf{x}}(p_s)$ 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 2.1b which is an all-subject combined analysis, Figure 2.3 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.

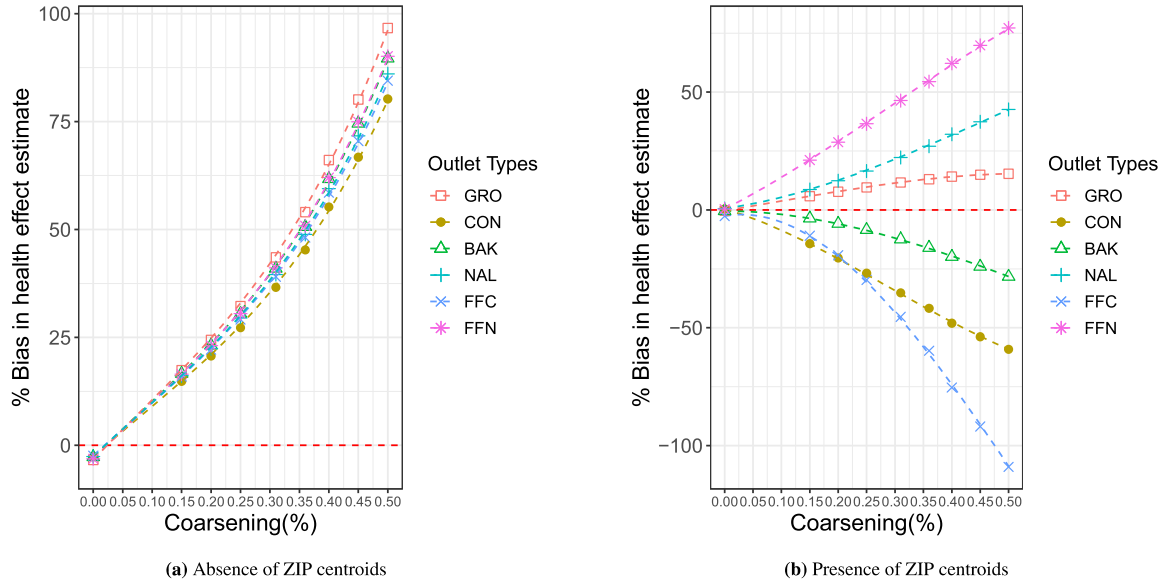


Figure 2.3: Percent bias in health estimate against each coarsening level p_s , stratified by whether schools do not have a ZIP centroid within their buffers (a) or have at least one (b). The naïve estimates occur at $p_0 = 15\%$ in both strata. Additional coarsening to original observed data magnifies 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 1,000 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 non-chain (FFN).

To see how the differences in the direction of bias arise, consider the ordinary least square (OLS)-estimators. Omitting $\mathbf{Z}$ for simplicity, the estimate of β_x can be derived as $\widehat{\beta}_x = \Sigma_W^{-1} \text{cov}(\mathbf{W}, \mathbf{Y}) \approx \Sigma_W^{-1} \text{cov}(\mathbf{W}, \mathbf{X}\beta_x) = \Sigma_W^{-1}(\Sigma_X + \Sigma_{XU})\beta_x = [\mathbf{I} - \Sigma_W^{-1}\Sigma_{WU}]\beta_x \equiv [\mathbf{I} - \Omega]\beta_x$. Attenuation or inflation in estimates is determined by $[\mathbf{I} - \Omega]$. As shown in Table 2.1, generally Ω can not be null because $\Sigma_{WU} (= \Sigma_{XU} + \Sigma_U)$ is not zero. Furthermore, as Table 2.1 shows, there are non-zero positive or negative covariances. The sign and magnitude of covariance matrices are determined by spatial locations of subjects and businesses. Appendix A.4.1 shows algebraic justifications for the varying directions of bias shown in Figure 2.3, for the case of regressions with a single exposure.

2.4.2 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 2.2 presents the percent bias, coverage probabilities of their 95% Wald confidence intervals (CI), and mean squared error $\times 100$ (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 non-chains, 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 (Carroll et al., 2006), 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 (Table A.3). In Table 2.2, 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 for examples: Alexeeff et al. (2016); Mao et al. (2017)). Despite this difference, SC-SIMEX produces smaller standard error estimates than other SIMEX

methods (Table A.5). To increase the Monte Carlo precision, increasing B could be helpful as recommended by Carroll et al. (2006), 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 $B=100$.

Table 2.2: Simulation results comparing naïve, traditional SIMEX, adjusted SIMEX, and split-combine SIMEX; Weighted cubic regression is used for extrapolation $f(\cdot)$. We used randomly sampled $N=1,000$ schools. The results are from 1,000 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 non-chain (FFN), Grocery (GRO), Fast food chain (FFC), Non-Alcohol drinking place (NAL).

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%

2.4.3 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 x 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 1,000 schools for 7th-grade children. Children in 7th grade typically attend middle school, which typically consists of grades 6-8. 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 (Table A.3 and Table A.3) 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 non-chain fast food (FFN) when considering all schools, the bias is negative when considering only the location of schools for 7th-grade children.

2.5 Application to Fitnessgram Data

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%), non-chain 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 (California Department of Education, 2019). 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) (Pietrobelli et al., 1998). 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 (Ingram and Franco, 2014), 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 5th, 7th, or 9th grade, who attended one of 1,166 schools. We focus the analysis on Asian children because previous research had documented a counter-intuitive finding among this population subgroup (Sánchez et al. (2012)). 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 (Wang et al., 1998). 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 non-positive 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 (Eckert et al., 1997). 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 $\mathbf{W}(p)$ has been simulated. Hence, we take advantage of direct simulation of $\mathcal{W}$ to generate $\log_2(\mathbf{W}(p) + 1)$.

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 (i.e., six separate predictors) to highlight efficiency gains of SC-SIMEX when multiple outlets are included, as demonstrated from the simulation study in Section 2.4.

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.

Results Figure 2.4 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 2.3 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.

2.6 Discussion

In this paper, we focused on two major objectives. First, we examined the measurement error distribution and bias in estimating the health impact of point-

Table 2.3: Multiple outlet model. 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 2.1 for outlet abbreviations.

Outlet Type	Naïve $\hat{\beta}(SE)$	Traditional $\hat{\beta}(SE)$	Trad+Centroid $\hat{\beta}(SE)$	SC-SIMEX $\hat{\beta}(SE)$
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)

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 estimation of the health effect is not always towards 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 (Carroll et al., 2006). However, its applications were until now limited to parametric measurement error distributions where estimated parameter values were available (e.g., 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 misclassification 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, see Carroll et al. (2006) for examples) or a normal approximation is used when the count is large such as in population estimation studies (Cheng et al., 2015). There are other works that use Poisson regression models for count *outcomes* with continuous erroneous predictors (Thamerus, 1997; Kukush et al., 2004). Other discrete error distributions such as misclassification errors have been studied for categorical predictors (Gustafson and Le Nhu, 2002; Buonaccorsi, 2010; Keogh et al., 2020). 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. A straightforward extension is also possible to measurement errors due to different types of geocode accuracy. In NETS database, there are five different levels of geocode precision: block face (most accurate), street level, block group, census-tract centroid, and ZIP code centroid (least accurate). As long as we have the coordinates that businesses will be assigned to after coarsening, we can easily extend our method to different types of area-level precision. However, since there are only few businesses that have either street level, block group, or census-tract centroids ($< 2\%$), we are not too concerned about these remaining levels of precision.

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 paper 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.

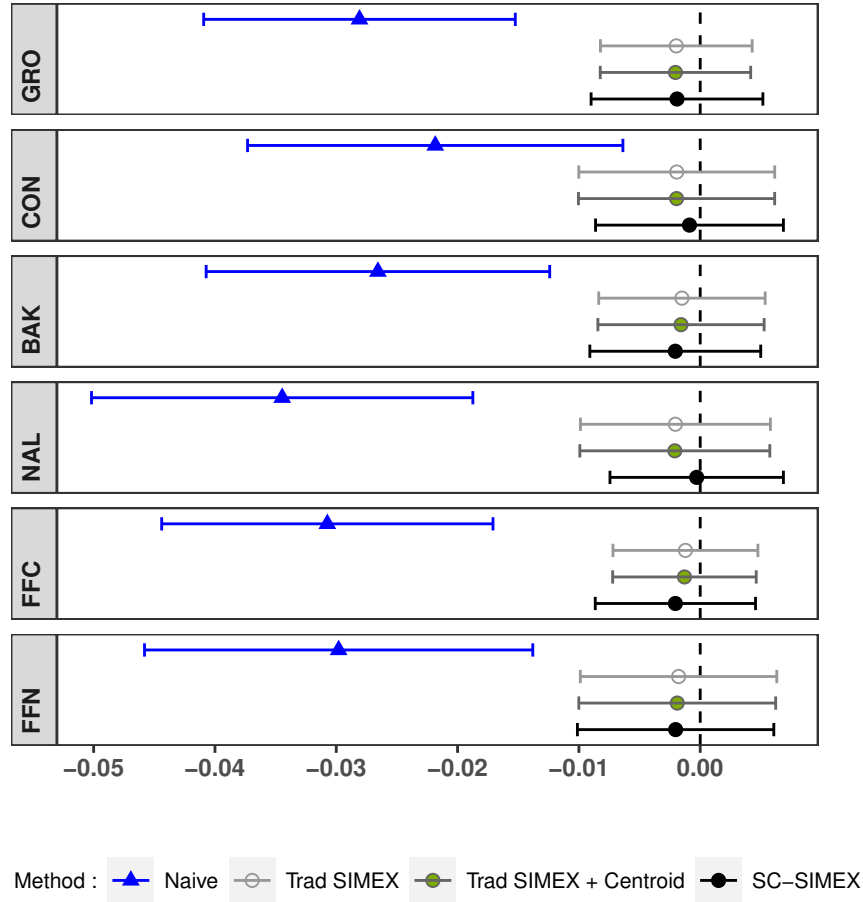


Figure 2.4: Single outlet models. Coefficient estimates and 95% confidence intervals from the naïve, traditional, covariate-adjusted, and SC-SIMEX approaches, illustrating the impact of geocode coarsening on the naïve regression. The estimates represent the difference in BMIz per each two-fold higher exposure to a given food outlet type. Each sub-panel represents the results of separate models, one for each food outlet type, among a sample of Asian children attending California public schools. See Figure 2.1 for outlet abbreviations.

CHAPTER III

Integrating Multiple Built Environment Data Sources

3.1 Introduction

Built environment and health studies focus on examining how the availability of community amenities influences health outcomes; for instance how the availability of junk food outlets near schools influences child obesity (Larson et al. (2009); Lee (2012)). Many of these studies rely on available commercial business data such as InfoUSA, Google Maps, Yelp, and others, to quantify the availability of food outlets. A practical challenge in the field is that exposure measurements differ depending on the food outlet data source, due to what is known as ascertainment error – i.e., data sources have incomplete listings or list businesses that no longer exist. There is a growing concern about the extent of bias in health effect estimates due to ascertainment error (Lucan et al. (2013); Jones et al. (2017)), and resource intensive-strategies have been recommended to minimize measurement error (Caspi and Friebe (2016)). Yet, the impact of ascertainment error on health effect estimates has not been systematically investigated, is rarely quantified in practice, and is often explained with simplified heuristics such as citing attenuation of regression coefficients (van Smeden et al. (2020)). In this paper, we show that exposure ascertainment error introduces

biases with more complex patterns, beyond simplified heuristics, and propose methods to integrate exposure information from two popular commercial business lists to construct more accurate school-level exposure measures and improved association estimates.

A large amount of literature has proposed methods to combine information from error-prone databases or multi-observer data. Among these, capture-recapture methods have been widely used for ecological population estimation and many extensions of the original approach are available (Fienberg (1972b); Pollock and Otto (1983); Manrique-Vallier (2016)). The key information required for capture-recapture methods is a capture history that identifies which list captured a particular entry. However, in our specific application, we do not know the capture histories for individual business records unless we merge listings. Due to the proprietary nature of the commercial databases and related data use agreements, business-level matching is not permitted, thus standard capture-recapture methods cannot be applied. In our motivating study, we can compare the aggregated exposure metrics at a given subject’s location, i.e., the number of available businesses within each school’s neighborhood. Although matching the sources is not allowed, field validation studies of such secondary commercial data sources are available (e.g., Liese et al. (2013); Fleischhacker et al. (2013)). External validation studies use direct field observations for a given geographic area, validate the listed businesses on databases, and publish validity scores for each source. Given this reference about the source quality, latent truth models can be one potential approach to improve exposure measures. Latent truth models are popular for crowdsourcing problems where data providers have different levels of credibility (Dong et al. (2009); Zhao et al. (2012)). This modeling approach uses the user’s sensitivity and specificity to evaluate whether the claim or input from each user (i.e., data source) is true or not. However, the latent truth model focuses on finding the “correct” data entry among multiple observations. In our case, however, it is

highly unlikely that true exposure values are captured by any one list or combination thereof.

Our goal is to minimize the bias in estimating the health effect of food environment exposure when direct observation or gold standard data are not available. After describing the motivating study in Section 3.2, in Section 3.3 we propose a method that accounts for errors in exposure measures derived from more than one business database. Specifically, for each database, we consider a model that treats the business counts as being obtained from the trivariate distribution of: observed true businesses, falsely listed businesses that don't exist, and unlisted but truly existing businesses. The number of latent true businesses is assumed to be the same across databases. A flexible Dirichlet process prior is used for the latent true exposure. The health outcome is modeled jointly with a regression model with the latent true business total as the primary predictor. Available estimates of the database's credibility (sensitivity and positive predictive value) allow for the identification of model parameters. Section 3.4 presents an estimation approach and Section 3.5 presents simulation studies to its properties. Section 3.6 applies our method to estimate the effect of availability of convenience stores and grocery stores near schools on BMI among children in California. In Section 3.7, we discuss final remarks and possible future extensions.

3.2 Food exposure in multiple sources

Our motivating example investigates the association between the availability of food outlets near schools and the bodyweight of 5th and 7th-grade children attending 2,651 urban public schools in California in 2008. Children's weight and height are collected as part of the state-mandated Fitnessgram test (California Department of Education (2019)), and used to calculate body mass index z-scores (BMIz). BMIz scores are obtained by standardizing BMI according to age and sex-specific reference distributions, and ensure comparability of body weight across children of different

ages since they are still growing (Must and Anderson (2006)). Our outcome variable is school-level mean body mass index z-scores (BMIz) among children in these schools. We use publicly available school-level characteristics as covariates (e.g., income), and the schools' latitude and longitude to assess the availability of outlets nearby.

Lists of food outlets are obtained from the two most popular commercial business lists used in built environment studies in the United States: the National Establishment Time-Series (NETS) (NETS (2021)) and InfoUSA (InfoUSA (2021)). The NETS is a longitudinal dataset produced by Walls & Associates, from annual Dun & Bradstreet (D&B) establishment data. InfoUSA is an independent data vendor (a competitor of D&B), that produces annual listings of businesses. Both databases include information regarding business names, their locations (addresses and latitude-longitude), dates of business openings and closings, and information on the goods or services offered by the business as denoted by standard industrial classification (SIC) codes. Validation studies show that these business lists have different validity and do not always provide matching information for a given region (Liese et al. (2013); Jones et al. (2017)).

Direct merging of these two databases at the business level (e.g., matching by business name or address or other statistical matching algorithms) is not allowed due to specifications in the data use agreements with the data providers. Instead, we compare the school-level exposures, which are aggregated/derived quantities. We use one of the popular choices to derive exposure measures, namely counting the number of community resources within a one-mile radius circular buffer around each school (Athens et al. (2016)). We focus on SIC codes that identify 'convenience stores' and separately, 'grocery stores', as defined in a field validation study (Powell et al. (2011)). For convenience stores, establishments with SIC codes 541102 , 55410000 , 55419901 , and 55419903 are used for NETS; establishments with SIC/NAICS codes 541103, 554101 , 554102 , and 554103 are used for InfoUSA. For grocery stores, SIC

codes 541101, 541100, and 541199 are used for NETS; SIC/NAICS codes 541101, 541102, 541104, 541105, 541106, 541107, and 541108 are used for InfoUSA. The exposure distributions of convenience stores and grocery stores are shown in Figure 3.1. The figure demonstrates the discordance between NETS and InfoUSA. For the following sections, we take convenience store exposures to illustrate and compare the distributions of the exposure data across the databases as well as overdispersion.

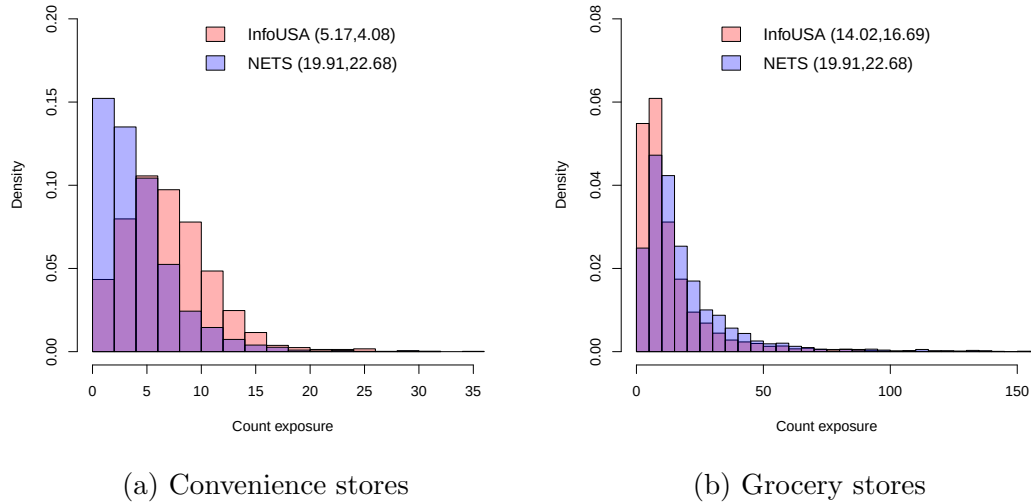


Figure 3.1: Distributions of exposure to convenience stores and grocery stores from different sources. Exposures are computed by counting the number of stores within 1 mile buffer around California public schools. Means and standard deviations are presented in the legend.

With the growing use of readily available business databases in community environment studies, there is a rich set of validation studies that compare the commercial databases with ground observations from field studies, or from gold-standard data such as local government databases (Powell et al. (2011); Liese et al. (2013); Lebel et al. (2017)). The agreement between the true availability and listings on commercial databases is measured by the number of correctly listed businesses (true positive), incorrectly omitted businesses (false negative), and incorrectly listed businesses (false positive). These measures are used to construct two major validity scores: (1)

sensitivity: the proportion of truly existing establishments that were listed in the commercial database and (2) positive predictive value (PPV): the proportion of establishments listed in the commercial database that were also present on the ground. The undercount (1-sensitivity) and overcount rates (1-PPV) are also commonly reported.

Considering the similarity of urbanization level of the study region and the year that data were collected, we use Powell et al.’s validity statistics as a reference for the California establishment data in 2008. According to the authors, the sensitivity of D&B and InfoUSA for convenience stores were 0.3769 and 0.5047, respectively, with PPVs of 0.4812 (D&B) and 0.6182 (InfoUSA). For grocery stores, the sensitivities were 0.4635 (D&B), 0.5414 (InfoUSA), and PPVs were 0.2880 (D&B), and 0.4446 (InfoUSA).

3.3 Modeling approach

3.3.1 Notation for measurement error model

Let X_i denote the true, unobserved outlet count within a one mile radius buffer of i th subject (e.g., school), and Y_i be a continuous outcome for subject $i = 1, \dots, n$. For simplicity, we assume a simple linear regression model: $Y_i = \beta_0 + \beta_x X_i + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma)$, although inclusion of covariates is straightforward. Our primary interest is estimating β_x .

Table 3.1: Contingency tables comparing exposure measures between truth and claims from two databases.

(a) Database 1				(b) Database 2			
	Listed	Not Listed	Total		Listed	Not Listed	Total
Exist	$N_{111,i} (\pi_{111})$	$N_{121,i} (\pi_{112})$	X_i	Exist	$N_{112,i} (\pi_{112})$	$N_{122,i} (\pi_{122})$	X_i
Not exist	$N_{211,i} (\pi_{211})$	-		Not exist	$N_{212,i} (\pi_{212})$	-	
Total	W_{1i}			Total	W_{2i}		

Instead of true X_i , contaminated measures W_{1i} and W_{2i} are derived from two databases. Their relationship with X_i is represented in Table 3.1. For database s ($s = 1, 2$), $N_{11s,i}$ denotes the number of outlets that are “correctly” captured (true positives) within the buffer area of the i th subject; $N_{12s,i}$ is the number of omitted outlets (false negatives), and $N_{21s,i}$ denotes the number of non-existing outlets that are falsely listed (false positives). Since outlets that do not exist cannot be incorrectly omitted from the databases, these tables have structural zeros. Each table has three cell probabilities $\boldsymbol{\pi}_s = (\pi_{11s}, \pi_{12s}, \pi_{21s})$ which sum to one. The tables share the row margin $X_i = N_{111,i} + N_{121,i} = N_{112,i} + N_{122,i}$ (true exposure), but have different column margins $W_{1i} = N_{111,i} + N_{211,i}$ and $W_{2i} = N_{112,i} + N_{212,i}$ (observed exposure). Since the companies providing establishment data are competitors, we assume the two contingency tables in Table 3.1 are conditionally independent given the row margin X_i . Thus, the data at each location can be represented as two conditionally independent contingency tables with structural zeros. Note that we can easily extend our model to more than two databases due to the conditional independence assumption.

To develop a modeling strategy for the contingency tables *at any one location*, we rely on the Poisson sampling scheme for tables with structural zeros, where each cell is assumed to be an independent Poisson variate (Fienberg (1972a); Levin (1981)). Since the sum of independent Poisson random variables also has a Poisson distribution with a sum of the mean parameters, the shared row margin X_i is Poisson distributed. We define $X_i \sim \text{Pois}(\lambda_i)$, with a location-specific mean number of stores within i th subject’s buffer λ_i .

Let $\boldsymbol{\phi}_1 = \{\text{sen}_1, \text{ppv}_1\}$ and $\boldsymbol{\phi}_2 = \{\text{sen}_2, \text{ppv}_2\}$ denote sensitivity and positive predictive values for source s respectively for $s = 1, 2$, which are available from published literature as described in Section 3.2. Although $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ could be further indexed by i , to reflect variation by school-level characteristics, for example, such granular data about PPV and sensitivity would rarely be available. Validation studies

report sensitivity and PPV as proportions calculated using the observed numbers of businesses in a large study region (e.g., an entire city). To model ϕ_{si} instead of ϕ_s , we would need additional data or need to rely on stronger modeling assumption; otherwise the model will not be identifiable. Furthermore, the majority of validation studies that examined the association between the source quality and geographical (e.g., urbanicity) or demographic characteristics (e.g., socio-economic status) of the locations found no significant difference in sensitivity or PPV values across these location characteristics (Lebel et al., 2017). Thus, we assume the agreement statistics are constant over all subject locations (schools), and express them based on the expected counts: $sen_s = \frac{E(N_{11s,i})}{E(X_i)}$ and $ppv_s = \frac{E(N_{11s,i})}{E(W_{si})}$ for $i = 1, \dots, n$. Then, as shown in Appendix B.1, each cell probability π_1, π_2 in Table 4.1 can be computed from ϕ_1 and ϕ_2 .

Since the shared row margin $X_i \sim Pois(\lambda_i)$, the cell counts of database s follow:

$$\begin{aligned} N_{11s,i} &\sim Pois\left(\frac{\pi_{11s}}{\pi_{11s} + \pi_{12s}}\lambda_i\right) \equiv Pois(sen_s\lambda_i), \\ N_{12s,i} &\sim Pois\left(\frac{\pi_{12s}}{\pi_{11s} + \pi_{12s}}\lambda_i\right) \equiv Pois((1 - sen_s)\lambda_i), \\ N_{21s,i} &\sim Pois\left(\frac{\pi_{21s}}{\pi_{11s} + \pi_{12s}}\lambda_i\right) \equiv Pois\left(\left(\frac{sen_s}{ppv_s} - sen_s\right)\lambda_i\right). \end{aligned} \quad (3.1)$$

Consequently, the local expectations of W_{1i} and W_{2i} are proportional to λ_i : $E(W_{1i}) = E(N_{111,i} + N_{211,i}) = \frac{sen_1}{ppv_1}\lambda_i$, $E(W_{2i}) = E(N_{112,i} + N_{212,i}) = \frac{sen_2}{ppv_2}\lambda_i$. The measurement error does not follow classical zero mean assumption because $E(W_{si} - X_i) = \left(\frac{sen_s}{ppv_s} - 1\right)E(X_i)$ is typically non-zero unless $sen_s = ppv_s$ or the expected true exposure is zero, which does not generally hold for count exposures.

3.3.2 Modeling mixture distribution of the latent exposure counts

We assumed the true exposure is locally Poisson distributed with mean parameter λ_i . However, not imposing any structure on the λ_i 's leads estimation problems,

while setting λ_i 's to a constant leads to an unrealistic marginal distribution for the X 's (given the overdispersion seen in Figure 3.1). One approach is to model λ_i as dependent on covariates for the i th subject. However, our primary interest is on β_x , not identifying predictors of exposure. Moreover, misspecification of such model could introduce bias in estimates of β_x . Thus, we aim to model λ_i flexibly, in order to reduce the parameters estimated while maintaining robustness. Specifically, we consider a Poisson mixture model, so that conditional on mixture component membership, the results in Section 3.3.1 hold, without requiring that we know in advance the specific mixture component to which a subject belongs.

Further, to avoid specifying the number of mixture components in advance, we use a Dirichlet process prior on λ_i (Ferguson (1973)), which enables clustering of λ_i based on shared information across sites and ultimately creates a flexible marginal distribution for the latent exposures, X , effectively handling overdispersion. Although this prior allows for infinitely many mixture components (clusters), in practical applications a finite number of clusters, K , is identified in a data-adaptive fashion. The cluster allocation of λ_i is denoted as $Z_i \in \{1, \dots, K\}$. Within the k th cluster, subjects share the same value λ_k so that $[X_i|Z_i = k] \sim \text{Pois}(\lambda_k)$. The model is: $X_i \sim \text{Pois}(\lambda_i)$; $\lambda_i \sim G$; $G \sim DP(\alpha G_0)$; $G_0 = \text{Gamma}(a_0, b_0)$, where α denotes a concentration parameter and G is a random measure on the parameter space. The choice of α will be discussed in the following Section 3.4.

Under the Bayesian framework, we are interested in the posterior distribution of latent X_i given observed data, which will be used as the basis of Gibbs sampling in Section 3.4. For each subject, we jointly model two independent trinomial distributions with known column margins (W_{1i}, W_{2i}) and unknown shared row margin X_i . Using Table 3.1 with our outcome model defined in Section 3.3.1, we can model the conditional distribution of X_i given observed data, cluster membership, and a set of

parameters as

$$[x_i|y_i, w_{1i}, w_{2i}, z_i] \propto [y_i|x_i; \beta_0, \beta_x, \sigma] \times [w_{1i}|x_i; \phi_1, \lambda_k] \times [w_{2i}|x_i; \phi_2, \lambda_k] \times [x_i|z_i = k; \lambda_k]. \quad (3.2)$$

By definition of the conditional distribution,

$$\begin{aligned} [w_{1i}|x_i; \phi_1, \lambda_k] &= \frac{[w_{1i}, x_i; \phi_1, \lambda_k]}{[x_i; \lambda_k]} = \frac{\sum_{n_{111}} [n_{111,i}, w_{1i}, x_i; \phi_1, \lambda_k]}{[x_i; \lambda_k]} \\ &= \frac{\sum_{n_{111}} [n_{111,i}; sen_1 \lambda_k] [x_i - n_{111,i}; (1 - sen_1) \lambda_k] [w_{1i} - n_{111,i}; (\frac{sen_1}{ppv_1} - sen_1) \lambda_k]}{[x_i|z_i=k; \lambda_k]}, \end{aligned}$$

since each cell count is independent Poisson random variables. Thus, the log-likelihood is

$$\begin{aligned} l_{x_i|\sim} &\propto \log N(y_i|x_i; \beta_0, \beta_x, \sigma) \\ &+ \log \left(\frac{\sum_{n_{111,i}=0}^{\min(w_{1i}, x_i)} Pois(n_{111,i}; sen_1 \lambda_k) Pois(x_i - n_{111,i}; (1 - sen_1) \lambda_k) Pois(w_{1i} - n_{111,i}; (\frac{sen_1}{ppv_1} - sen_1) \lambda_k)}{Pois(x_i|z_i=k; \lambda_k)} \right) \\ &+ \log \left(\frac{\sum_{n_{112,i}=0}^{\min(w_{2i}, x_i)} Pois(n_{112,i}; sen_2 \lambda_k) Pois(x_i - n_{112,i}; (1 - sen_2) \lambda_k) Pois(w_{2i} - n_{112,i}; (\frac{sen_2}{ppv_2} - sen_2) \lambda_k)}{Pois(x_i|z_i=k; \lambda_k)} \right) \\ &+ \log Pois(x_i|z_i = k; \lambda_k). \end{aligned} \quad (3.3)$$

For the outcome model $N(y_i|x_i; \beta_0, \beta_x, \sigma)$, we assumed a non-differential measurement error so that observed exposures are independent of the outcome given the true data. Non-differential measurement error of exposure is most commonly used assumption in measurement error models, since it enables regression coefficients to be identifiable even when true covariates are not observable. When measurement errors are differential, outcome status can influence the measurement errors in exposure (e.g., case-control study with self-reported exposures) (Carroll, 1998). If we assume the differential error, $[w_{1i}|x_i; \phi_1, \lambda_k] \times [w_{2i}|x_i; \phi_2, \lambda_k]$ in equation (3.2) should be replaced with $[w_{1i}|x_i, y_i; \phi_1, \lambda_k] \times [w_{2i}|x_i, y_i; \phi_2, \lambda_k]$. This modeling is practically

not possible due to limited identification. Imai and Yamamoto (2010) suggests more assumptions such as restricting bounds of parameters. Differential models with a validation study that has (W_1, W_2, Y, X) all observed are typically best analyzed by techniques for handling missing data (Carroll, 1998). For example, Carroll et al. (1993) proposed a pseudo-likelihood approach where parametric measurement error model is chosen based on the validation dataset and marginal distribution of X is estimated by empirical distribution from the complete data.

Although we could also estimate X_i by maximizing the log likelihood in equation (3.3) by treating λ_k nuisance parameter, the estimation process was not stable because we only have two observations (W_{1i}, W_{2i}) per site. However, having a distribution of X_i in the model with clustered rate parameters showed more stable estimation. Alternatively, the EM algorithm can be considered for a frequentist inference. However, unlike the Dirichlet process mixture model, the EM algorithm requires a fixed number of clusters, which is not desirable in our problem with unknown degree of overdispersion in true exposure data. Also, the EM algorithm may suffer from the multimodality of the model that typically arises with discrete-valued latent variable (Bartolucci et al., 2022).

3.4 Estimation

We use a Bayesian approach for estimation. We chose a noninformative prior for regression parameters $\pi(\beta_0, \beta_x, \sigma) \propto \sigma^{-2}$ and hyperparameters (a_0, b_0) for the base measure G_0 are set to (1,1). Given that the Dirichlet process mixture (DPM) model is used to model the distribution of the *latent* exposure, there is relatively little information in the data to identify the concentration parameter α . Thus, although more recent applications of DPM models estimate α , we instead use a fixed value, chosen as follows. Assuming that observed exposures have a similar shape of distribution to the true exposure, we fit a DPM model to the observed exposures, W 's,

from each data source separately. In these models, we use a $\text{Gamma}(1,1)$ prior for a source-specific concentration parameter α_s , and thus obtain a data-driven estimate of a concentration parameter. We use 10,000 iterations for each model, and after discarding the first half as a burn-in, each chain produces the posterior mean $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for source 1 and 2, respectively. We average the separate estimates and obtain $\bar{\alpha} = \text{mean}(\hat{\alpha}_1, \hat{\alpha}_2)$, which we plug into the model. This is a reasonable approach because the observed exposures have information on the number of components and mixing proportions that make up the mixture distribution for X_i . Specifically, consider that $W_{si} \sim \text{Pois}(\frac{sen_s}{ppv_s} \lambda_i)$ at i th subject's location and the $\lambda_i = \lambda_k$ when the i th subject belongs to k th cluster. Thus, the distribution of W_{si} has the same number of components, although the mean parameters of the mixing components are multiplied by a nonzero scalar $\frac{sen_s}{ppv_s}$.

We implement a Gibbs sampling algorithm to draw values from the posterior distribution and use standard summaries of the draws for inference (see Appendix B.2 for details). Within the algorithm, we implement the DPM model for the latent exposure model using the Chinese restaurant process (CRP) (Aldous (1985); Teh et al. (2010)). Loosely speaking, the CRP finds a flexible number of clusters by allowing observations to be assigned to new clusters as needed. The conditional posterior probability of allocation to a new cluster is proportional to the concentration parameter $\bar{\alpha}$ in the previous section.

3.5 Simulations and bias analysis

The measurement error in the exposure predictor leads to bias in the estimation of regression coefficients. Here we describe algebraically the nature of the bias in linear regression coefficients when a naïve regression analysis that disregards the measurement error is used. Having described the factors that drive the direction and magnitude of the bias, we demonstrate our model's estimation properties, under

various data generation scenarios.

3.5.1 Bias in OLS

When exposure measures derived from a single source s are used as a predictor in a linear regression model, and the observed exposures arise from the hypothesized model (Table 3.1), the ordinary least squares (OLS) estimate of the target coefficient is

$$\text{plim } \hat{\beta}_{w_s} = \frac{\frac{\text{Var}(X)}{E(X)} - (1 - ppv_s)}{\left(\frac{sen_s}{ppv_s}\right) \frac{\text{Var}(X)}{E(X)} + (1 - \frac{sen_s}{ppv_s})} \beta_x = \left(\frac{(ppv_s - 1)E(X) + (1 - \frac{sen_s}{ppv_s})\{\text{Var}(X) - E(X)\}}{E(X) + (\frac{sen_s}{ppv_s})\{\text{Var}(X) - E(X)\}} + 1 \right) \beta_x \quad (3.4)$$

(see Appendix B.2 for details). The scaling factor multiplying β_x in (3.4), also known as the *reliability ratio* in the measurement error literature, consists of source credibility statistics and the dispersion of X relative to its mean, or, equivalently, the difference between the mean and variance of X . Consequently, the direction and magnitude of bias depend upon the source credibility and the true exposure distribution, as follows.

When $sen_s > ppv_s$, the term $(ppv_s - 1)E(X) + (1 - \frac{sen_s}{ppv_s})\{\text{Var}(X) - E(X)\}$ is negative since $\{\text{Var}(X) - E(X)\} \geq 0$, and therefore $\hat{\beta}_{w_s}$ is always attenuated towards zero. On the other hand, when $sen_s < ppv_s$, either attenuation or inflation can occur, depending on the amount of overdispersion. In the case X is marginally equidispersed (e.g., $\lambda_i = \lambda$ thus overall X has a simple Poisson distribution), the reliability ratio simplifies to ppv_s . Since $0 < ppv_s < 1$, the OLS estimate is always attenuated. In this case, unbiasedness can only occur when there are no false positives (i.e., $ppv_s = 1$), irrespective of sensitivity.

Given $sen_s < ppv_s$, the naïve coefficient estimate is attenuated when $\text{Var}(X) < \left\{ \frac{ppv_s(1 - ppv_s)}{ppv_s - sen_s} + 1 \right\} E(X)$. Otherwise, an inflated coefficient $\hat{\beta}_{w_s}$ is obtained. Interestingly, when the overdispersion is approximated by $\left\{ \frac{ppv_s(1 - ppv_s)}{ppv_s - sen_s} + 1 \right\} E(X)$, the naïve coefficient estimator is approximately unbiased. (Note that this is a special form of

negative binomial overdispersion, where the mean of the prior gamma distribution varies with shape and the scale remains constant – see (Nelder and Lee (1992))).

Considering the fact that W_1 and W_2 are independent given X , the average of the two exposures could be used as a predictor. However, the resultant estimate will also have a bias that depends on the combination of source credibility parameters, $E(X)$, and $Var(X)$.

3.5.2 Simulation design

We used the analytical results from the previous section to design our simulation study, which varies three factors: (i) the distribution of X , (ii) the ratio of sensitivity to PPV, and (iii) the value of agreement statistics (i.e., high or low source quality).

We aimed to use a realistic distribution of the latent exposure, while at the same time having a way to control the variance of X relative to its mean, given the dependence of the bias of β_w on these moments of X (equation (3.4)). Our strategy is to have $E(X)$ fixed and investigate the effect of the magnitude of variance on the amount of bias in the naïve estimate of β_x . To be specific, we fixed $E(X) = 5.71$ and created distributions of X with a varying degree of variance using mixture distributions (see Table 3.2). To obtain realistic mixture components from which to simulate X , we fitted a Poisson mixture distribution to the observed exposures from the InfoUSA data alone using a DPM model. The DPM was fitted using a concentration parameter $\alpha = 0.1$ and a Gamma(1,1) base measure for the rate parameter. The estimated median number of mixture components was four, with mixing probabilities (0.23, 0.42, 0.33, 0.02) and mean parameters (1,4,10,25).

The mixtures were selected to have increasing variance to mimic various degrees of overdispersion. This was achieved by setting ‘Mixture 1’ to be a simple Poisson(5.71). The remaining mixtures had four components with mixing proportions (0.23, 0.42, 0.33, 0.02) with the mean parameters of the components given in Table 3.2.

Table 3.2: Different distributions of X with the same expectation for the simulation study.

Type	$E(X)$	$Var(X)$	Components
Mixture 1	5.71	5.71	Pois(5.71)
Mixture 2	5.71	10.72	$0.23\text{Pois}(3) + 0.42\text{Pois}(5) + 0.33\text{Pois}(8) + 0.02\text{Pois}(14)$
Mixture 3	5.71	25.56	$0.23\text{Pois}(1) + 0.42\text{Pois}(4) + 0.33\text{Pois}(10) + 0.02\text{Pois}(25)$
Mixture 4	5.71	35.98	$0.23\text{Pois}(2) + 0.42\text{Pois}(5) + 0.33\text{Pois}(7) + 0.02\text{Pois}(42)$
Mixture 5	5.71	41.74	$0.23\text{Pois}(3) + 0.42\text{Pois}(5) + 0.33\text{Pois}(6) + 0.02\text{Pois}(47)$

Source quality parameters

To select *realistic* values of source quality, we used Powell et al.’s agreement statistics for convenience stores, namely $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. However, since the ratio of sensitivity to PPV determines the direction of bias in naïve analysis (Section 3.5.1), we reversed the sensitivity and PPV values, i.e., $sen_1 = 0.48$, $sen_2 = 0.62$, $ppv_1 = 0.37$, and $ppv_2 = 0.5$, to show the opposite scenario. Although other scenarios such as $(sen_1 > ppv_1, sen_2 < ppv_2)$ can exist, we found that the relative size of sensitivity to PPV is mostly consistent across different data sources (see Lebel et al. (2017) for systematic review). The source credibility reported for different food store type by different databases also suggest that it is most likely to observe the dominating nature of undercount error so that sensitivities are less than PPV (Powell et al., 2011; Liese et al., 2013). Hence, for practical importance, we only present $(sen_1 < ppv_1, sen_2 < ppv_2)$ or $(sen_1 > ppv_1, sen_2 > ppv_2)$. The extra simulation study with $(sen_1 > ppv_1, sen_2 < ppv_2)$ is available in Table B.3.

To investigate the extent to which bias varies with varying quality of the source, in addition to the ‘realistic’ scenario, we also defined ‘extremely optimistic’, ‘optimistic’, ‘pessimistic’, and ‘extremely pessimistic’ scenarios for data sources by multiplying the *realistic* sensitivity and PPV values by 1.5, 1.25, 0.75, and 0.5, respectively. For

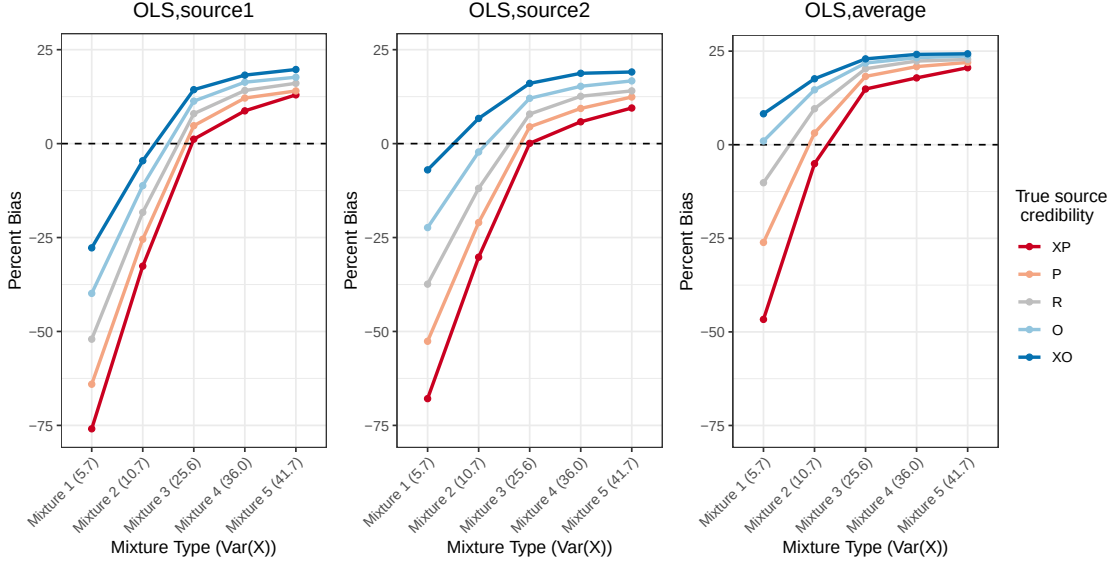
example, for the extremely pessimistic scenario we used $sen_1 = 0.185$, $sen_2 = 0.25$, $ppv_1 = 0.24$, and $ppv_2 = 0.31$. This range of source quality covers the values of agreement statistics reported by field studies of various other business types, aside from convenience and grocery stores considered here.

Data generation

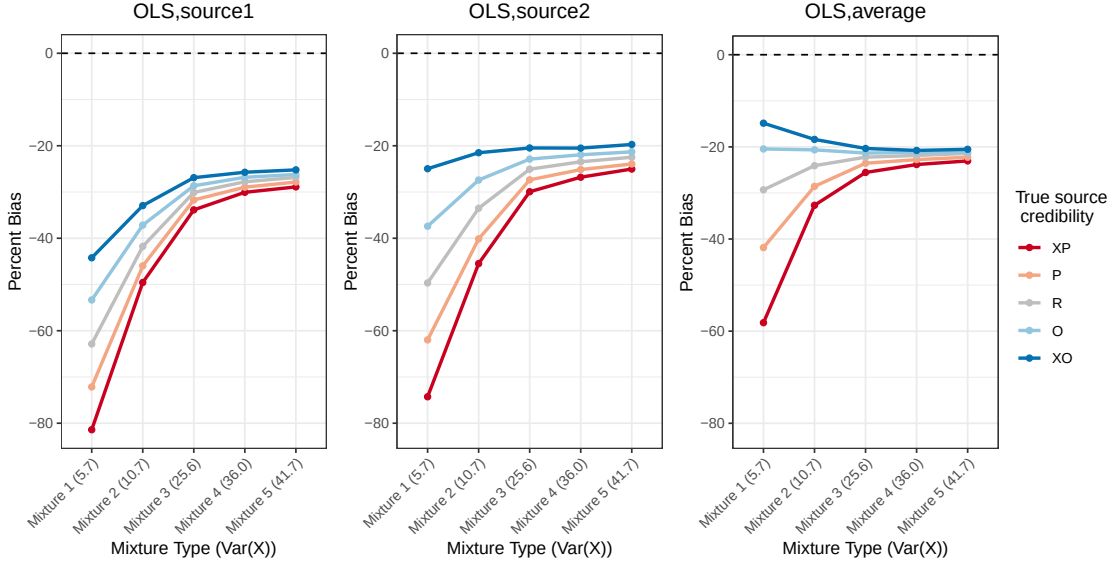
For the single component model, all X_i were generated from $Poisson(5.71)$. For the mixture component models, we first generated cluster assignments $\{Z_i\}_{i=1}^n$ to one of four mixture components from Multinomial distribution with given mixing probabilities (0.23, 0.42, 0.33, 0.02). Then we generated $X_i|Z_i = k$ from $Poisson(\lambda_k)$ using a set of mean parameters defined for each mixture distribution in Table 3.2. For each distribution of X , we generated 100 data sets of sample size $n = 1,000$. For each subject in each data set, we simulated health outcome Y_i from $N(1.5 + 0.5X_i, 1.5)$. To generate observed exposure value, W_{si} , we first obtained $N_{11s,i}$ from its conditional distribution $N_{11s,i}|X_i \sim Binom(X_i, sen_s)$ to prevent drawing a value that is larger than X_i . For $N_{21s,i}$, we draw a value from $Pois((\frac{sen_1}{ppv_1} - sen_1)\lambda_k)$ according to equation (3.1). Then W_{si} is obtained by adding $N_{11s,i}$ and $N_{21s,i}$.

3.5.3 Simulation results for naïve bias

The percent bias of the naïve slope coefficient is shown in Figure 3.2. Given the source quality, the naïve estimates for the slope have the largest attenuation when X has a simple Poisson distribution (‘Mixture 1’). When sensitivity is smaller than PPV in both sources, Figure 3.2(a) shows that the bias of OLS estimators switch from attenuation to inflation as $Var(X)$ exceeds $\{\frac{ppv_s(1-ppv_s)}{ppv_s-sen_s} + 1\}E(X)$ as described in Section 3.5.1. On the other hand, as displayed in Figure 3.2(b), the bias of all naïve estimators are attenuated towards zero when the ratio of sensitivity to PPV > 1 in both sources.



(a) $sen_1 < ppv_1, sen_2 < ppv_2$



(b) $sen_1 > ppv_1, sen_2 > ppv_2$

Figure 3.2: Percent bias of slope (β_x) estimates from naïve regression models (“OLS, source 1”, “OLS, source 2”, “OLS, average”) under the different X distribution, the ratio of sensitivity to PPV, and the magnitude of agreement statistics. Values for *realistic* (labeled as R in the legend) sensitivity and PPV are (a) $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, $ppv_2 = 0.62$, and (b) $sen_1 = 0.48$, $sen_2 = 0.62$, $ppv_1 = 0.37$, $ppv_2 = 0.5$.

Regarding the source quality, mean $\hat{\beta}_{w_s}$ is larger with more optimistic sources. It may seem counter-intuitive that we still observe biased naïve estimates from high-quality sources. However, this result is explained by equation (3.4) which shows that

higher PPV results in higher point estimates when the ratio $\frac{sen_s}{ppv_s}$ stays the same. To be more specific, the reliability ratio in equation 3.4 is a concave function with respect to sen_s when holding $Var(X)/E(X)$ and ppv_s constant (Figure B.1). The concavity of reliability ratio with respect to sen_s implies that we have smaller $\hat{\beta}_{w_s}$ as we have higher sensitivity, given first two moments of X and the value of PPV. This counter-intuitive result can be explained with probabilities that generate a contingency table in Table 3.1. Increasing $\frac{sen_s}{ppv_s}$ indicates higher probability of adding false stores in the database (false positives) than the uncaptured true stores (false negatives) and this overcounting leads to attenuated $\hat{\beta}_{w_s}$. The ideal ratio that gives minimum bias is achieved when $sen_s = \frac{ppv_s(ppv_s + Var(X)/E(X) - 2)}{Var(X)/E(X) - 1}$. Therefore, even though both sensitivity and PPV are high for source s , $\hat{\beta}_{w_s}$ can be biased depending on the values of $\frac{sen_s}{ppv_s}$ and $Var(X)/E(X)$. In our case, we simulated data while holding values of $\frac{sen_s}{ppv_s}$ at 0.77 ($s = 1$) and 0.81 ($s = 2$) for Figure 3.2(a); 1.30 ($s = 1$) and 1.24 ($s = 2$) for Figure 3.2(b).

3.5.4 Evaluation of proposed model

We evaluated our model when the true exposures are simulated according to the five mixture distributions in Section 3.5.2 and realistic values of sensitivity and PPV. Our simulations take into account two aspects of the model. First, given our model's use of known values for sensitivity and PPV, we additionally evaluate the consequences of misspecification of these values. In practice, the model misspecification occurs when working credibility parameters, denoted as ϕ_1^* and ϕ_2^* , differ from the true but unknown agreement statistics. In our case, given the observed exposures are generated with *realistic* true source credibility, the model was evaluated under five different sets of working parameters $\{\phi_1^*, \phi_2^*\}$. The five sets of working parameters are defined by multiplying the *realistic* sensitivity and PPV by (0.5, 0.75, 1, 1.25, 1.5) to generate extremely pessimistic, pessimistic, realistic (correctly specified), optimistic,

and extremely optimistic working models, respectively. Second, given that the model relies on the estimation of parameters of the latent exposure distribution, we compare the performance of the model when all parameters are estimated to the case when the parameters of the distribution of X are known. This comparison shows the effect of estimating λ_i 's on the estimation of regression parameters. Third, we show the benefits of using Dirichlet process priors for λ_i by comparing our model results to the model that uses a common Gamma(1,1) prior for all λ_i 's. Note that our main goal of using the DP prior is nonparametric density estimation of X , not inference on the number of components or cluster memberships. It is well known that the DPM performs well for density estimation under general conditions (Ghosal (2010)). Moreover, using a DP prior generates a mixture of Poisson distributions for the underlying X . A mixture of Poissons is a robust choice for modeling overdispersed count data (Hougaard et al. (1997); Joe and Zhu (2005)) and is consistent with our observed built-environment data. For each simulated dataset, we estimated $\bar{\alpha}$ for a concentration parameter as described in Section 3.4. For brevity, we present the results from mixture 1,3, and 5 only.

3.5.4.1 Sensitivity < PPV in both sources

We first demonstrate the estimation properties of our model when complete information about true exposure distribution, namely cluster assignments and cluster means, is given. Table 3.3 (middle) shows that having knowledge of the underlying distribution of latent X is particularly helpful with overdispersed X (columns labeled Mixture 3 and 5), as the model shows improved estimates of β_x compared to naïve approaches, even with extreme misspecifications of credibility parameters in either direction.

Table 3.3: ($sen_1 < ppv_1, sen_2 < ppv_2$) Posterior summaries of the proposed model for β_x when true source credibility is $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. True $\beta_x = 0.5$. The proposed model is evaluated with and without estimating X distribution. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications

Method	Mixture 1			Mixture 3			Mixture 5		
	Average			Average			Average		
	Naive Est.	SD	CP	Naive Est.	SD	CP	Naive Est.	SD	CP
	(%Bias)			(%Bias)			(%Bias)		
OLS, source 1	0.240 (-52.05%)	0.028	0%	0.540 (7.96%)	0.016	8%	0.580 (16.06%)	0.012	16%
OLS, source 2	0.313 (-37.40%)	0.026	0%	0.539 (7.86%)	0.014	8%	0.570 (14.04%)	0.011	14%
OLS, average	0.449 (-10.15%)	0.033	72%	0.602 (20.31%)	0.015	20%	0.614 (22.77%)	0.011	23%
Proposed Model									
when X distribution known									
Working credibility	Average			Average			Average		
	Posterior Mean	SD	CP	Posterior Mean	SD	CP	Posterior Mean	SD	CP
	(%Bias)			(%Bias)			(%Bias)		
XP	0.706 (41.18%)	0.042	0%	0.516 (3.18%)	0.014	76%	0.523 (4.68%)	0.018	82%
P	0.607 (21.31%)	0.044	28%	0.510 (2.06%)	0.013	88%	0.513 (2.65%)	0.016	95%
R (Correctly specified)	0.506 (1.12%)	0.037	91%	0.503 (0.68%)	0.013	97%	0.502 (0.36%)	0.014	98%
O	0.442 (-11.59%)	0.032	58%	0.496 (-0.82%)	0.012	93%	0.493 (-1.38%)	0.013	91%
XO	0.391 (-21.76%)	0.029	7%	0.486 (-2.86%)	0.012	76%	0.486 (-2.79%)	0.011	76%
Proposed Model									
when X distribution unknown									
Working credibility	Average			Average			Average		
	Posterior Mean	SD	CP	Posterior Mean	SD	CP	Posterior Mean.	SD	CP
	(%Bias)			(%Bias)			(%Bias)		
XP	0.709 (41.75%)	0.040	0%	0.651 (30.22%)	0.021	0%	0.632 (26.50%)	0.030	0%
P	0.605 (20.99%)	0.043	28%	0.576 (15.27%)	0.017	0%	0.558 (11.50%)	0.022	14%
R (Correctly specified)	0.504 (0.90%)	0.037	91%	0.532 (6.49%)	0.014	36%	0.521 (4.12%)	0.017	77%
O	0.441 (-11.71%)	0.032	56%	0.509 (1.76%)	0.013	91%	0.502 (0.41%)	0.014	97%
XO	0.401 (-19.81%)	0.028	16%	0.490 (-1.93%)	0.012	84%	0.491 (-1.87%)	0.012	81%

However, we typically do not have such information about the underlying distribution of X . We need to estimate the X density with the DPM sampler as described in Section 3.4. Table 3.3 (bottom) shows the posterior mean for the slope parameter β_x when the DPM sampler estimates the distribution of X (results for β_0 are available

in Table B.1). For Mixture 1 (Poisson X), the model performance is similar to the performance when X density is known since estimating density for the single cluster Poisson distribution is easier than estimating complex distribution such as Mixture 3 or Mixture 5. For these mixture distributions, the result in Table 3.3 shows that the optimistic working model has the smallest bias and a higher coverage rate. This result is likely due to bias that arises from the X density estimation via DPM sampler since the correctly specified model outperforms other misspecified models with Poisson distributed X . Yet, the DPM model has a strong advantage over common Gamma prior for λ_i 's by allowing flexible X distribution (Table B.2). Using common prior assumes incorrect equidispersion and leads to more biased estimates than naïve models.

When the working credibility parameters are much worse compared to the true values, the proposed model shows less correction, and sometimes results in estimates with higher bias compared to OLS. That is, when data sources are assumed to be much worse than they truly are, little can be gained from our proposed model. In summary, when the working credibility is at least at the level of true sensitivity and PPV, our model corrects bias (Table 3.3). With correctly specified values of working sensitivity and PPV, standard error sometimes increase on bias correction (Mixture 1 and Mixture 5), which is not uncommon in measurement error literature. However, the increase in variability is small compared to the reduction in bias.

3.5.4.2 Sensitivity > PPV in both sources

As shown in Section 3.5, the naïve model can already attain nearly unbiased estimation when sensitivity is less than PPV, depending on the distribution of the latent X . This section evaluates our model when sensitivity is larger than PPV; that is, when naïve OLS estimator is attenuated towards zero regardless of X distribution. The results from the previous subsection 3.5.4.1 imply that estimates from pessimistic

and optimistic working models are in between the realistic and the two extreme cases, so we only present the results of extremely pessimistic (XP), correctly specified (R), and extremely optimistic (XO) misspecification.

Table 3.4: $(sen_1 > ppv_1, sen_2 > ppv_2)$ Posterior summaries of the proposed model for β_x when true source credibility is $sen_1 = 0.48$, $sen_2 = 0.62$, $ppv_1 = 0.37$, and $ppv_2 = 0.5$. True $\beta_x = 0.5$.

Method	Mixture 1			Mixture 3			Mixture 5		
	Average			Average			Average		
	Naive Est.	SD	CP	Naive Est.	SD	CP	Naive Est.	SD	CP
	(%Bias)			(%Bias)			(%Bias)		
OLS, source 1	0.186 (-62.87%)	0.021	0%	0.350 (-30.10%)	0.009	0%	0.366 (-26.84%)	0.007	0%
OLS, source 2	0.252 (-49.67%)	0.021	0%	0.375 (-25.07%)	0.009	0%	0.388 (-22.48%)	0.007	0%
OLS, average	0.353 (-29.32%)	0.026	0%	0.389 (-22.24%)	0.009	0%	0.393 (-21.42%)	0.007	0%
Proposed Model									
when X distribution unknown									
Working credibility	Average			Average			Average		
	Posterior Mean	SD	CP	Posterior Mean	SD	CP	Posterior Mean	SD	CP
	(%Bias)			(%Bias)			(%Bias)		
XP	0.695 (38.98%)	0.041	2%	0.600 (20.09%)	0.018	0%	0.600 (20.02%)	0.026	0%
R (Correctly specified)	0.504 (0.83%)	0.037	95%	0.509 (1.76%)	0.013	91%	0.510 (1.93%)	0.015	96%
XO	0.393 (-21.30%)	0.029	7%	0.481 (-3.74%)	0.012	66%	0.487 (-2.60%)	0.011	72%

Table 3.4 shows that all naïve OLS estimators for the slope parameter are attenuated, as shown in Section 3.5.1. With the correct specification of working credibility, our model substantially reduces the bias and has a high coverage rate. Similar to Table 3.3, simple Poisson X is more sensitive to the choice of working credibility compared to the overdispersed X . However, as overdispersion is more common in the real data, our model will produce less biased estimates with the belief that true source credibility is similar to working credibility.

3.6 Application to children’s BMI study in California

The influence of local convenience stores and grocery stores on obesity risk among young children has been studied in many prior studies, using observed exposure estimates (W 's), typically from one data source only (Grafova (2008); Howard et al. (2011)). In this section, we use our model to obtain health effect estimates of the association between these two markers of the food environment near schools on childhood obesity among children in grades 5 and 7 attending urban schools in California (also see Section 3.2), in order to integrate information from both data sources and correct for ascertainment error. We conducted a stratified analysis by sex and grade, given that boys and girls have different obesity prevalence by age (Flegal et al., 2009), and each grade has different age groups. Although our main analysis is based on the stratification, we also performed a collapsed data analysis as well in order to compare the results. Among urban public schools in California, 1,773 schools and 1,758 schools provided BMI z-scores of male and female 5th graders, respectively. For 7th graders, 623 schools and 622 schools reported BMI z-scores of male and female students, respectively. Separate analyses were conducted for exposures to convenience stores and grocery stores.

Since the responses are mean BMI z-scores of schools and student sizes are different by the school, our outcome model needs weighting to adjust for the heterogenous error variance. In each stratum, a weighted linear model is used to regress school-level mean BMI z-scores on the count of food stores of interest. For easier interpretation, we divided the exposure variable by 10, so that ten stores become the unit used to interpret regression coefficients. We also adjusted for school-level demographic characteristics to the model, including schools’ percentage age groups, percentage of children who meet different fitness standards, and percentage of children’s ethnicity.

Based on exact match agreement statistics by Powell et al., both NETS and InfoUSA have sensitivity less than PPV for convenience stores (see Section 3.2 for source

credibilities of both sources). Hence the naive estimators of the effect of convenience stores on childhood BMI could be either attenuated, inflated, or even approximately unbiased. For supermarket/grocery stores, both databases have sensitivity $>$ PPV (Section 3.2). Therefore, the naive estimators of the effect of supermarket/grocery stores on BMI should be attenuated (we use the term ‘grocery stores’ to broadly include both supermarkets and grocery stores).

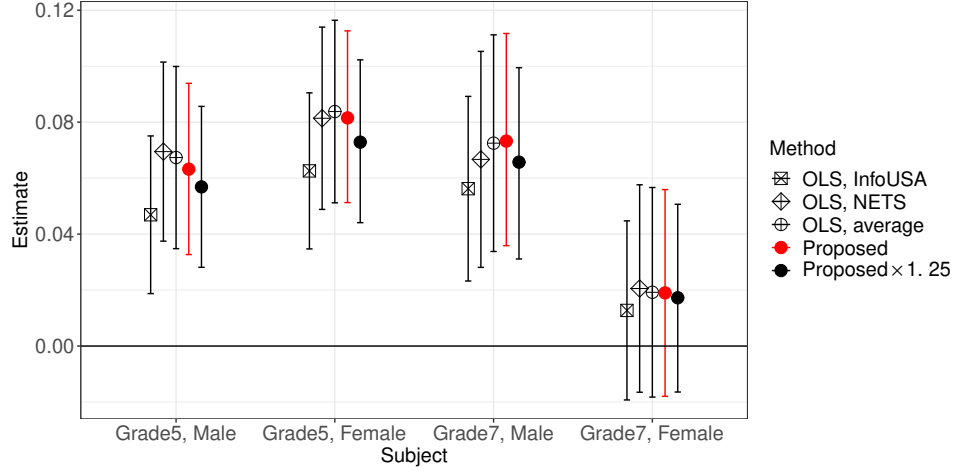
By following the SIC and NAICS codes chosen to compute the reference statistics (Powell et al. (2011)), 5th-grade schools had on average 4.56 (SD 3.55) convenience stores within 1 mile around school according to NETS in 2008. However, InfoUSA claims the mean convenience store exposure for these schools is 7.36 (4.12). Similarly, 7th-grade schools had mean exposure of 4.47 (3.38) according to NETS and 7.18 (3.93) according to InfoUSA. For supermarket/grocery stores, the mean exposure for 5th-grade schools is 19.98 (22.42) by NETS and 14.22 (16.86) by InfoUSA. Grade 7 had a mean exposure of 18.71 (18.84) by NETS and 13.15 (13.96) by InfoUSA.

Figure 3.3 shows the corrected results from our model and the naïve estimates for the effect of additional ten food stores on the school’s mean BMIz. In addition to reference values for working credibility, we present results from our model that use $1.25 \times$ reference values for working credibility (defined as optimistic misspecification in Table 3.3), based on simulation results in Table 3.3. The correction under optimistic credibility shows lower point estimates as we could expect from the simulation results in Table 3.3 and Table 3.4. The graphical posterior predictive checks shows that the replicated W ’s from both of our working models approximate the observed exposures, suggesting that our models fit the data well (Figure B.2).

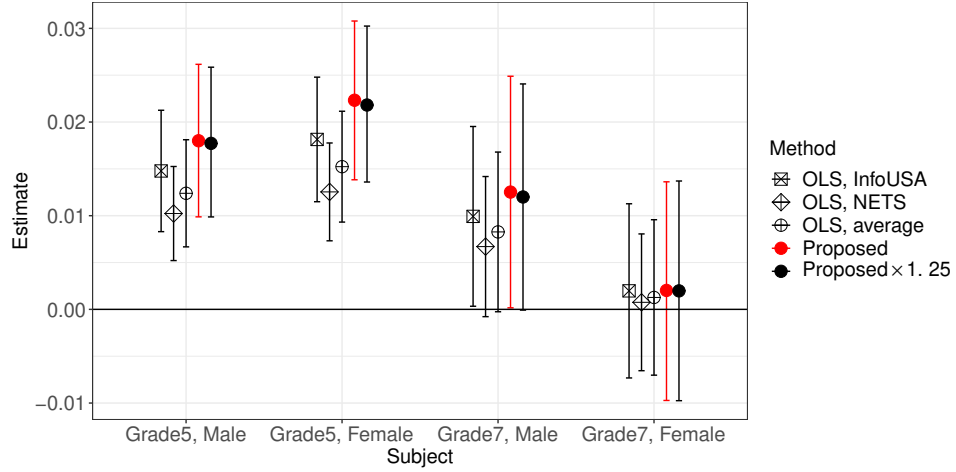
The results from our approach are qualitatively similar to the naïve analyses. Strong positive associations are found between exposure to convenience stores and school-level obesity risk, except for the female students in 7th grade. For the effect of grocery stores, we expect that naïve estimates are attenuated in all strata, given

the ratio of source credibilities. The corrected estimates from our model that use reference values for working credibility indicate that naïve use of secondary databases may fail to capture the meaningful association, as shown in the result from 7th-grade male students. For both working models, the proposed method finds a higher effect size for availability to grocery stores on 5th graders' obesity.

When we use all students without stratification, Figure 3.4 shows positive association between exposure and school-level BMIz for both convenience stores and grocery stores. Due to large sample size, all estimates have narrower confidence intervals. While stratified analyses in Figure 3.3 show weak differences in exposure effects by strata, collapsed analyses have not found significant effects of interactions between exposure and stratum factors from all naïve and proposed methods.

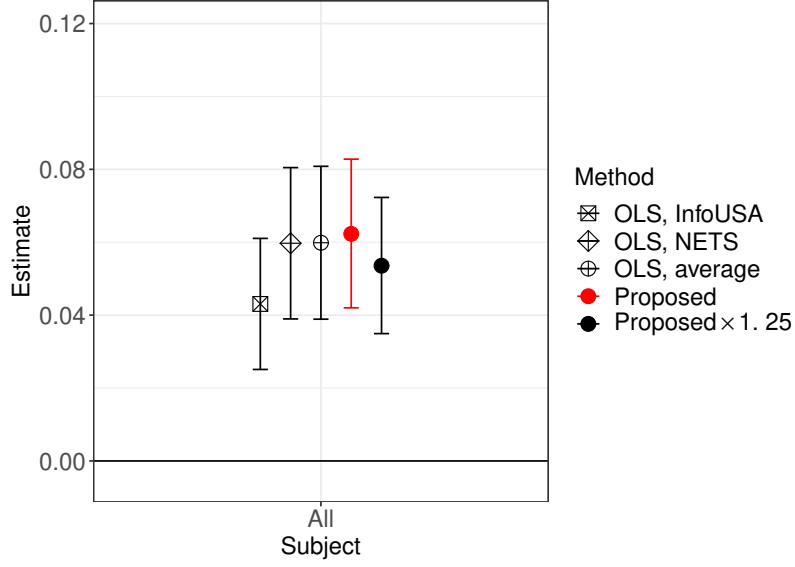


(a) Difference in BMIz per ten additional convenience stores

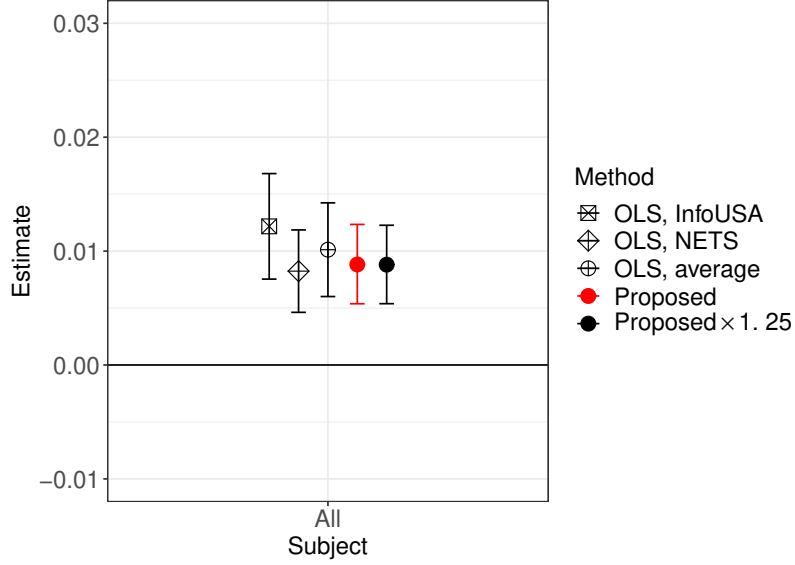


(b) Difference in BMIz per ten additional supermarket/grocery stores

Figure 3.3: Comparison of naïve OLS estimators for β_x and posterior mean of β_x from the proposed model under different misspecification. For naïve approaches, “OLS, InfoUSA” and “OLS, NETS” denote OLS estimators for simple linear regression that use $\mathbf{w}_{\text{InfoUSA}}$ and $\mathbf{w}_{\text{NETS}}$, respectively. “OLS, average” denotes an OLS estimator for a regression using average of $\mathbf{w}_1$ and $\mathbf{w}_2$. For the proposed model, reference values (denoted as “Proposed”) and reference values $\times 1.25$ (denoted as “Proposed $\times 1.25$ ”) are used for working credibility. Plotted are the naïve estimates with 95% confidence interval (black unfilled dots) and the posterior means with 95% credible intervals from our proposed model with reference values (red filled dot), and with $1.25 \times$ reference values (black filled dot).



(a) Difference in BMIz per ten additional convenience stores



(b) Difference in BMIz per ten additional supermarket/grocery stores

Figure 3.4: Comparison of naïve OLS estimators for β_x and posterior mean of β_x from the proposed model without data stratification. See caption of Figure 3.3 for details of each model.

3.7 Discussion

In this paper, we proposed a data integration method for cases when multiple exposure sources (e.g., business lists) yield different exposure measures of an otherwise

unobservable true exposure count. We also analytically and empirically showed the extent of bias in linear regression parameter estimates when a mismeasured exposure predictor is used. The bias in naïve estimates is linked to the unknown true exposure density in addition to the source quality measures. One advantage of our approach is that we can obtain a correct estimate of the health effect of exposure from multiple erroneous sources, without requiring gold-standard data.

Although our proposed method showed substantial bias correction, there are several shortcomings. First, the choice of working sensitivity and positive predictive value is critical. While the simulation study showed that the modest misspecification of source credibility still produces corrected estimates, we suggest finding validity measures for the same businesses with similar geographical and time contexts. As there are many validation studies available, obtaining a good reference for source credibility should not be extremely difficult. Our data example used source quality parameters from a single validation study that matched the type of location (urban) and time period. If multiple validation studies were available for the study location and time, it would be useful to extend our approach to incorporate source quality parameters from multiple validation studies. Furthermore, as confidence intervals of the sensitivity and PPV are also reported in the validation studies, researchers are recommended to use the upper and lower bound of the intervals to conduct sensitivity analyses in practice. Second, our method uses a fixed concentration parameter that is obtained from the observed exposures for the DPM model. Although we used a plug-in value for the concentration parameter, our assumption can be relaxed in our Bayesian method by proposing a prior distribution for α , for example, basing the mean and variance as a function of $\hat{\alpha}_1, \hat{\alpha}_2$. However, from our sensitivity analyses that tried drawing α from its posterior distribution with different priors for α , we found negligible gain in the correction of regression parameters (result not shown). Finally, our model relies on the conditional independence of the two sources, a com-

mon assumption in the measurement error literature (Richardson and Gilks, 1993a,b). However, this is a reasonable assumption in our data since the data vendors are competitors. When more data are available, it may be possible to relax the conditional independence assumption (Wang et al., 1996), though this would need to be addressed in future research.

Even with the limitations, our method has contributions to measurement error studies and built environment research. Unlike the existing measurement error literature, we incorporate the measurement error into the model without specifying a parametric model for the error variable. From a view of missing data problem, we showed that we can construct an identifiable model given the source credibility, although true exposure X is completely missing. Furthermore, our model specifies the marginal distribution of X in a data-adaptive way. Our work is one of few applications that use the DPM model for *latent* Poisson data (Dorazio et al. (2008)). The DPM model is well-suited for nonparametric Bayesian density estimation and it makes our model more generalizable by approximating overdispersed count data using Poisson mixtures. Our model may struggle with *underdispersed* X distributions, but this is unlikely to happen in built-environment studies. We anticipate that our work provides a guidance to handle a complicated count exposure variable that is prone to error.

We suggest some future directions of the proposed method. First, our outcome model can be extended with random effects or non-normal residual errors, although we used school-level responses with normality assumption for ecological analysis in this paper. These extensions require different Gibbs samplers and more complex computation, but conceptually, more general outcome models can be fitted. Second, there may be some situations where it may be desirable to infer subjects' cluster membership. Although not required in the context of our current approach, which only relies on the marginal distribution of the latent X , correctly identifying clusters would

require modifications for the prior on λ_i . Existing literature shows the traditional DPM has limitations for making inferences about clusters, and emerging approaches address this issue (Miller and Harrison (2014, 2018)). Third, our model can easily incorporate three or more lists. By adding one more source (even of medium quality), simulation studies show our model provides lower bias and higher efficiency (result not shown). As an increasing number of commercial lists provide point-referenced data of built environment features, our model would improve the inference with more sources.

We note that the practical concern about varying source quality by location. In our motivating example of urban schools, we assumed the source credibility stays constant for all school locations. This assumption was based on the inconsistent conclusion regarding a systematic relationship between source quality and area-level characteristics (Fleischhacker et al. (2013), Lebel et al. (2017), Powell et al. (2011), Liese et al. (2010), Bader et al. (2010)). Although there are mixed findings in source validity across different urbanization levels, high density areas are found to have slightly higher agreement (see Fleischhacker et al. (2013) and Lebel et al. (2017) for systematic reviews). For example, the metropolitan Chicago study by Powell et al. found convenience store and supermarket/grocery agreement were similar in rural and urban areas in Dun & Bradstreet, while convenience stores agreement by InfoUSA was lower in rural areas. Other validation study in South Carolina reported greater sensitivity of Dun & Bradstreet and InfoUSA in urban than rural areas (Liese et al., 2010). However, as we only used urban schools in this chapter, the agreement statistics should be roughly same for all schools. Another aspect of location-specific validity is neighborhood characteristics. A number of validation studies that investigated the association between database accuracy and neighborhood socio-economic factors or race/ethnicity composition have found little evidence for the different validity across neighborhood characteristics (Liese et al., 2010; Powell et al., 2011; Bader et al., 2010;

Cummins and Macintyre, 2009).

CHAPTER IV

Estimating the Effect of Latent Time-Varying Count Exposures Using Multiple Lists

4.1 Introduction

The majority of longitudinal built-environment health studies contextualize the population health within the neighborhood’s health-relevant resources over time. Recent studies have taken advantage of retrospective study designs with the benefit of having longer follow-up periods that typically rely on historical data on the business establishments provided by private data vendors (Zenk et al., 2018; Wang et al., 2019). For example, the association between obesity status of individuals with availability of food outlets that influence dietary choices is of interest. Although such secondary commercial databases often provide longitudinal point-referenced data, preceding studies raised concerns about measurement error when quantifying environmental influences with secondary commercial lists (Kaufman et al., 2015b; Hirsch et al., 2020). Counting the number of built environments within the neighborhood around subjects is one of the most widely used tools to quantify the environmental exposure (Berke et al., 2007; Rutt and Coleman, 2005). Imperfect listings of businesses in the commercial databases such as false listings of non-existing businesses or falsely omitted businesses yield undercount or overcount of exposures.

Ignoring the measurement error in exposures are well known to impair the estimation of exposure-outcome association (Greenland, 1980; Song and Huang, 2006). However, we often cannot remove the measurement error by directly correcting the databases as it is expensive to obtain primary data about true point-referenced environment features. Also, matching the business information across different databases is not permitted as they are proprietary data. Instead, we use aggregated level of information (counting measure). Then, we have replicated and error-prone measures of exposures from different databases. Figure 4.1 shows that the distribution of observed exposures differs between the databases and Figure 4.2 shows the disagreeing exposure measurements over time because one of the databases shows more change from year to year. Two figures indicate that the inference about exposure-outcome association can be inconsistent by the choice of database. As source quality of each listing can be found by external field validation studies (Powell et al., 2011; Liese et al., 2013; Fleischhacker et al., 2012), we suggest integrating the multiple observed exposures based on prior information about source quality in order to estimate unbiased association between the true exposures and time-varying health outcomes. Here, challenges arise with unknown temporal dependencies in true count exposures and time-varying measurement error mechanisms because source qualities are likely to change over years.

In the existing measurement error literature, methods for measurement errors in time-varying covariates have been developed with survival data models (Liao et al., 2011; Huang and Wang, 2000; Thomas, 2009). Integrating time-varying information from different sources has also been studied in the field of official statistics when multiple annual surveys produce conflicting estimates for the same quantity (Cheng et al., 2015; Lohr and Raghunathan, 2017). However, most of these studies use classical error assumption that comes with normality of measurement errors. This is not desirable in our case because we have time-varying count exposures and integer-

valued errors. Conceptually similar to misclassification errors in discrete covariates (Spiegelman et al., 2000; Greenland, 2000), we use trinomial distribution to model the relationship between observed and true exposures, and we allow cell probabilities vary over time to address the time-dependent measurement error mechanism.

Models for imperfect count data for a given period of follow-up have been developed in ecological studies that aim to estimate the abundance of species using the count of detected animals from repeated visits to multiple sites (Royle, 2004). To address this estimation problem, N-mixture models have been developed as an alternative to expensive abundance estimation tools such as capture-recapture (Royle, 2004; Dail and Madsen, 2011; Martin et al., 2011). Animal detection can be translated to business detection in our problem. Although the framework of using temporal ‘observed’ count to estimate latent true count data is similar to the N-mixture, the major differences between our work and the N-mixture model lie in (1) the goal of the model, (2) the modeling assumptions, and (3) the availability of data. While the N-mixture model is interested in accurate estimation of latent count data, we are further interested in estimating the association between the latent count and health outcomes. Also, the traditional N-mixture model assumes a closed population, i.e., abundance at each location does not change over time, and the absence of false positives, i.e., no false detection of ‘phantom’ animals. In contrast, we have dynamic population sizes (opening and closing businesses) and false positive listings. Furthermore, unlike the N-mixture model that uses a single observation for each location and time, we have multiple observations from different sources, health outcomes, and external information about the source quality.

In this study, we propose joint modeling of health outcomes, observed exposures, and latent exposures that incorporates source quality to integrate multiple longitudinal databases and provides a better estimate of health effects associated with changes in exposure to the nearby food environment. Bayesian framework is used to effectively

incorporate external information into the model and to deal with the heterogeneity in expected exposures and business dynamics by different subject locations. For flexible modeling of such heterogeneity in the data, we use Bayesian nonparametric approach. Having introduced the data in Section 4.2, we further describe our model in Section 4.3 and evaluate the proposed model in Section 4.4 with simulation studies. Our method is demonstrated with childhood obesity data in California from 2001 to 2008 in Section 4.5.

4.2 Exposure measurements using commercial lists

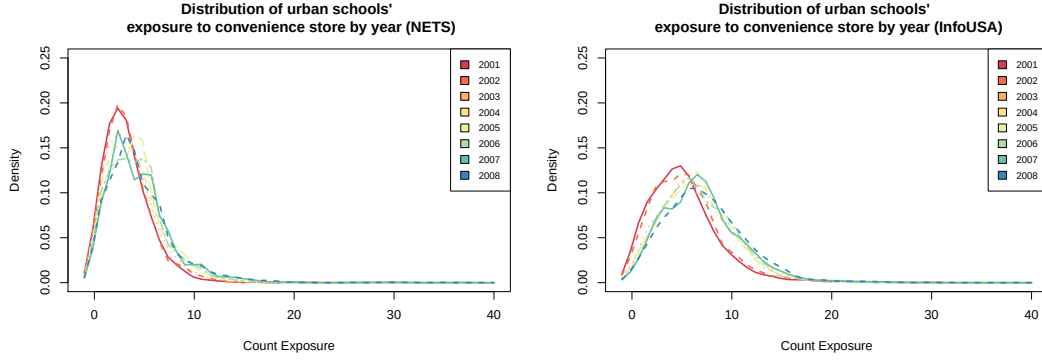
Our motivating example aims to estimate the association of childhood obesity with school neighborhoods' dynamic food environments, using convenience stores as an exposure marker. To compute school-level exposures, we count the number of convenience stores within 1-mile radius buffer around the school separately each year from 2001 to 2008. We also have the mean BMI z-scores at school-level that are calculated from cross-sectionally collected children's fitness data (California Department of Education, 2019).

We calculate the count exposure by using geocodes (latitude, longitude) of urban public schools in California ($N=1,834$) and geocodes of convenience stores listed in the National Establishment Time-Series (NETS) database and InfoUSA. The NETS database is a product derived from Dun & Bradstreet's (D&B) for longitudinal analysis. Both NETS and InfoUSA are commonly used secondary retail food outlet data sources that offer an annual register of business information. Two databases are designed to collect entire establishments each year, and both data companies provide business names, service types such as standard industrial classification (SIC) codes, as well as geographic locations.

The discrepancy between the two databases is easily shown in Figure 4.1 and Figure 4.2. Figure 4.1 presents the overlay of annual exposure density from 2001 to

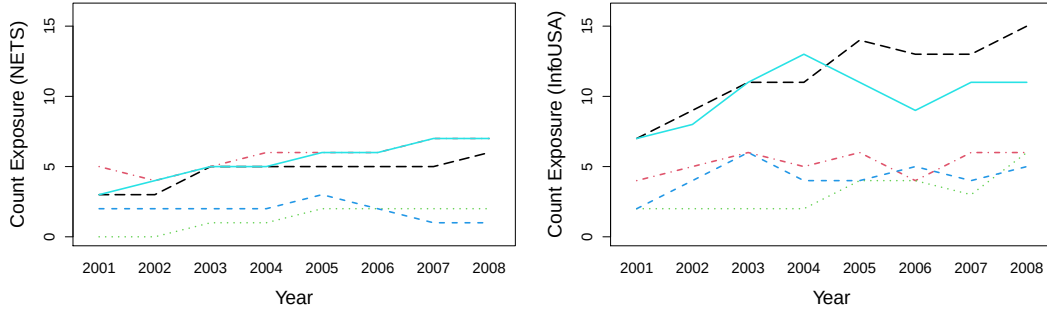
2008. In both sources, exposure density shifts to the right from the baseline (red to blue), implying increased availability of convenience stores in many schools although the NETS database shows less amount change than InfoUSA. Figure 4.2 shows the discrepancy between two sources for the school-level trajectory of exposures in a given period of time. By taking five randomly sampled schools as an example, the figure implies that the conflict between exposures from two databases occurs almost every year.

The incompleteness of commercial databases is one major reason for the discordance in exposure measurements between data sources. A rich amount of validation studies compare the commercial databases with ground observations from field studies, or from gold-standard data such as local government databases (Powell et al. (2011); Liese et al. (2013)). The agreement between the true availability and listings on commercial databases is measured by the number of correctly listed businesses (true positive), incorrectly omitted businesses (false negative), and incorrectly listed businesses (false positive). These measures are used to calculate agreement statistics: (1) sensitivity: the proportion of ground-truthed establishments that were also listed in the commercial database and (2) positive predictive value (PPV): the proportion of establishments listed in the commercial database that were also observed on the ground. In the following section, we describe our model to integrate these conflicting exposures using source credibilities to obtain a valid inference for the health effect of latent historical true exposures.



(a) Exposure density based on NETS (b) Exposure density based on InfoUSA

Figure 4.1: Annual distribution of exposure to convenience stores between the calendar year 2001 to 2008, for $N=1,834$ urban schools in California, using NETS and InfoUSA.



(a) Exposure to convenience stores based on NETS (b) Exposure to convenience stores based on InfoUSA

Figure 4.2: The number of convenience stores within 1-mile buffer by year, for randomly sampled five schools. Each school is represented by a colored line. The convenience store is defined as establishments with SIC code 541102 (convenience stores), 55410000 (gasoline service stations), 55419901 (filling stations, gasoline), and 55419903 (truck stops) are treated as convenience stores in the NETS database. For the InfoUSA database, convenience stores are defined as establishments that have NAICS codes 541103 (convenience stores), 554101 (service stations—gasoline and oil), 554102 (gas—diesel), and 554103 (truck stops and plazas).

4.3 Modeling strategy

For each school $i = 1, \dots, n$, we observe continuous health outcome $\{Y_{it}\}$ and observed exposures $\{W_{it}^A\}$ and $\{W_{it}^B\}$ from source A and source B, respectively, for

equally spaced measurement times $t = 1, \dots, T$. Let X_{it} denote the true, unobserved outlet count within a one mile radius buffer of i th subject (e.g., school) at time t . Here we interchange the term subject and site, as the subject stays at the fixed location. For simplicity of outcome model, we assume a simple multivariate linear regression model: $Y_{it} = \beta_0 + \beta_x X_{it} + \epsilon_{it}$ with $(\epsilon_{i1}, \dots, \epsilon_{iT}) \sim MVN_T(\mathbf{0}, \Sigma_i)$. In the real data application in Section 4.5, we assumed $\Sigma_i = \Sigma$ with the autoregressive covariance structure of order 1 (AR(1)) with a standard deviation σ and a correlation parameter ρ . We will use this covariance assumption for the rest of this section as well. Our primary interest is in estimation of β_x .

4.3.1 Measurement error model

Table 4.1: Contingency tables comparing truth and observed exposures from two databases ($s = A, B$) for subject i at time t . Each table has TP_{it}^s (true positive of subject i at time t from source s), FN_{it}^s (false negative), FP_{it}^s (false positive), and corresponding cell probabilities $(\theta_{11t}^s, \theta_{12t}^s, \theta_{21t}^s)$.

(a) Database 1				(b) Database 2			
	Listed	Not Listed	Total		Listed	Not Listed	Total
Exist	$TP_{it}^A (\theta_{11t}^A)$	$FN_{it}^A (\theta_{12t}^A)$	X_{it}	Exist	$TP_{it}^B (\theta_{11t}^B)$	$FN_{it}^B (\theta_{12t}^B)$	X_{it}
Not exist	$FP_{it}^A (\theta_{21t}^A)$	-		Not exist	$FP_{it}^B (\theta_{21t}^B)$	-	
Total	W_{it}^A			Total	W_{it}^B		

At each location at time t , we can model $(W_{it}^A, W_{it}^B, X_{it})$ with two trinomial contingency tables with structural zeros that are not observable by the study design. Since two data vendors collect business data independently, we assume conditional independence of the two contingency tables in Table 4.1 given the true exposure X_{it} . We allow the sensitivity and PPV of each data source change over time, but are not site-specific, i.e., $sen_{it} = sen_t, ppv_{it} = ppv_t$. Although one might consider that different source quality by the subjects due to geographic characteristics, there is inconsistent findings about association between the source quality and neighborhood

characteristics such as census-tract level socio-economic status or urbanicity (Lebel et al., 2017). Therefore, we let sensitivity and PPV values are shared across all individuals. Then, the cell probabilities in each table at time t for any one location can be written as a function of sen_t and ppv_t : $\theta_{11t}^s = \frac{1}{(1-ppv_t^s)/ppv_t^s + (1-sen_t^s)/sen_t^s + 1}$, $\theta_{12t}^s = \frac{(1-sen_t^s)/sen_t^s}{(1-ppv_t^s)/ppv_t^s + (1-sen_t^s)/sen_t^s + 1}$, $\theta_{21t}^s = \frac{(1-ppv_t^s)/ppv_t^s}{(1-ppv_t^s)/ppv_t^s + (1-sen_t^s)/sen_t^s + 1}$ where the superscript s stands for source ($s = A$ or B).

The role of source quality in the relationship between the W_{it}^s and X_{it} can be explained by the familiar measurement error framework. Under additive error assumption, $W_{it}^s = X_{it} + (FP_{it}^s - FN_{it}^s)$ where $(FP_{it}^s - FN_{it}^s)$ is a measurement error. Consider Poisson sampling framework for contingency tables in Table 4.1 and let $X_{it} \sim Pois(\lambda_{it})$. Since FP_{it}^s and FN_{it}^s are independent Poisson random variables, the error has Skellam distribution. The Skellam distribution is the discrete probability distribution of the difference of two independent Poisson random variables with mean m_1 and m_2 . The probability mass function of Skellam distribution is $p(k; m_1, m_2) = \exp\{-(m_1 + m_2)\} \left(\frac{m_1}{m_2}\right)^{k/2} I_k(2\sqrt{m_1 m_2})$, where the $I_k(\cdot)$ is the modified Bessel function and the support of k is $\{\dots, -2, -1, 0, 1, 2, \dots\}$ (Skellam, 1946). The mean and variance of Skellam distribution are $(m_1 + m_2)$ and $(m_1 - m_2)$, respectively. The error $FP_{it}^s - FN_{it}^s$ follows Skellam distribution with mean $(\frac{sen_t^s}{ppv_t^s} - 1)\lambda_{it}$ and variance $(\frac{sen_t^s}{ppv_t^s} - 1 + 2(1 - sen_t^s))\lambda_{it}$.

Therefore, if $sen_t^s < ppv_t^s$, measurement error is negative on average and undercount is more likely. On the other hand, when $sen_t^s > ppv_t^s$, the observed count exposure likely overcounts the true exposure. When $sen_t^s = ppv_t^s$, $E(W_{it}^s | X_{it}) = X_{it}$ and the observed count is unbiased for the true count due to false positive and false negative counts canceling on average.

Note that this measurement error model violates the classical independence assumption between the measurement error and the truth, and thus we cannot apply traditional measurement error models. Instead, we can specify the joint distribution

of (W_{it}^A, X_{it}) as a bivariate Poisson. The conditional distribution of W_{it}^A given X_{it} is then specified by integrating out the true positive:

$$f(W_{it}^A; X_{it}, sen_t^A, ppv_t^A, \lambda_{it}) = \sum_{TP_{it}^A=0}^{\min\{X_{it}, W_{it}^A\}} Binom(TP_{it}^A; X_{it}, sen_t^A) \times Pois\left(W_{it}^A - TP_{it}^A; \left(\frac{sen_t^A}{ppv_t^A} - sen_t^A\right) \lambda_{it}\right). \quad (4.1)$$

For source B, $f(W_{it}^B; X_{it}, sen_t^B, ppv_t^B, \lambda_{it})$ can be similarly constructed. If there was no false positive in the listings so that $W_{it}^s = TP_{it}^s$, our problem resembles the assumed data in Royle's N-mixture model. Although the false positives makes our data more complicated by adding either positive or negative measurement errors, we have multiple databases which will help correct for the consequences of the measurement errors. With the conditional independence assumption between observed measurements given the true exposures, we can model observed exposures given the truth as

$$f(W_{it}^A, W_{it}^B; X_{it}, sen_t^A, ppv_t^A, sen_t^B, ppv_t^B, \lambda_{it}) = f(W_{it}^A; X_{it}, sen_t^A, ppv_t^A, \lambda_{it}) \times f(W_{it}^B; X_{it}, sen_t^B, ppv_t^B, \lambda_{it}) \quad (4.2)$$

for subject i at time t .

4.3.2 Dependencies in true exposures

We aim to model both heterogeneity of latent longitudinal count data $\mathbf{X}_i = (X_{i1}, \dots, X_{iT})$ across different subjects and serial correlations within the count series. We define $\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\lambda}_i, \mathbf{w}_i^A, \mathbf{w}_i^B$ as length- T vectors of observed outcomes, latent true exposures, expected true exposures, and observed exposures of i th school during the time $t = 1, \dots, T$. Using the measurement error model described in the previous

section, the school-specific full likelihood is given as:

$$\begin{aligned}
& L(\mathbf{X}_i, \boldsymbol{\beta}, \sigma, \rho, \text{sen}_t^A, \text{ppv}_t^A, \text{sen}_t^B, \text{ppv}_t^B, \boldsymbol{\lambda}_i | \mathbf{Y}_i, \mathbf{W}_i^A, \mathbf{W}_i^B) \\
& = f(\mathbf{Y}_i; \mathbf{X}_i, \boldsymbol{\beta}, \sigma, \rho) \left(\prod_{t=1}^T f(W_{it}^A; X_{it}, \text{sen}_t^A, \text{ppv}_t^A, \lambda_{it}) f(W_{it}^B; X_{it}, \text{sen}_t^B, \text{ppv}_t^B, \lambda_{it}) \right) \times \\
& \quad f(\mathbf{X}_i; \boldsymbol{\lambda}_i).
\end{aligned} \tag{4.3}$$

Instead of independent exposures, $f(\mathbf{X}_i; \boldsymbol{\lambda}_i) = \prod_{t=1}^T \text{Pois}(X_{it}; \lambda_{it})$, we need to model time dependencies of true exposures because X_{it} is likely dependent to the past exposure X_{it-1} . A class of Poisson autoregressive models can be one potential approach for time series of counts (Brandt and Williams, 2001; Held et al., 2006; Fokianos and Tjøstheim, 2011). We extend the integer-valued Poisson autoregressive model, first introduced by (Alzaid and Al-Osh, 1988), as it was effectively used in modeling multiple time-series counts data (Böckenholt, 1998; Aldor-Noiman et al., 2016) and the conceptually similar migration decomposition approach was used for an open population in the generalized N-mixture model to allow birth and immigrants (Dail and Madsen, 2011).

We specify the conditional distribution X_{it} given X_{it-1} after the baseline as

$$\begin{aligned}
X_{i1} & \sim \text{Pois}(\lambda_{i1}), \\
X_{it} | X_{it-1} & = p \circ X_{it-1} + e_{it}, \quad t > 1, \\
e_{it} & \sim \text{Pois}(\mu_i).
\end{aligned} \tag{4.4}$$

The $p \circ X_{it-1}$ in equation (4.4) models dependencies between past and present exposures by the binomial thinning operator $\circ$. The binomial thinning operation gives $p \circ X_{it-1} = \text{Bin}(X_{it-1}; p)$ where p represents the probability of business survival from $t-1$ to t and $p \in [0, 1]$. Another part, e_{it} , is a Poisson random variable with rate μ_i which reflects the rate at which new stores open near site i since time

$t - 1$ (hereafter we refer to this as store opening rate). Here, we assume the independence between the new arrivals e_{it} and X_{it-1} . Then, combined with our assumption about the baseline $f(X_{i1}; \lambda_{i1}) \sim \text{Pois}(\lambda_{i1})$, the marginal distribution of X_{it} is Poisson distribution because of additivity of Poisson distribution and the fact that Poisson distribution is invariant to binomial thinning (Puig and Valero, 2007). The mean parameter of marginal X_{it} can be written as $\lambda_{it} = p^{t-1}\lambda_{i1} + \frac{1-p^{t-1}}{1-p}\mu_i$ for $t > 1$ (Proof available in Appendix C.1). The Markov assumption leads to $f(\mathbf{X}_i; \mu_i, \lambda_i, p) = f(X_{i1}; \lambda_{i1}) \prod_{t=2}^T f(X_{it}; X_{it-1}, p, \mu_i)$ and the $f(X_{it}; X_{it-1}, p, \mu_i)$ is a discrete convolution of Binomial and Poisson: $\sum_{e=0}^{\min\{x_{it}, x_{it-1}\}} \text{Bin}(x_{it} - e; x_{it-1}, p) \text{Pois}(e; \mu_i)$. Unlike the classic INAR(1) process, we allow non-stationarity of the series so that the mean and variance are changing over time as we could see the changing exposure distribution from Figure 4.1.

In addition to time dependencies *within* a school's exposures over time, we need a flexible modeling of expected baseline exposure λ_{i1} and arrival rate μ_i instead of using simple Poisson distribution with $\lambda_{i1} = \lambda$ and $\mu_i = \mu$. To allow heterogeneity of school-specific expected baseline exposure, we impose the Dirichlet process (DP) prior on the λ_{i1} (Ferguson, 1973). The Dirichlet process, $DP(\alpha, G_0)$, with baseline distribution G_0 and concentration parameter α provides a distribution over countably infinite probability measures (Teh et al., 2010). The resulting Dirichlet process mixture model can be considered as a limit of a finite mixture model (Neal, 2000). The DP prior groups the schools according to their corresponding λ_{i1} into an unknown number of clusters and schools with same group have a common value of λ_{i1} . Although Dirichlet process mixture models are popularly used for clustering problems, our model uses DP priors to estimate density nonparametrically, not to find true clusters.

We also separately model the heterogeneity of arrival rate μ_i with a similar manner. Independent processes are assumed for λ_{i1} and μ_i because we do not have any prior knowledge about the relationship between the two parameters. However, when

the arrival rate is assumed to be dependent on the initial expected exposure, we can consider modeling μ_i as a function of λ_{i1} or use site-specific covariates as in animal population research (Dail and Madsen, 2011). The DP prior for μ_i has been previously studied to estimate the location-specific crime rates in crime event time-series data and it is shown to well perform in different values of rates (Aldor-Noiman et al., 2016). The DP priors allow nonparametric formulation of business dynamics given survival rate p . The two independent DP priors are given as follows.

$$X_{i1} \sim \text{Pois}(\lambda_{i1}), \lambda_{i1} \sim G, G \sim DP(\alpha, G_0), \quad (4.5)$$

$$e_{it} \sim \text{Pois}(\mu_i), \mu_i \sim G', G' \sim DP(\alpha', G'_0). \quad (4.6)$$

The Dirichlet process requires the specification of a base measure and a concentration parameter. We choose base measures $G_0 = \text{Gamma}(1, 1)$ and $G'_0 = \text{Gamma}(1, 1)$ for the conjugacy. Prior of concentration parameters α and α' are specified as $\text{Gamma}(a_0, b_0)$ and $\text{Gamma}(a'_0, b'_0)$ as suggested by Escobar and West (1995). In particular, we assume exposures are overdispersed across the schools, but fewer unique patterns in exposure trends are found. Hence, we set $a_0 = a'_0 = 1$, $b_0 = 1$, and $b'_0 = 10$. The computational algorithm is based on the Chinese restaurant process (CRP) and detailed steps are available in Appendix C.2 (Teh et al., 2010; Neal, 2000).

We also tried an additional DP prior for the site-specific p_i , but the model was unstable due to an identifiability issue. Conditional on unknown X_{it-1} , distinction among different representations of $p_i \circ X_{it-1}$ is difficult without a common site effect governing their mean (Barker et al., 2018). To alleviate this issue, we instead fixed the autocorrelation p across locations while allowing location-specific μ_i 's as in the generalized N-mixture model (Dail and Madsen, 2011). However, a logit model for p_i can be considered to model additional spatial variation when site-specific covariates are available (Böckenholt, 1998; Kery et al., 2009).

4.3.3 Model for time-varying source quality

Given the historical nature of the data, it is sensible to assume the different quality of commercial lists over time. A practical concern arises as not all values of sensitivity and PPV throughout the study time period are provided by external field studies. Instead, we have partial knowledge of agreement statistics for a few time points. In this paper, we assume the source agreement statistics are available for the last year T of the study period (i.e., estimates for sen_T , ppv_T are available).

Although partial knowledge of the multinomial cell entries such as Table 4.1 causes a problem in identifiability (Swartz et al., 2004; Neath and Samaniego, 1997), Bayesian approaches with an introduction of the prior information are shown to rectify the nonidentifiability problem and make a reasonable inference (Neath and Samaniego, 1997; Gustafson, 2005). Therefore, under the Bayesian framework, we construct a prior distribution for $(sen_t^s, ppv_t^s), t = 1, \dots, T$. Here, note that by including T , we assume that true values of last year's source quality can be different from reference values because those values are not from our direct observation. However, the last year's reference values help us model informative priors instead of flat priors. We use the identical and independent prior distributions for each year's source quality using the last year's reference values because we believe that historical source qualities are not largely deviated from the last year's value and we do not know underlying trends in sensitivity or PPV. One might also extend the prior with constraints such as strictly increasing sensitivity or PPV (e.g., $sen_1 < sen_2 < \dots < sen_T$). As discussed in Chapter III, however, the information about changes in source quality is not known as there is no field studies that examined annual source quality for the same or similar geographic area (Fleischhacker et al., 2013; Lebel et al., 2017). Therefore, we assumed independence between annual source qualities to have flexibility in the scope of this paper.

Both sensitivity and PPV share the θ_{11t}^s because $sen_t^s = \frac{\theta_{11t}^s}{\theta_{11t}^s + \theta_{12t}^s}$ and $ppv_t^s =$

$\frac{\theta_{11t}^s}{\theta_{11t}^s + \theta_{21t}^s}$ by definition. Therefore, we consider the joint prior distribution $\pi(\theta_{11t}^s, \theta_{12t}^s, \theta_{21t}^s)$ instead of independent priors for sensitivity and PPV. To model the joint prior, we use reparameterization of a Dirichlet distribution into Gamma distribution (Narayanan, 1990). We define three independent standard Gamma variables $U_{kt}^s \sim \text{Gamma}(\gamma_{kt}^s, 1) = \text{Gamma}(\gamma_k^s, 1), k = 1, 2, 3$ for each cell in Table 4.1. As previously described, we use identical prior distribution for all time points, so that $\gamma_{kt}^s = \gamma_k^s$ for $k = 1, 2, 3$. Then, $(\theta_{11t}^s, \theta_{12t}^s, \theta_{21t}^s) = (U_{1t}^s/R_t, U_{2t}^s/R_t, U_{3t}^s/R_t)$ where $R_t = \sum_k U_{kt}^s$, follows Dirichlet distribution with parameters $(\gamma_1^s, \gamma_2^s, \gamma_3^s)$. This reparameterization gives $sen_t^s = \frac{U_{1t}^s}{U_{1t}^s + U_{2t}^s} \sim \text{Beta}(\gamma_1^s, \gamma_2^s)$ and $ppv_t^s = \frac{U_{1t}^s}{U_{1t}^s + U_{3t}^s} \sim \text{Beta}(\gamma_1^s, \gamma_3^s)$. Since we can have reference values for sensitivity and PPV of year T , we can obtain $(\gamma_1^s, \gamma_2^s, \gamma_3^s)$ by matching the mean and standard errors of the reference values. Choosing hyperparameters that match the prior guess is one recommended strategy for nonidentifiable models (Gustafson et al., 2001). Sampling of (sen_t^s, ppv_t^s) is then fairly straightforward using Metropolis-Hastings algorithm (Details available in Appendix C.2).

4.4 Simulation study

We conducted a simulation study to evaluate the ability of our proposed method to reduce the bias in exposure effect estimation due to the presence of measurement errors. The simulation used a full factorial design with two different effect sizes for β_x , two values for the number of follow-ups, T , three different values for the business survival probability p , and two different settings of the relative size of sensitivities to PPVs, giving 24 different scenarios.

True exposure data are simulated with $N = 300$ subjects, $T = \{4, 8\}$ longitudinal observations per subject, and with different survival rate values $p = \{0.5, 0.8, 0.99\}$. In all scenarios, we use two mixture components to generate baseline exposure, i.e., λ_{i1} taking the value of either 5 or 15, with mixing probability 0.9 and 0.1. Store opening rates are fixed to either of two realistic values, $\mu_i = (0.5, 1.5)$ with equal

mixing probabilities. For example, $\mu_i = 0.5$ represents one new store is expected to open every two years within the subject's 1-mile buffer area. Therefore, we have total four different combinations of (λ_{i1}, μ_i) for individual exposure trajectory for each scenario. True exposure trends with different business survival probabilities, p , are shown in Appendix C.1. Continuous outcome data are then generated according to the model in Section 4.3 with $\beta_0 = 0.85$, $\beta_x = \{0.05, 0.5\}$, $\rho = 0.5$, and $\sigma = 0.35$, with AR(1) correlation within the repeated measurements.

Source credibilities at time T are set to $(sen_T^A, ppv_T^A, sen_T^B, ppv_T^B) = (0.38, 0.48, 0.5, 0.62)$ based on published sensitivity and PPV values of NETS and InfoUSA for convenience stores in 2008 (Powell et al., 2011). The remaining $T - 1$ years' sensitivities and PPVs are randomly sampled from Beta distributions with the mean centered around $(sen_T^A, ppv_T^A, sen_T^B, ppv_T^B)$ and standard deviations that match the standard errors of sample proportions. Observed exposures are then generated from the trinomial distributions in Table 4.1.

Although $sen < ppv$ is dominant according to validation studies (Lebel et al., 2017), we also investigate the impact of having more false positives than false negatives. Another set of observed exposures are generated with sensitivities and PPVs that are sampled with reversed values, i.e., $(sen_T^A, ppv_T^A, sen_T^B, ppv_T^B) = (0.48, 0.38, 0.62, 0.5)$. Reversing sensitivity and PPV values exchanges the probability of false positives and false negatives, while maintaining the probability for true positives. Other scenarios with conflicting agreement such as $(sen_t^A < ppv_t^A, sen_t^B > ppv_t^B)$ can exist, but we found that the relative size of sensitivity to PPV is mostly consistent across different data sources (Lebel et al., 2017; Powell et al., 2011). Hence, we only present $(sen_T^A < ppv_T^A, sen_T^B < ppv_T^B)$ or $(sen_T^A > ppv_T^A, sen_T^B > ppv_T^B)$.

Additionally, a natural concern may arise with prior misspecification, especially when the true source qualities take values different than expected by the external information. Therefore, we run correctly and incorrectly specified priors for

source credibility for each of the 24 scenarios. Correctly specified priors use true $(sen_T^A, ppv_T^A, sen_T^B, ppv_T^B)$ to construct priors, whereas incorrectly specified models use values that are 20% higher or lower than the truth so that $(sen_T^A \times 1.2, ppv_T^A \times 0.8, sen_T^B \times 0.8, ppv_T^B \times 1.2)$ are used to calculate hyperparameters for priors. True sensitivities and PPVs with prior densities from correct and incorrect information are displayed in Appendix C.2.

The model performance was evaluated by comparing root mean square error (RMSE) for our target coefficient β_x , based on 50 simulated dataset each scenario. Estimates for β_x are obtained from the proposed methods and naïve generalized least square methods using either a single-sourced exposure predictor (only source 1 is presented) or an averaged exposure predictor. Using the Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992), multivariate potential scale reduction factors were all near 1.1 after 200,000 iterations based on five independent Markov chains with random starting values, indicating convergence. The convergence was checked for a subset of scenarios ($\beta_x = 0.05, T=8$). Simulation results from 24 scenarios are summarized in Table 4.2.

4.4.1 Simulation results

Table 4.2: Summaries of the sampling distribution of β_x based on 50 simulated data sets. Root mean square error (RMSE), percent bias, and coverage probability (CP) are presented for each scenario. $N=300$ sites, $T=\{4,8\}$, and business dynamics $p = \{0.99, 0.8, 0.5\}$, $\mu_i = (0.5, 1.5)$, and $\lambda_{i1} = (5, 15)$. Regression parameters are $\beta_0 = 0.85$, $\beta_x = \{0.05, 0.5\}$ $\sigma = 0.35$, and $\rho = 0.5$.

	β_x	T	p	Naïve: source 1			Naïve: averaged W			Correctly specified prior			Misspecified prior		
				RMSE	%Bias	CP	RMSE	%Bias	CP	RMSE	%Bias	CP	RMSE	%Bias	CP
sen<ppv	0.05	4	0.5	0.0154	-29.8%	4%	0.0044	1.3%	96%	0.0042	-1.5%	98%	0.0064	-10.2%	92%
sen<ppv	0.05	4	0.8	0.0273	-54.0%	0%	0.0122	-22.8%	24%	0.0073	10.0%	88%	0.0052	2.5%	96%
sen<ppv	0.05	4	0.99	0.0248	-49.3%	0%	0.010	-18.9%	20%	0.0039	-4.6%	92%	0.0046	-6.9%	92%
sen<ppv	0.05	8	0.5	0.0205	-40.4%	0%	0.0065	-10.3%	76%	0.0053	6.6%	92%	0.0042	-3.4%	94%
sen<ppv	0.05	8	0.8	0.0317	-63.2%	0%	0.0164	-32.3%	0%	0.0059	8.4%	88%	0.0039	1.4%	96%
sen<ppv	0.05	8	0.99	0.0292	-58.3%	0%	0.0131	-25.6%	0%	0.0030	3.4%	94%	0.0025	-1.5%	95%
sen<ppv	0.5	4	0.5	0.1414	-27.8%	0%	0.0439	7.6%	20%	0.0476	9.1%	15%	0.0220	3.4%	88%
sen<ppv	0.5	4	0.8	0.3945	-78.9%	0%	0.2547	-50.8%	0%	0.0459	8.8%	18%	0.0217	3.4%	84%
sen<ppv	0.5	4	0.99	0.4132	-82.6%	0%	0.3225	-64.4%	0%	0.0162	1.6%	92%	0.0174	-2.1%	88%
sen<ppv	0.5	8	0.5	0.2092	-41.7%	0%	0.0187	-2.2%	60%	0.0185	2.9%	74%	0.0152	-2.2%	82%
sen<ppv	0.5	8	0.8	0.4159	-83.2%	0%	0.3044	-60.8%	0%	0.0230	4.1%	58%	0.0104	-0.7%	98%
sen<ppv	0.5	8	0.99	0.4413	-88.3%	0%	0.3790	-75.8%	0%	0.0179	2.6%	66%	0.0134	-1.5%	83%
sen>ppv	0.05	4	0.5	0.0230	-45.6%	0%	0.0150	-29.6%	0%	0.0047	-4.2%	92%	0.0061	-9.1%	88%
sen>ppv	0.05	4	0.8	0.0310	-61.7%	0%	0.0201	-39.6%	0%	0.0055	-6.3%	92%	0.0068	-10.7%	76%
sen>ppv	0.05	4	0.99	0.0293	-58.5%	0%	0.0197	-39.1%	0%	0.0036	-0.3%	98%	0.0047	-5.7%	94%
sen>ppv	0.05	8	0.5	0.0258	-51.3%	0%	0.0154	-30.1%	0%	0.0046	-1.7%	92%	0.0057	-7.3%	82%
sen>ppv	0.05	8	0.8	0.0346	-69.1%	0%	0.0236	-46.8%	0%	0.0068	-10.9%	60%	0.0085	-15.2%	42%
sen>ppv	0.05	8	0.99	0.0317	-63.3%	0%	0.021	-41.8%	0%	0.0041	-6.0%	70%	0.0063	-11.4%	40%
sen>ppv	0.5	4	0.5	0.2151	-42.9%	0%	0.1369	-27.3%	0%	0.0493	-9.6%	6%	0.0626	-12.2%	0%
sen>ppv	0.5	4	0.8	0.4039	-80.7%	0%	0.2710	-54.0%	0%	0.0478	-9.2%	12%	0.0665	-13.1%	0%
sen>ppv	0.5	4	0.99	0.4151	-83.0%	0%	0.3307	-66.1%	0%	0.0420	-7.9%	28%	0.0626	-12.2%	0%
sen>ppv	0.5	8	0.5	0.2495	-49.8%	0%	0.1367	-27.3%	0%	0.0559	-11.0%	0%	0.0731	-14.5%	0%
sen>ppv	0.5	8	0.8	0.4281	-85.6%	0%	0.3306	-66.1%	0%	0.0731	-14.5%	0%	0.0917	-18.2%	0%
sen>ppv	0.5	8	0.99	0.4470	-89.4%	0%	0.3874	-77.5%	0%	0.0645	-12.7%	0%	0.0852	-16.9%	0%

Naïve results show that RMSE is larger when we have a higher true effect size for β_x as it makes the model more sensitive to errors in exposures. Our model performance such as coverage probability is also limited by the large effect size, but we are not too concerned about this finding because the effect size is likely to be very small in

the real data (< 0.05). Longer T and higher autocorrelation ρ generally increases naïve RMSE, as successive measurement errors for a longer sequence are more likely to make observed exposure trajectory far from the true data and thus result in biased estimation of β_x .

When sensitivity is smaller than PPV, this simulation study presents attenuated coefficient estimates when a single source is used and mixed direction of bias when exposures are averaged. However, we should recall the findings from Chapter III that the direction and size of bias are dependent on dispersion of true exposures when we have more undercount errors (sensitivity $<$ PPV). The relationship between the bias and the distribution of latent true exposure is more complicated in longitudinal design, but findings from Chapter III implies that both attenuation and inflation can occur with time-varying measurement errors as well. We should also note that naïve methods show negative bias with larger magnitude and larger RMSE when sensitivity is greater than PPV. This is consistent to the simulation results from Chapter III that bias is always negative when sensitivity $>$ PPV. When this occurs, there are relatively more incorrectly included businesses than incorrectly omitted business, which creates a noisier exposure measure than the scenario when sensitivity is lower than PPV. Formally, recall that the measurement error variance is $\left(\frac{sen_t^s}{ppv_t^s} - 1 + 2(1 - sen_t^s)\right) \lambda_{it}$. This variance increases when $sen_t^s > ppv_t^s$ thus producing more biased estimates because errors with higher variability are added to the true exposures compared to the scenario with $sen_t^s < ppv_t^s$. Despite more erroneous observed exposure, our model greatly reduces bias and RMSE compared to naïve models. Our model provides approximately correct coverage for β_x when sensitivity is smaller than PPV and the effect size is small (0.05). In this setting, performance of our model is robust to the mild prior misspecification. However, the model is more sensitive to the misspecification in the opposite case (sensitivity $>$ PPV). As the $sen_t^s < ppv_t^s$ with small effect size is the most plausible scenarios in the built-environment studies, we can expect

that our proposed model will well perform with the real data application.

4.5 Childhood BMI study in California from 2001 to 2008

In this section, our proposed model is applied to find the association between the changing availability of convenience stores near schools and the changes in body weight of 5th and 7th-grade children attending urban public schools in California from 2001 to 2008 (also see Section 4.2) using two available commercial databases. Children’s weight and height are collected by repeated cross-sectional Fitnessgram test (California Department of Education (2019)). Body mass index z-scores (BMIz) were then calculated by standardizing BMI according to age and sex-specific reference distributions because children have varying development of body compositions by age and sex (Must and Anderson (2006)). For the same reason, we also considered a stratified analysis by grade and sex. Our outcome variable is school-level average body mass index z-scores (BMIz) among children in urban public schools for $T = 8$ years. There were 1,369 schools and 1,341 schools that provided 8 years of BMI z-scores of male and female 5th-grade students, respectively. For 7th graders, 397 schools and 394 schools reported 8 years of BMI z-scores of male and female students, respectively. Other covariates such as publicly available school-level characteristics as covariates (e.g., income, ethnicity) were adjusted in the regression model.

To allow the nonlinear effect of one store increase in the school neighborhood, we used $\log_2(1 + x_{it})$ as an exposure predictor that copes with zero values in a count variable. Then the regression coefficient is interpreted as the difference in the outcome associated with a doubling in exposure. The model for each stratum is specified as follows.

$$Y_{it} = \beta_0 + \beta_x \log_2(x_{it} + 1) + \beta_t \text{Time}_{it} + \beta_{sp}(\text{Time}_{it} - 5)_+ + \mathbf{z}'_i \boldsymbol{\beta}_z + \mathbf{v}'_{it} \boldsymbol{\beta}_v + \epsilon_{it}, \quad (4.7)$$

where $Corr(\epsilon_{it}, \epsilon_{it+s}) = \rho^s$ for all t and s . To weight different number of students each year when averaging BMI-z scores on the school level, we assume $Var(\epsilon_{it}) = \frac{\sigma^2}{n_{it}}$ where n_{it} is the number of children in the i th school at time t at a given grade and sex. *Time* variable takes a numeric value between 1 (year 2001) to 8 (year 2008) and we added a linear spline knot at $Time_{it} = 5$ as there was a new school nutrition policy introduced in year 2005 (Ortiz, 2003). The bill prohibited the sale of sugar-added beverages and certain snacks in elementary schools.

We also adjusted for time-invariant school-level demographic characteristics $\mathbf{z}_i$ such as baseline community household income, and baseline community education level. Time-varying covariates $\mathbf{v}_{it}$ include schools' percentage ethnicity and school's enrollment size at time t .

Considering the similarity of urbanization level of the study region and the year that data were collected, Powell et al.'s validity statistics for Chicago metropolitan area in 2009 are used as reference for the California establishment data in 2008. According to the authors, the sensitivity of D&B and InfoUSA for convenience stores were 0.3769 (SE: 0.0263) and 0.5047 (SE: 0.0271), respectively, with PPVs of 0.4812 (SE: 0.0313, D&B) and PPV 0.6182 (SE: 0.0286, InfoUSA). Therefore, the prior distribution for $(sen_t^{NETS}, ppv_t^{NETS})$ and $(sen_t^{InfoUSA}, ppv_t^{InfoUSA})$ for $t < 8$ were constructed using these external information. For example, hyperparameter calculation by matching moments according to Section 4.3.3 gives us $(\gamma_1^{InfoUSA}, \gamma_2^{InfoUSA}) = (171.3, 168.1)$ from sensitivity and $(\gamma_1^{InfoUSA}, \gamma_3^{InfoUSA}) = (177.8, 109.8)$ from PPV. The value of $\gamma_1^{InfoUSA}$ computed with sensitivity and PPV are slightly different due to potential measurement errors in data, but we simply take an average of these two values and finally gain $(\gamma_1^{InfoUSA}, \gamma_2^{InfoUSA}, \gamma_3^{InfoUSA}) = (\frac{171.3+177.8}{2}, 168.1, 109.8)$. Then, the prior distribution $\pi(\theta_{11t}^{InfoUSA}, \theta_{12t}^{InfoUSA}, \theta_{21t}^{InfoUSA}) = Dirichlet(\gamma_1^{InfoUSA}, \gamma_2^{InfoUSA}, \gamma_3^{InfoUSA})$. Similarly, $(\gamma_1^{NETS}, \gamma_2^{NETS}, \gamma_3^{NETS}) = (124.9, 210.9, 131.7)$. We investigated the benefit of estimating source qualities instead of using fixed values for annual source qualities.

By comparing graphical posterior checks of replicated W 's (Appendix C.3), the model that estimates each year's source quality has a better model fit than the model that simply plugs in the reference values for all years. Therefore, this comparison emphasizes that the values of historical sensitivity and PPV are varying over time and we should estimate them.

Figure 4.3 shows the corrected results from our model and the naïve estimates for the effect of doubling convenience stores on the school's mean BMIz. The results shows that the naïve use of data give smaller magnitude of effect size compared to our approach. Stronger positive associations between exposure to convenience stores and school-level obesity risk are found by our proposed model and the model with the averaged predictor. The length of intervals is shortest in the proposed model and shorter for 5th grader data due to the larger sample size. The effect of adjusted covariates implied that schools with a higher percentage of Hispanic and Black students have significantly higher mean BMIz. Also, all strata showed a significant positive time effect with a negative spline effect, indicating the effect of policy intervention on school-level obesity. Posterior summaries of other covariates are available in Appendix C.1.

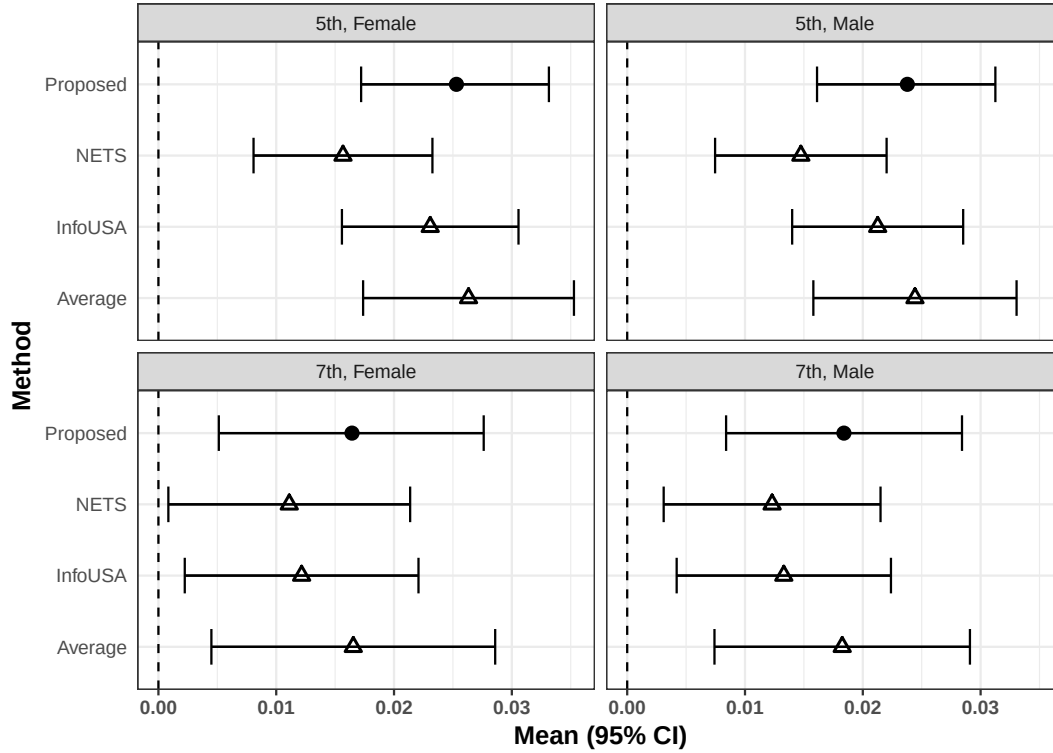


Figure 4.3: Comparison of estimated change in mean standardized BMI per doubling convenience stores around 1-mile buffer from a school between 2001 and 2008 after fitting the proposed model, naïve model with single-sourced exposure (NETS, InfoUSA), and naïve model with averaged exposure (Average). Stratified analyses are conducted by grade (5th, 7th) and sex. Plotted are posterior means (filled circle dots) with 95% credible intervals for the proposed model, and generalized least square estimates (open triangle dots) with 95% confidence intervals for naïve models.

In addition to stratified analyses, we also conducted an unstratified analysis using all subjects together and adjusting for stratum-factors in the regression model. When using all subjects combined, we found a differential effect of doubling exposure on school-level BMIz by grade. Increased exposure to conveniences stores had stronger influence for fifth graders and this is also shown in Figure 4.3 from stratified analyses. This implies that nearby convenience stores around schools are more detrimental to younger students.

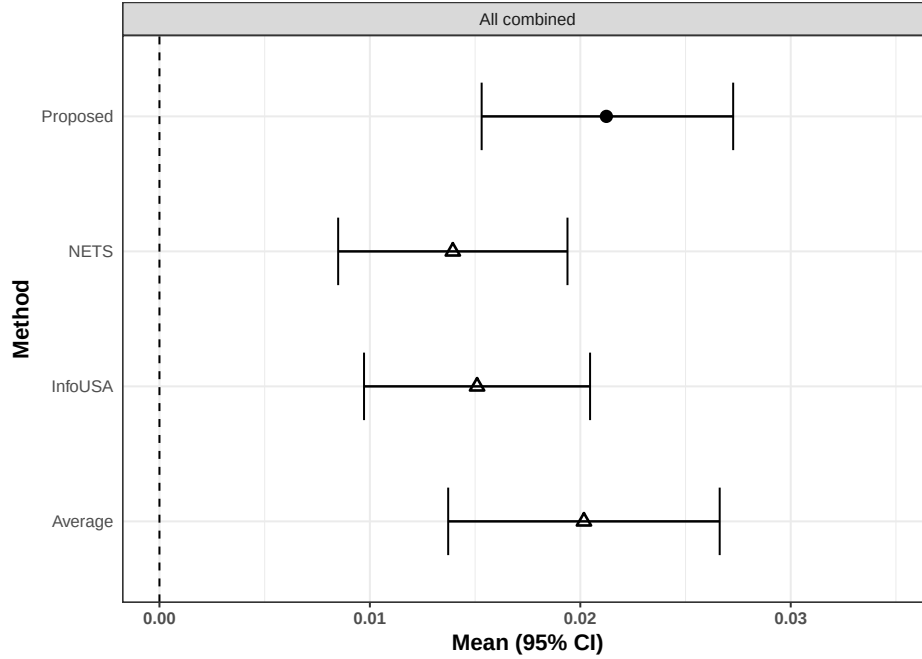


Figure 4.4: Estimated difference in mean standardized BMI per doubling convenience stores around schools between 2001 and 2008, using unstratified analysis. See caption of Figure 4.3 for the description of methods.

4.6 Discussion

By accounting for business dynamics in latent time-varying exposures, the proposed method showed that we can significantly reduce the bias of health effect estimates. Our method takes advantage of additional information about source quality that is reported from field studies and can limit costly data cleaning or data matching. Another attractive feature of the presented model is its parsimonious structure about sources of dependencies without requiring site- or time-dependent covariates. While this may sound appealing, it is important to carefully consider the assumption of the data generation mechanism and the availability of external source agreement statistics before applications of the model. Fortunately, the presented Bayesian framework greatly facilitates estimations of target parameters and reduces considerable bias in the naïve inference that can be misleading due to measurement errors.

One straightforward extension is a different outcome model $f(\mathbf{y}_i|\mathbf{x}_i)$ although derivation and estimation of posterior may not be simple. Another practical consideration centers around the choice of the prior distribution for sensitivity and PPV. Considering the amount of field studies, having a good reference for one or more years will not be extremely difficult. However, in practice, one might carry out the analysis for several different priors for sensitivity analyses.

Further investigation about measurement errors due to imputed data, specifically for the longitudinal databases such as NETS, can be another future direction. Some information in NETS is known to be imputed based on the past year’s database for the purpose of longitudinal use (Barnatchez et al., 2017). However, there is limited information about the mechanism of missing information and imputation of data in NETS, leading to ignoring the ‘smoothness’ of observed exposures in the scope of our paper. If we want to extend our model to introduce an extra layer of smoothness on observed exposures, we can use multivariate Poisson distributions or copulas (Joe, 2014) to model temporally dependent false positives.

We have suggested and shown the benefits of our model relative to the naïve use of commercial business databases in retrospective longitudinal environmental epidemiologic studies. While the issue of the incompleteness of data lists is often raised as a concern, there has been only sporadic attention in the literature to develop methods to reduce the bias without tedious and resource-consuming cleaning of data. Our model provides one promising approach to using count data in imperfect listings and provide valid inferences for regression coefficients.

CHAPTER V

Discussion and Future Work

With growing availability of secondary commercial databases, comprehensive information about community resources can be easily obtained for a large-scale geographic area and for a long follow-up time period. Based on this digitized built-environment information, a variety of observational health research could be conducted. Food-environment health research is one of the field that actively use the secondary commercial databases. Given documented inaccuracies in these databases, this dissertation has explored measurement errors that arise in using commercial databases and their effect on regression coefficient estimates. Each chapter proposed measurement error model to correct the biased inference, particularly for built food environment studies that use the secondary data sources. While the applications of each model were illustrated with food environment and obesity outcomes, the proposed models can be used to any built-environment research that use secondary commercial databases.

This dissertation focused on measurement errors from geocoding errors and ascertainment errors in the commercial databases. In Chapter II, we extended classical SIMEX by Cook and Stefanski to address complex measurement error due to geocoding coarsening. Key features of the method are that it does not assume a parametric form of measurement error and that it can accommodate a stratified error process, in

our case differences in the effects of geocode coarsening depending on whether or not a ZIP code centroid is near a school location. The SC-SIMEX is not just restricted to handle measurement errors from coarsened geocodes, but this method can be used to any count measurement errors whose distribution can be decomposed into several components. Chapter III and IV focused on the other critical source of error: faulty records in the databases. In Chapter III, joint modeling of outcomes, observed exposures, and latent exposures is proposed for a single time point. Chapter IV extends Chapter III for the longitudinal problem where exposures are varying over time. Both Chapter III and Chapter IV use external information about source credibility which can be found by existing validation studies. Having prior knowledge about source credibility led us to a Bayesian framework. We have shown that we can effectively use the prior source agreement statistics to reconcile the disagreeing multiple sources. Considering the availability of business listings provided by private sectors and the resource constraints of most epidemiologic studies, our proposed joint model can be easily applied to tackle erroneous exposures due to ascertainment errors.

This dissertation has practical and scientific contributions to future investigators who confront integer-valued measurement errors in count exposures. Even in the recent epidemiologic research with error-prone exposures (Shaw et al., 2018), only vague discussions are given for the exposure measurement errors because the origin and impact of the errors are not known. We hope this dissertation may serve as a guidance on the source and structure of measurement error, size and direction of bias on estimated associations, and methods to handle these issues. While only a mismeasured regressor with classical measurement error in a simple regression model leads to bias of regression coefficients toward the null, there is a perhaps oversimplified misconception that nearly all measurement error leads to an attenuated coefficient estimate (van Smeden et al., 2020). This thesis reminds that measurement error due to misclassification of count data is not necessarily symmetric and random.

All three chapters have shown that the direction and the magnitude of bias cannot be simplified by a few simple factors. Each of methods proposed in this dissertation also has contribution in wider class of measurement error literature by providing new modeling frameworks to non-classical count measurement errors. Historical measurement error models have been heavily focused on continuous measurement errors with distributional assumptions on measurement error. Errors in latent count variable is complicated due to its nature of discreteness when modeling mean and variance. Presented methods in this dissertation do not require specific distributional assumption on measurement error term and allows flexible modeling with count data.

In the future research, we hope our methods to be extended to other regression problems with different link functions in the outcome models. Although we demonstrated our methods with motivating examples that use continuous BMI scores as outcome data, all three models in this dissertation can be extended to different outcome types. Also, modeling additional dependencies is deserving of future investigation. The dependencies can include spatial dependencies of geocode accuracies, or time-dependence in the change of source quality over years. This extension will require extra data or strong modeling assumptions.

APPENDICES

APPENDIX A

Chapter II Supplementary Materials

A.1 Additional measurement error and bias analysis

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

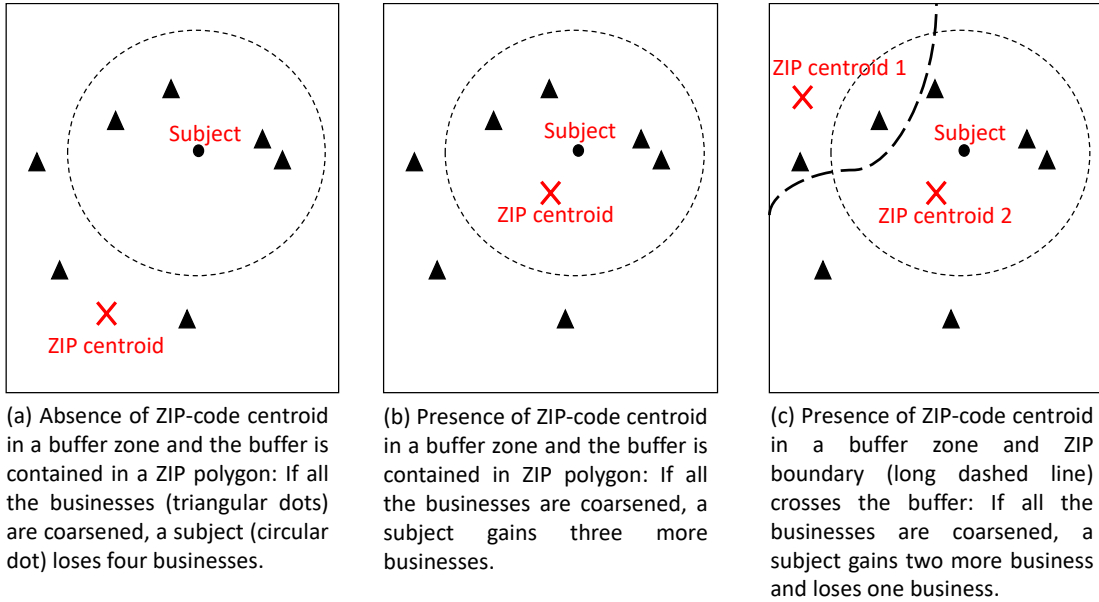


Figure A.2: Percent bias in health effect estimates from naïve multiple linear regression by coarsening level and food outlet type. Error-prone exposures to six outlet types are regressed in the same regression model. True effect size in the simulation is 3.2 for all outlet types.

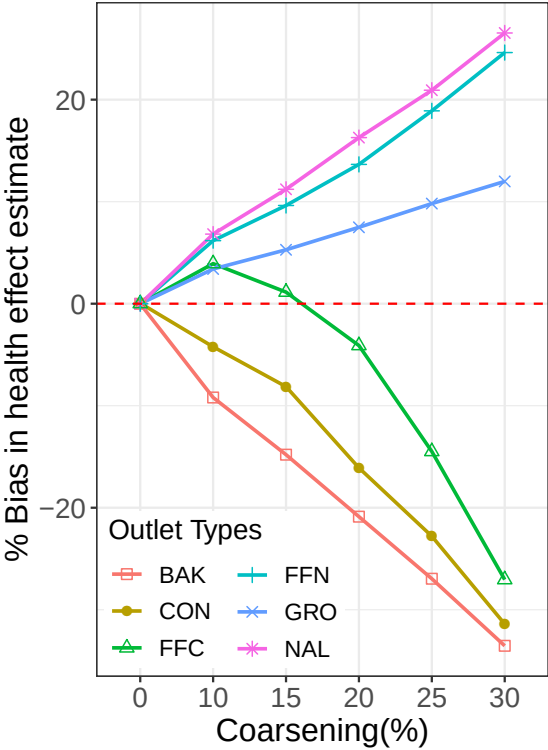
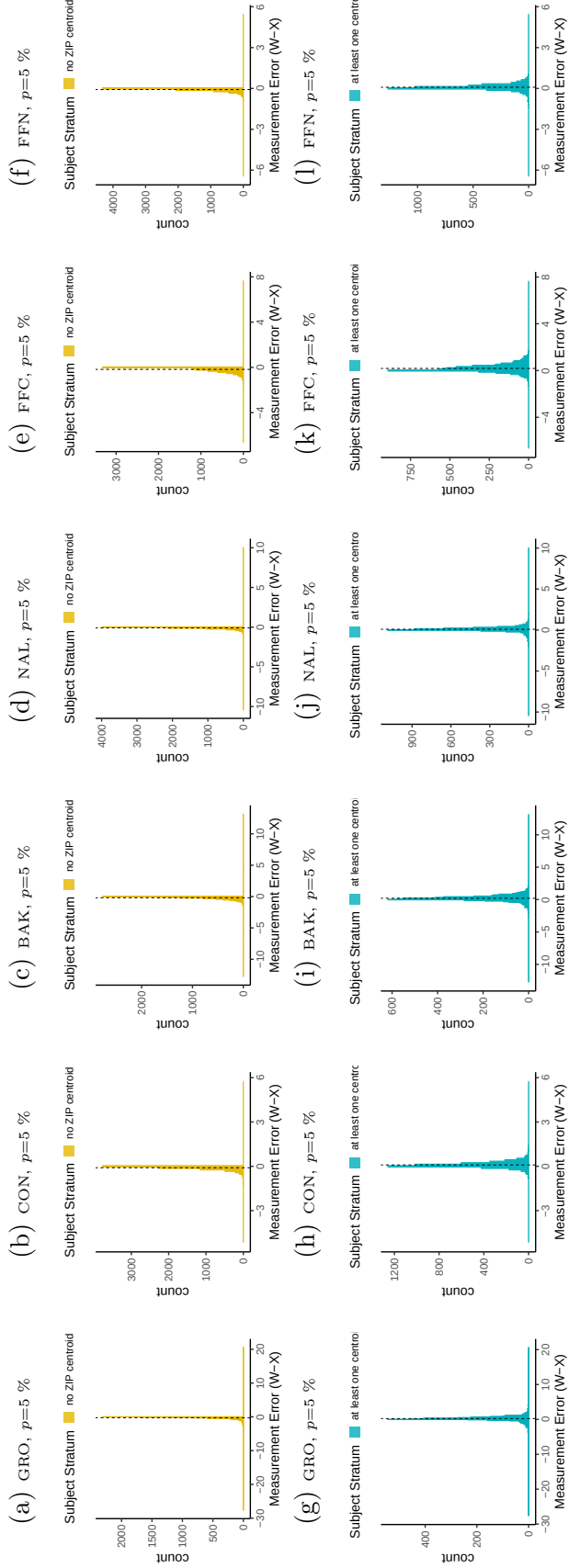


Figure A.3: Strata-specific histogram of measurement error mean, among 12,370 schools when coarsening probability is $p = 5\%$ and $p = 25\%$ (subfigures (a)-(x)). Measurement error means are obtained by averaging across 50 simulated values of U_i for each school. Dotted vertical line represents the 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.



(Figure A.3 continued) Strata-specific histogram of measurement error mean, among 12,370 schools when coarsening probability is $p = 5\%$ and $p = 25\%$ (subfigures (a)-(x)). 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.

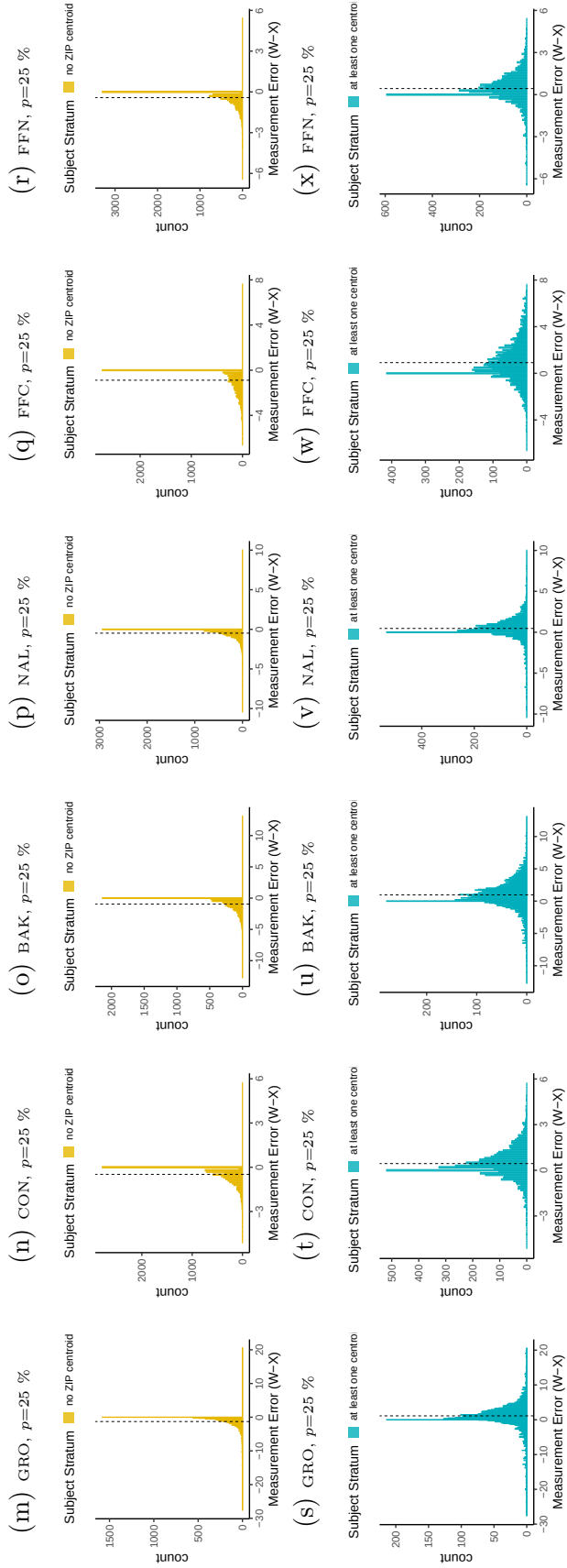
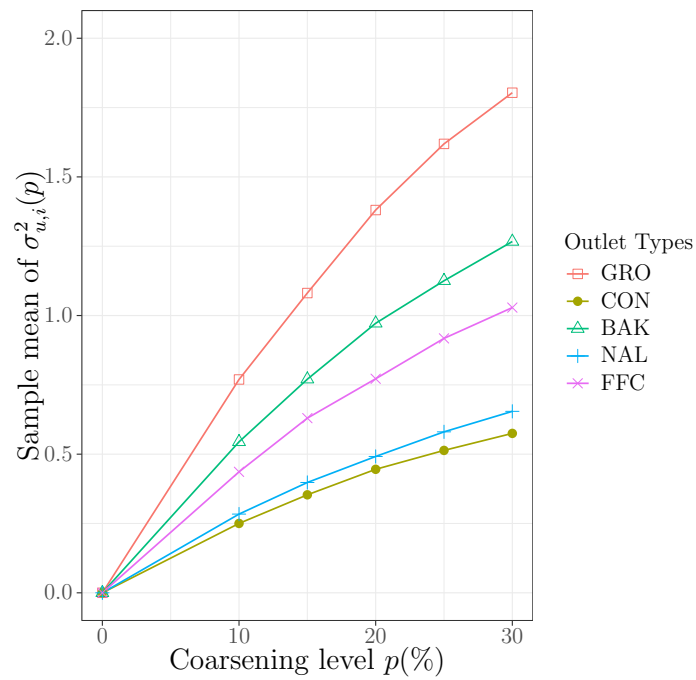


Figure A.4: Sample mean of individual level measurement error variance by coarsening level, for a single randomly selected school. Error variance on each subject was estimated from 50 replications.



A.2 Additional data analysis results

Table A.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.

Urbanization Level	Number of ZIP centroids within 1 mile buffer of school		
	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

A.3 Additional simulation results

Table A.2: Extension of Table 2.2, 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 +centroid	Quad	-2.0%	1.4%	4.2%	-1.6%	-13.3%	-0.5%
	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 +centroid	Quad	4.6	12.5	11.8	1.2	22.5	4.6
	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 +centroid	Quad	98%	99%	96%	94%	25%	99%
	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%

Table A.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 2.2, 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%

Table A.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 2.2, 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%

Table A.5: Sample standard deviation of the estimates (SE) and the average of bootstrapped standard error estimates (ESE) of SIMEX methods in Table 2.2. 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.

<i>SE</i>	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
<i>ESE</i>	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.4 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: $U_i^{10} \sim \text{Bin}(|\mathcal{S}_i^{10}|, p)$ and similarly $U_i^{01} \sim \text{Bin}(|\mathcal{S}_i^{01}|, p)$. 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 $\sigma_{u_i}^2 = \text{Var}(U_i) = \text{Var}(U_i^{01} - U_i^{10}) = \text{Var}(U_i^{01}) + \text{Var}(U_i^{10}) - 2\text{Cov}(U_i^{01}, U_i^{10}) = p(1-p)\mathbf{C}_{ii}$ where $\mathbf{C}_{ii} = (|\mathcal{S}_i^{01}| + |\mathcal{S}_i^{10}|)$. Let $\boldsymbol{\mu}_u$ denote the vector of measurement error means for all subjects, and $\boldsymbol{\Sigma}_u$ denote the variance-covariance matrix of measurement errors, with diagonal entries

equal to $\sigma_{u_i}^2$.

The remaining entries of Σ_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 i th subject's buffer, but incorrectly leave the r th subject's buffer; and analogously for $\mathcal{S}_{ir}^{10,01}$ and $\mathcal{S}_{ir}^{10,10}$. 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., $U_{ik}^{m(1-m),l(1-l)} \sim \text{Binom}(|\mathcal{S}_{ij}^{m(1-m),l(1-l)}|, p)$. 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,

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

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

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, Σ_U will be a sparse matrix.

In summary, $\Sigma_U = p(1-p)\mathbf{C}$, with the entries of $\mathbf{C}$ given above. Given Σ_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 $\mathbf{C}$ is a constant. Clearly $p \rightarrow 0$ implies $\Sigma_U \rightarrow 0$. Figure A.4 shows the measurement error variance $\hat{\sigma}_{u,i}^2$ for each store type, for a single randomly selected school, whereas Figure 2.2c,d 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, Σ_{UX} , can be shown to be:

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

Hence, $\Sigma_{UX} = -p(1-p)\mathbf{E}$ where $\mathbf{E}$ is a square matrix with $\mathbf{E}_{ii} = |\mathcal{S}_i^{10}|$ in diagonal and $\mathbf{E}_{ir} = \{|\mathcal{S}_{ir}^{10,10}| + |\mathcal{S}_{ir}^{10,11}| - |\mathcal{S}_{ir}^{01,10}| - |\mathcal{S}_{ir}^{01,11}|\}$ 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 $\mathbf{U}$ be the vector of measurement

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

$$\begin{aligned}
(N-1)\sigma_u^2(p) &= E[(\mathbf{U} - \mathbf{1}_N \bar{u})^T (\mathbf{U} - \mathbf{1}_N \bar{u})] \\
&= E[\mathbf{U}^T \mathbf{A} \mathbf{U}], \text{ where } \mathbf{A} = \mathbf{I} - \frac{1}{N} \mathbf{1} \mathbf{1}^T \\
&= \text{tr}\{\mathbf{A} \boldsymbol{\Sigma}_U\} + \boldsymbol{\mu}_U^T \mathbf{A} \boldsymbol{\mu}_U \\
&= p(1-p)\text{tr}\{\mathbf{A} \mathbf{C}\} + p^2 \mathbf{d}^T \mathbf{A} \mathbf{d}
\end{aligned} \tag{A.2}$$

where $\mathbf{d}$ is a N-dimensional vector with i th entry $|\mathcal{S}_i^{01}| - |\mathcal{S}_i^{10}|$

$$= p(1-p)c_1 + p^2 c_2.$$

Similarly,

$$\begin{aligned}
(N-1)\sigma_{ux}(p) &= E[(\mathbf{X} - \mathbf{1}_N \bar{x})^T (\mathbf{U} - \mathbf{1}_N \bar{u})] = E[\mathbf{X}^T \mathbf{A} \mathbf{U}] \\
&= \text{tr}\{\mathbf{A} \boldsymbol{\Sigma}_{UX}\} + \boldsymbol{\mu}_X^T \mathbf{A} \boldsymbol{\mu}_U \\
&= -p(1-p)\text{tr}\{\mathbf{A} \mathbf{E}\} + p \boldsymbol{\mu}_X^T \mathbf{A} \mathbf{d}
\end{aligned} \tag{A.3}$$

where $\boldsymbol{\mu}_X$ is a N-dimensional vector with i th entry $|\mathcal{S}_i^{11}| + |\mathcal{S}_i^{10}|$

$$= -p(1-p)c_3 + pc_4.$$

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

A.4.1 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

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

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,

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

with the quantities $\sigma_u^2(p)$ and $\sigma_{ux}(p)$ as defined above. The total variance of X is similarly written as $(N-1)\sigma_x^2 = E[(\mathbf{x} - \mathbf{1}_N\bar{x})^T(\mathbf{x} - \mathbf{1}_N\bar{x})]$. Given the quadratic nature of the dependence of $\sigma_u^2(p)$ and $\sigma_{ux}(p)$ on p , it is obvious that $\sigma_u^2(p) \rightarrow 0$ and $\sigma_{ux}(p) \rightarrow 0$ as $p \rightarrow 0$, and that $\text{plim}\hat{\beta}_w(p)$ 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 (A.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 $c_1, \dots, c_4$, 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 2.1(b) 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 (A.5) has a more predictable pattern as a function of p , as shown empirically in Figure 2.3.

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,

$$\begin{aligned} C_{ii} &= |\mathcal{S}_i^{10}|, \quad C_{ir} = |\mathcal{S}_{ir}^{10,10}|, \\ d_i &= -|\mathcal{S}_i^{10}|, \quad \mu_{xi} = |\mathcal{S}_i^{10}|, \\ E_{ii} &= |\mathcal{S}_i^{10}|, \quad E_{ir} = |\mathcal{S}_{ir}^{10,10}|. \end{aligned}$$

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

$$\begin{aligned}(N-1)(\sigma_u^2(p) + \sigma_{ux}(p)) &= (c_1 - c_3)p(1-p) + c_2p^2 + c_4p \\ &= p(p-1)c_2 < 0 \text{ for } 0 < p < 1.\end{aligned}$$

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 $0 \leq p < 1$.

APPENDIX B

Chapter III Supplementary Materials

B.1 Detailed proof for Section 3.3

$$sen_s = \frac{E(N_{11s,i})}{E(X_i)} = \frac{\pi_{11s}}{\pi_{11s} + \pi_{12s}} \quad (\text{B.1})$$

$$PPV_s = \frac{E(N_{11s,i})}{E(W_{si})} = \frac{\pi_{11s}}{\pi_{11s} + \pi_{21s}} \quad (\text{B.2})$$

Equation (B.1) and (B.2) result in $\pi_{12s} = \frac{1-sen_s}{sen_s} \pi_{11s}$ and $\pi_{21s} = \frac{1-ppv_s}{ppv_s} \pi_{11s}$, respectively. Since $\pi_{11s} + \pi_{12s} + \pi_{21s} = 1$, $\pi_{11s} = \frac{1}{\frac{1-sen_s}{sen_s} + \frac{1-ppv_s}{ppv_s} + 1}$.

B.2 Detailed estimation steps for Section 3.4

B.2.1 Prior and Initial Values

We use a noninformative prior for regression parameters $\pi(\beta_0, \beta_x, \sigma) \propto \sigma^{-2}$ and set initial values $(\hat{\beta}_0^{(0)}, \hat{\beta}_x^{(0)}, \hat{\sigma}^{(0)})$ to the estimates from ordinary least square for the regression using the average of $\mathbf{w}_1$ and $\mathbf{w}_2$ as a predictor. We initialize the true exposure values with $\hat{x}_i^{(0)} = \max(\lceil ppv_1 w_{1i} \rceil, \lceil ppv_2 w_{2i} \rceil)$ for all i , where $\lceil a \rceil$ defines the smallest integer that is not smaller than a . Initial cluster means are set to $\lambda^{(0)} = \overline{W}$, where $\overline{W}$ denotes an total mean

of observed exposures from all sources. Hyperparameters (a_0, b_0) for the base measure G are set to $(1, 1)$. The initial clustering of marginal model begins with a single cluster so that $\mathbf{z} = (1, 1, \dots, 1)$.

B.2.2 Details of Dirichlet Process Mixture model

In each MCMC step, two key steps of CRP are

1. Update z_i by multinomial sampling with

$$Pr(z_i = k | \mathbf{z}_{-i}, x_i, \boldsymbol{\lambda}) \propto \begin{cases} \frac{n_{k,-i}}{n+\alpha-1} Pois(x_i | \lambda_k) & \text{if } k \leq K, \\ \frac{\alpha}{n+\alpha-1} \int Pois(x_i | \lambda) G_0(d\lambda) & k = K + 1 \end{cases}$$

where $n_{k,-i} = \sum_{j \neq i} I(z_j = k)$ and K is the number of distinct z_{-i} 's. The details of computed probabilities are as follow.

- (a) The probability that a x_i falls into one of the *active* cluster k with parameter λ_k is derived as follows.

$$\begin{aligned} p(z_i = k | \mathbf{X}, \alpha, \mathbf{z}_{-i}, \{\lambda_k\}_{k=1}^\infty) \\ &= p(z_i = k | x_i, \alpha, \mathbf{z}_{-i}, \lambda_k) \\ &= p(z_i = k | \alpha, \mathbf{z}_{-i}, \lambda_k) p(x_i | z_i = k, \alpha, \mathbf{z}_{-i}, \lambda_k) \\ &= p(z_i = k | \alpha, \mathbf{z}_{-i}) p(x_i | \lambda_k) \\ &= \frac{n_{k,-i}}{n + \alpha - 1} Pois(x_i | \lambda_k) \\ &\propto n_{k,-i} Pois(x_i | \lambda_k). \end{aligned}$$

n_k denotes the size of cluster which is the number of observations assigned to the k th cluster. Note that the cluster disappears on the next cycle if i th observation is the only element of the cluster and the number of cluster varies across the samples.

- (b) The probability that z_i gets a *new* value of cluster index $K + 1$ where already

K clusters are active,

$$\begin{aligned}
& p(z_i = K + 1 | \mathbf{x}, \alpha, \mathbf{z}_{-i}, a_0, b_0) \\
&= p(z_i = K + 1 | x_i, \alpha, \mathbf{z}_{-i}, a_0, b_0) \\
&= p(z_i = K + 1 | \alpha, \mathbf{z}_{-i}, a_0, b_0) p(x_i | z_i = K + 1, \alpha, \mathbf{z}_{-i}, a_0, b_0) \\
&= p(z_i = K + 1 | \alpha, \mathbf{z}_{-i}) p(x_i | a_0, b_0) \\
&= \frac{\alpha}{n + \alpha - 1} \int \text{Pois}(x_i | \lambda) G_0(\lambda | a_0, b_0) d\lambda \\
&\propto \alpha \int_0^\infty \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \frac{b_0^{a_0} \lambda^{a_0-1} e^{-b_0 \lambda}}{\Gamma(a_0)} d\lambda \\
&\propto \alpha \frac{b_0^{a_0}}{(b_0 + 1)^{a_0+x_i}} \frac{1}{x_i!} \frac{\Gamma(a_0 + x_i)}{\Gamma(a_0)}.
\end{aligned}$$

2. Update λ_k with

$$[\lambda_k | z_i = k] \propto \begin{cases} G_0(d\lambda) \prod_{i: z_i = k} \text{Pois}(x_i; \lambda_k) & \text{if } k \leq K, \\ \frac{\text{Pois}(x_k; \lambda_k) G_0(d\lambda)}{\int \text{Pois}(x_k; \lambda_k) G_0(d\lambda)} & k = K + 1. \end{cases}$$

The cluster mean parameter λ_k is sampled from $\text{Gamma}(a_0 + \sum_{i=1}^{n_k} x_{k,i}, b_0 + n_k)$. Here, z_i represents the index of the cluster chosen for the data x_i . In our case, since x_i is not observable, we replace x_i with $\hat{x}_i$ that is sampled from its posterior distribution.

B.2.3 Algorithm outline

Gibbs sampling at t th iteration proceeds as follows:

1. Sample x given other random variables

For each school location, we sample $x_i^{(t+1)}$ from histogram sampling by computing equation (3.3) for possible support of x_i . The support of true exposure can be different by the urbanicity of the subject's location. We assume that observed exposures w_{1i}, w_{2i} reflect some level of urbanicity of i th location so that we build the support to be $[0 : \text{MAX}(\text{MAX}(w_{1i}, w_{2i}) * 3, 10)]$. This support avoids a degenerate case when

$w_{1i} = w_{2i} = 0$ where all databases failed to capture the real existing businesses. Here we assume the support $[0:10]$ covers the plausible value of falsely omitted x_i , when $w_{1i} = w_{2i} = 0$.

Once we have sampled values $\hat{\mathbf{x}}^{(t+1)} = (\hat{x}_1^{(t+1)}, \dots, \hat{x}_n^{(t+1)})$, we draw the parameter of interest. Then, we also update cluster assignment $\mathbf{z}$ and draw distinct cluster mean λ_k for the k th cluster from the DPM sampler.

2. Draw a sample for σ^2, β from

$$\begin{aligned} \hat{\sigma}^{2(t+1)} | \hat{\mathbf{x}}^{(t+1)}, \mathbf{y} &\sim \text{Inv}\chi^2(n-p, \frac{1}{n-p}(y - \hat{\mathbf{x}}^{(t+1)}\hat{\beta}^{(t)})'(y - \hat{\mathbf{x}}^{(t+1)}\hat{\beta}^{(t)})) \\ \hat{\beta}^{(t+1)} | \hat{\mathbf{x}}^{(t+1)}, \mathbf{y} &\sim N_2((\hat{\mathbf{x}}^{(t+1)T}\hat{\mathbf{x}}^{(t+1)})^{-1}\hat{\mathbf{x}}^{(t+1)T}\mathbf{y}, \hat{\sigma}^{2(t+1)}(\hat{\mathbf{x}}^{(t+1)T}\hat{\mathbf{x}}^{(t+1)})^{-1}). \end{aligned}$$

3. Draw a sample for the cluster index z_i for all i with

$$\begin{aligned} p(z_i = k, k \leq \hat{K}^{(t)}) &\propto \frac{n_{k,-i}^{(t)}}{n+\alpha-1} \text{Pois}(\hat{x}_i^{(t+1)} | \lambda_k^{(t)}) \text{ and} \\ p(z_i = \hat{K}^{(t)} + 1) &\propto \alpha \int_0^\infty \frac{\lambda^{\hat{x}_i} e^{-\lambda}}{\hat{x}_i!} \frac{b_0^{a_0} \lambda^{a_0-1} e^{-b_0\lambda}}{\Gamma(a_0)} d\lambda = \frac{b_0^{a_0}}{(b_0+1)^{a_0+x_i}} \frac{1}{x_i!} \frac{\Gamma(a_0+x_i)}{\Gamma(a_0)} \\ \text{if a new cluster is formed, } &\hat{K}^{(t+1)} = \hat{K}^{(t)} + 1. \end{aligned}$$

4. Draw a sample for the cluster mean λ_k from $k=1$ to maximum number of active

clusters $\hat{K}^{(t+1)}$ clusters is as below where $n_k^{(t+1)}$ denotes the size of the k th cluster,
 $\lambda_k | \hat{\mathbf{x}}^{(t+1)} \sim \text{Gamma}(a_0 + \sum_{i=1}^{n_k^{(t+1)}} \hat{x}_{k,i}^{(t+1)}, b_0 + n_k^{(t+1)})$.

5. Save $\hat{\beta}^{(t+1)}, \hat{\sigma}^{2(t+1)}, \hat{K}^{(t+1)}, \hat{\mathbf{z}}^{(t+1)}, \hat{\lambda}_k^{(t+1)}$ and go back to step 1.

Estimation and inferences are straightforward using conventional MCMC methods. We generated a chain length of 20,000 and discarded the first half as a burn-in. We implemented the model in R (version 4.0.4) and we computed the Brooks-Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992) using the output from the five chains of random starting values. The diagnostic statistic was less than 1.05 for all parameters of the model which indicates the convergence.

B.3 Detailed proof for Section 3.5

The mixture model that we consider for the predictor X is a K -component mixture Poisson distribution.

$$Z \sim Multinom(\omega_1, \dots, \omega_K) \quad (\text{B.3})$$

$$X|Z = k \sim Poisson(\lambda_k) \quad (\text{B.4})$$

Within k th cluster, X follows a simple Poisson distribution with λ_k so that $E(X|Z) = \sum_{k=1}^K I(Z = k)\lambda_k$.

For simple linear regression, the slope estimate $\hat{\beta}_{w_1}$ using OLS is

$$plim \hat{\beta}_{w_1} = \frac{Cov(W_1, Y)}{Var(W_1)}. \quad (\text{B.5})$$

We want to write (B.5) in terms of true coefficient β_x to explore the direction and magnitude of bias when we use W_1 in place of X in the outcome model. Similar argument is possible for W_2 .

- $Cov(W_1, Y)$

$$Cov(W_1, Y) = E\{Cov(W_1, Y|X)\} + Cov\{E(W_1|X), E(W_1|X)\} \quad (\text{B.6})$$

by law of total covariance

$$= 0 + Cov\{E(W_1|X), \beta_0 + \beta_x X\} \text{ by surrogacy: } W_1 \perp Y|X \quad (\text{B.7})$$

Using Tower property,

$$E(W_1|X) = E\{E(W_1|X, Z)|X\} \quad (\text{B.8})$$

$$= E\{E(N_{111} + N_{211}|X, Z)|X\} \text{ since } W_1 = N_{111} + N_{211} \quad (\text{B.9})$$

$$= E\{E(N_{111}|X, Z) + E(N_{211}|X, Z)|X\} \quad (\text{B.10})$$

Here, note that $N_{111}|X = x \sim Binom(x, sen_1)$ regardless of cluster membership Z .

Therefore, $E(N_{111}|X, Z) = E(N_{111}|X)$.

For $E(N_{211}|X, Z)$, recall that we have three independent cells

$$N_{111}|Z = k \sim \text{Pois}(sen_1 \lambda_k),$$

$$N_{121}|Z = k \sim \text{Pois}((1 - sen_1) \lambda_k),$$

$$N_{211}|Z = k \sim \text{Pois}\left(\left(\frac{sen_1}{ppv_1} - sen_1\right) \lambda_k\right)$$

within k th cluster. Therefore, we have $N_{211} \perp X (= N_{111} + N_{121})$ given Z , and this leads to $E(N_{211}|X, Z) = E(N_{211}|Z)$.

Equation (B.10) becomes

$$(B.10) = E\{E(N_{111}|X) + E(N_{211}|Z)|X\} \tag{B.11}$$

$$= E\{sen_1 X|X\} + E\left\{\sum_{k=1}^K I(Z = k) \left(\frac{sen_1}{ppv_1} - sen_1\right) \lambda_k | X\right\} \tag{B.12}$$

$$= sen_1 X + \sum_{k=1}^K I(Z = k) \left(\frac{sen_1}{ppv_1} - sen_1\right) \lambda_k \tag{B.13}$$

Let p denotes $(\frac{sen_1}{ppv_1} - sen_1)$. Finally, equation (B.7) can be written as

$$(B.7) = Cov\{sen_1 X + \sum_{k=1}^K I(Z = k)p\lambda_k, \beta_0 + \beta_x X\} \quad (B.14)$$

$$= sen_1 \beta_x Var(X) + p \beta_x Cov(\sum_{k=1}^K I(Z = k)\lambda_k, X) \quad (B.15)$$

$$= sen_1 \beta_x Var(X) + p \beta_x Cov(E(X|Z), X) \text{ by definition of } E(X|Z) \quad (B.16)$$

$$= sen_1 \beta_x Var(X) + p \beta_x \{E\{Cov(E(X|Z), X|Z)\} + Cov\{E(E(X|Z)|Z), E(X|Z)\}\} \quad (B.17)$$

by law of total covariance

$$= sen_1 \beta_x Var(X) + p \beta_x \{E\{E(E(X|Z)X|Z) - E(E(X|Z)|Z)E(X|Z)\} + Cov\{E(X|Z), E(X|Z)\}\} \quad (B.18)$$

using definition of covariance

$$= sen_1 \beta_x Var(X) + p \beta_x \{E\{E(X|Z)E(X|Z) - E(X|Z)E(X|Z)\} + Var\{E(X|Z)\}\} \quad (B.19)$$

$$= sen_1 \beta_x Var(X) + p \beta_x \{E\{0\} + Var\{E(X|Z)\}\}$$

$$= sen_1 \beta_x Var(X) + p \beta_x \{Var(X) - E(X)\}$$

$$\text{since } Var(X) = E(Var(X|Z)) + Var(E(X|Z)) \quad (B.20)$$

$$= (sen_1 Var(X) + p\{Var(X) - E(X)\})\beta_x. \quad (B.21)$$

- $Var(W_1)$

$$Var(W_1) = E\{Var(W_1|Z)\} + Var\{E(W_1|Z)\} \quad (B.22)$$

$$= E\{Var(N_{111} + N_{211}|Z)\} + Var\{E(N_{111} + N_{211}|Z)\} \quad (B.23)$$

$$= E\{Var(N_{111}|Z) + Var(N_{211}|Z)\} + Var\{E(N_{111}|Z) + E(N_{211}|Z)\},$$

$$\text{by independence between cells} \quad (B.24)$$

$$= E\left\{\sum_{k=1}^K I(Z=k)sen_1\lambda_k + \sum_{k=1}^K I(Z=k)p\lambda_k\right\} + \\ Var\left\{\sum_{k=1}^K I(Z=k)sen_1\lambda_k + \sum_{k=1}^K I(Z=k)p\lambda_k\right\} \quad (B.25)$$

$$= (sen_1 + p)E(X) + Var\left(\sum_{k=1}^K I(Z=k)(sen_1 + p)\lambda_k\right) \quad (B.26)$$

$$= (sen_1 + p)E(X) + (sen_1 + p)^2 Var(E(X|Z)) \quad (B.27)$$

$$= (sen_1 + p)E(X) + (sen_1 + p)^2 \{Var(X) - E(X)\}. \quad (B.28)$$

Therefore, $plim \hat{\beta}_{w_1}$ in equation (B.5) is $\frac{(sen_1 Var(X) + p\{Var(X) - E(X)\})}{(sen_1 + p)E(X) + (sen_1 + p)^2\{Var(X) - E(X)\}} \beta_x$.

When the factor $\frac{(sen_1 Var(X) + p\{Var(X) - E(X)\})}{(sen_1 + p)E(X) + (sen_1 + p)^2\{Var(X) - E(X)\}}$ denoted as f , is greater than 1, then the naive estimate for β_x is inflated. On the other hand, when the factor is less than 1, we have attenuation.

- Simple Poisson X

When X has simple Poisson distribution with mean parameter λ , then $E(X) = Var(X) = \lambda$. The factor is written as

$$\frac{sen_1 Var(X)}{\frac{sen_1}{ppv_1} E(X)} \quad (B.29)$$

$$= ppv_1 < 1. \quad (B.30)$$

- Overdispersed X

When X is overdispersed, $Var(X) > E(X)$, so that $\{Var(X) - E(X)\} > 0$. In this

case, the factor f is written as

$$\frac{(sen_1 Var(X) + p\{Var(X) - E(X)\})}{(sen_1 + p)E(X) + (sen_1 + p)^2\{Var(X) - E(X)\}} \quad (B.31)$$

$$= \frac{(sen_1 Var(X) + (\frac{sen_1}{ppv_1} - sen_1)\{Var(X) - E(X)\})}{(\frac{sen_1}{ppv_1})E(X) + (\frac{sen_1}{ppv_1})^2\{Var(X) - E(X)\}} \quad (B.32)$$

$$= \frac{(sen_1\{Var(X) - E(X)\} + sen_1E(X) + (\frac{sen_1}{ppv_1} - sen_1)\{Var(X) - E(X)\})}{(\frac{sen_1}{ppv_1})E(X) + (\frac{sen_1}{ppv_1})^2\{Var(X) - E(X)\}} \quad (B.33)$$

$$= \frac{(sen_1E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\})}{(\frac{sen_1}{ppv_1})E(X) + (\frac{sen_1}{ppv_1})^2\{Var(X) - E(X)\}} \quad (B.34)$$

$$= \frac{(ppv_1E(X) + \{Var(X) - E(X)\})}{E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}} \quad (B.35)$$

$$= \frac{(ppv_1 - 1)E(X) + E(X) + \{Var(X) - E(X)\}}{E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}} + \frac{(\frac{sen_1}{ppv_1})\{Var(X) - E(X)\} - (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}}{E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}} \quad (B.36)$$

$$= \frac{(ppv_1 - 1)E(X) + (1 - \frac{sen_1}{ppv_1})\{Var(X) - E(X)\} + E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}}{E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}} = \frac{E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}}{E(X) + (\frac{sen_1}{ppv_1})\{Var(X) - E(X)\}} \quad (B.37)$$

Therefore, the attenuation occurs when

$$f < 1 \equiv (ppv_1 - 1)E(X) + (1 - \frac{sen_1}{ppv_1})\{Var(X) - E(X)\} < 0 \quad (B.38)$$

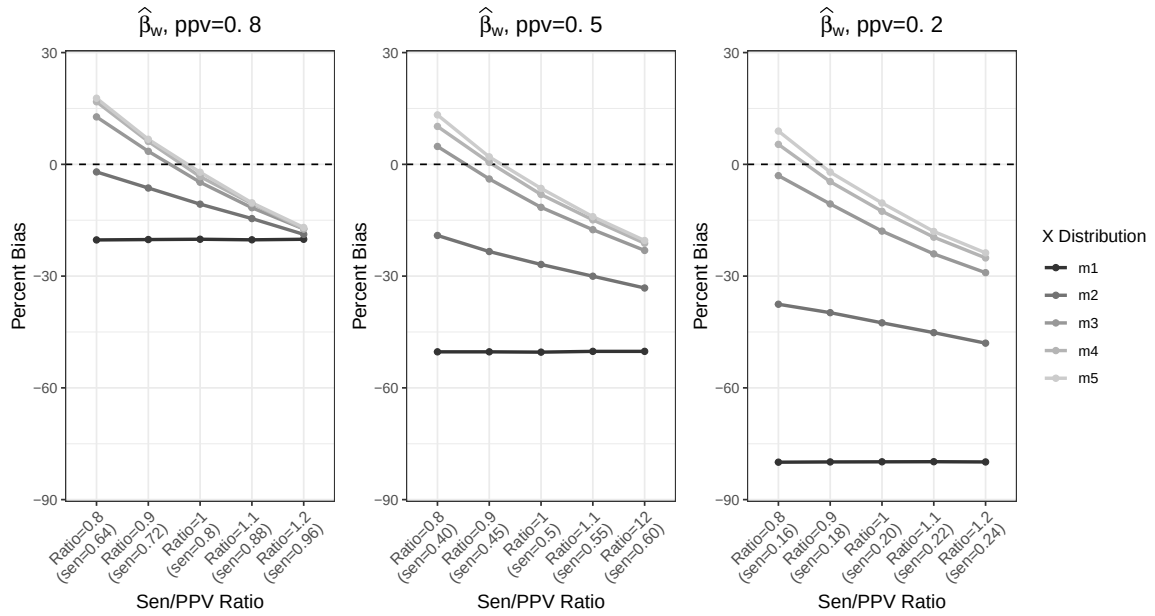
$$\frac{ppv_1 - sen_1}{ppv_1} Var(X) < (1 - \frac{sen_1}{ppv_1} + 1 - ppv_1)E(X) \quad (B.39)$$

$$\frac{ppv_1 - sen_1}{ppv_1} Var(X) < (\frac{ppv_1 - sen_1 + ppv_1(1 - ppv_1)}{ppv_1})E(X) \quad (B.40)$$

$$Var(X) < \{\frac{ppv_1(1 - ppv_1)}{ppv_1 - sen_1} + 1\}E(X). \quad (B.41)$$

B.4 Relationship between naïve bias and Sensitivity/PPV

Figure B.1: Percent bias of slope (β_x) estimates from single-source naïve regression models $OLS_s, s = 1$ by the ratio of sensitivity to PPV, and the magnitude of agreement statistics. Grayscale lines represent different X distributions that are defined in Section 3.5.2. For each X distribution, source is created with different values of sen_s , holding ppv_s at 0.8, 0.5, or 0.2. Note that when sensitivity is greater than PPV (i.e., $\frac{sen_s}{ppv_s} > 1$), naïve estimate for the slope is always attenuated as described in Section 3.5.4 and displayed in Figure 3.2(b), regardless of values of agreement statistics and X distribution. When X is equidispersed (Mixture 1), naïve bias only depends on the value of PPV (Section 3.5).



B.5 Model evaluation for β_0

Table B.1: ($sen1 < ppv1, sen2 < ppv2$) Posterior summaries of the proposed model for β_0 when true source reliability is $sen_1 = 0.37$, $sen_2 = 0.4$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. True $\beta_0=1.5$. The proposed model is evaluated with and without estimating X distribution. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications

Method	Mixture 1			Mixture 3			Mixture 5		
	Average Naive Est. (%Bias)	SD	Coverage Probability	Average Naive Est. (%Bias)	SD	Coverage Probability	Average Naive Est. (%Bias)	SD	Coverage Probability
OLS, source 1	3.30 (120%)	0.14	0%	1.98 (32%)	0.09	0%	1.81 (21%)	0.08	9%
OLS, source 2	2.91 (94%)	0.13	0%	1.87 (25%)	0.09	1%	1.74 (16%)	0.08	22%
OLS, average	2.33 (55%)	0.16	0%	1.65 (10%)	0.09	58%	1.60 (7%)	0.08	70%
Proposed Model									
when X distribution known									
Working credibility	Average Posterior Mean (%Bias)	SD	Coverage Probability	Average Posterior Mean (%Bias)	SD	Coverage Probability	Average Posterior Mean (%Bias)	SD	Coverage Probability
XP	0.35 (-76.57%)	0.24	0%	1.43 (-4.36%)	0.08	89%	1.40 (-6.98%)	0.10	82%
P	0.91 (-39.59%)	0.25	34%	1.46 (-2.78%)	0.08	92%	1.45 (-3.64%)	0.10	91%
R	1.47 (-1.77%)	0.22	94%	1.49 (-0.63%)	0.08	95%	1.50 (0.26%)	0.09	98%
O	1.83 (22.06%)	0.19	62%	1.53 (1.72%)	0.08	95%	1.55 (3.17%)	0.09	92%
XO	2.11 (40.45%)	0.18	7%	1.57 (4.36%)	0.09	87%	1.57 (4.70%)	0.08	77%
Proposed Model									
when X distribution unknown									
Working credibility	Average Posterior Mean (%Bias)	SD	Coverage Probability	Average Posterior Mean (%Bias)	SD	Coverage Probability	Average Posterior Mean (%Bias)	SD	Coverage Probability
XP	0.36 (-75.92%)	0.23	1%	0.62 (-58.46%)	0.12	0%	0.82 (-45.20%)	0.17	0%
P	0.93 (-38.30%)	0.25	35%	1.06 (-29.20%)	0.10	0%	1.21 (-19.23%)	0.12	31%
R	1.48 (-1.20%)	0.21	94%	1.32 (-12.27%)	0.09	45%	1.41 (-6.31%)	0.10	87%
O	1.83 (22.26%)	0.19	56%	1.45 (-3.11%)	0.09	92%	1.50 (0.16%)	0.09	95%
XO	2.05 (36.90%)	0.17	16%	1.54 (2.83%)	0.09	91%	1.55 (3.13%)	0.08	84%

B.6 Results of the model with single Gamma prior for λ_i

Table B.2: (Same data as Table 3.3) Posterior summaries of the proposed model for β_x when Gamma(1,1) prior is assumed for all $\lambda_i = \lambda$ instead of Dirichlet process prior. True $\beta_x=0.5$ and the true source credibility is $sen_1 = 0.37$, $sen_2 = 0.5$, $ppv_1 = 0.48$, and $ppv_2 = 0.62$. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications

Method	Mixture 1			Mixture 3			Mixture 5					
	Average			Average			Average					
	Naive Est.	SD	CP	Naive Est.	SD	CP	Naive Est.	SD	CP			
	(%Bias)			(%Bias)			(%Bias)					
OLS, source 1	0.240 (-52.05%)	0.028	0%	0.540 (7.96%)	0.016	8%	0.580 (16.06%)	0.012	16%			
OLS, source 2	0.313 (-37.40%)	0.026	0%	0.539 (7.86%)	0.014	8%	0.570 (14.04%)	0.011	14%			
OLS, average	0.449 (-10.15%)	0.033	72%	0.602 (20.31%)	0.015	20%	0.614 (22.77%)	0.011	23%			
Proposed Model												
when X distribution unknown and assumed simple Poisson												
	Average			Average	Average			Average	Average			Average
Working credibility	Posterior Mean	SD	CP	Posterior Mean	Posterior Mean	SD	CP	Posterior Mean	Posterior Mean	SD	CP	Posterior Mean
	$\hat{\beta}_x$ (%Bias)			$\hat{\lambda}$	$\hat{\beta}_x$ (%Bias)			$\hat{\lambda}$	$\hat{\beta}_x$ (%Bias)			$\hat{\lambda}$
XP	0.70 (39.81%)	0.02	0%	5.67	0.91 (81.65%)	0.01	0%	5.57	0.86 (72.76%)	0.01	0%	5.55
P	0.60 (20.04%)	0.02	0%	5.68	0.79 (57.01%)	0.01	0%	5.55	0.75 (49.12%)	0.01	0%	5.55
R	0.50 (0.49%)	0.02	95%	5.70	0.67 (33.29%)	0.01	0%	5.53	0.65 (30.82%)	0.01	0%	5.55
O	0.44 (-11.99%)	0.01	0%	5.73	0.59 (18.83%)	0.01	0%	5.54	0.60 (20.64%)	0.01	0%	5.59
XO	0.39 (-22.01%)	0.01	0%	5.81	0.55 (9.21%)	0.01	0%	5.63	0.57 (14.05%)	0.01	0%	5.69

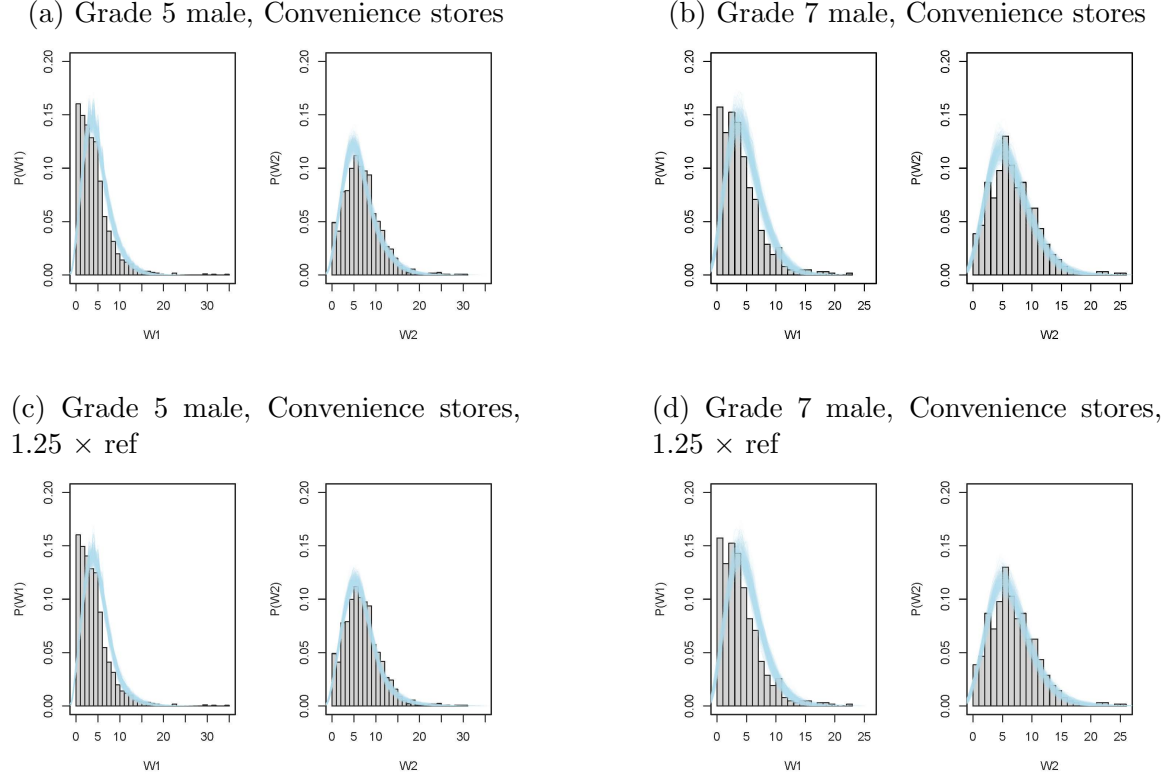
B.7 Results of the model with ($sen_1 > ppv_1, sen_2 < ppv_2$)

Table B.3: ($sen_1 > ppv_1, sen_2 < ppv_2$) Posterior summaries of the proposed model for β_x when true source credibility is $sen_1 = 0.48$, $sen_2 = 0.5$, $ppv_1 = 0.37$, and $ppv_2 = 0.62$. True $\beta_x = 0.5$. The proposed model is evaluated with unknown X distribution. The results shown are: average of point estimates, posterior standard deviation (SD), and coverage probability (CP) from 100 replications

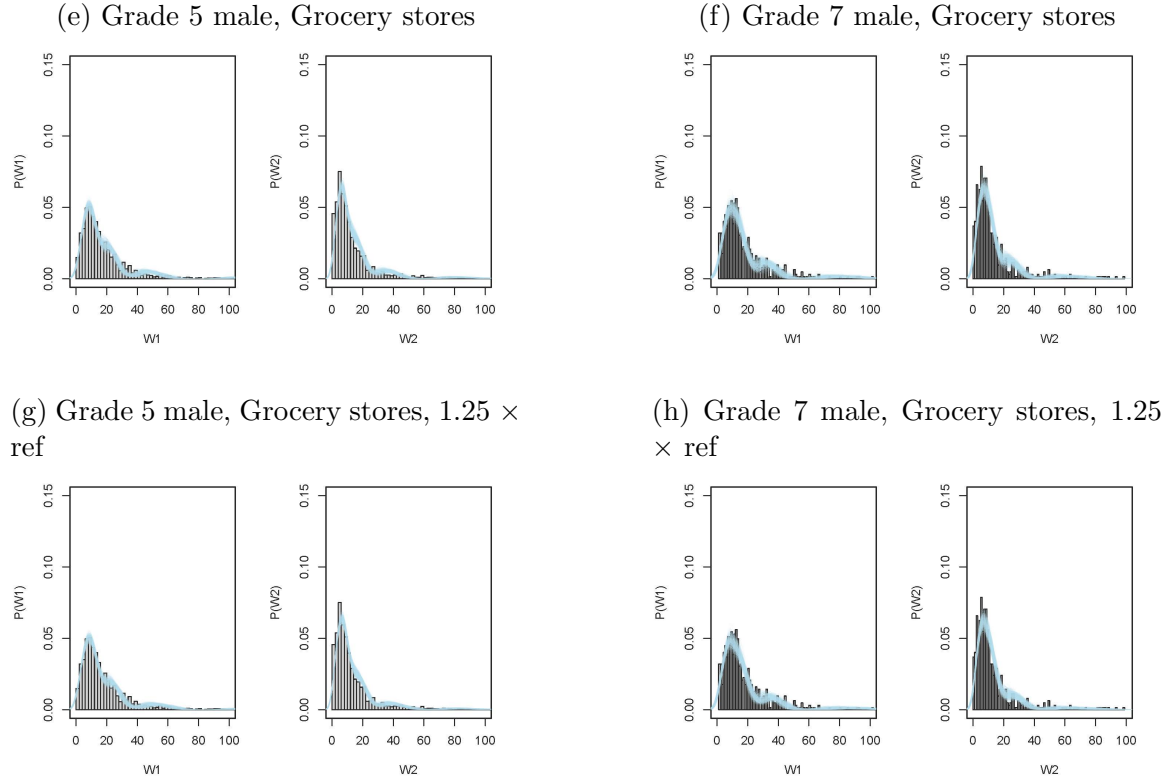
Method	Mixture 1			Mixture 3			Mixture 5		
	Average			Average			Average		
	Naive Est.	SD	CP	Naive Est.	SD	CP	Naive Est.	SD	CP
	(%Bias)			(%Bias)			(%Bias)		
OLS, source 1	0.186 (-62.87%)	0.022	0%	0.349 (-7.96%)	0.011	0%	0.365 (-27.10%)	0.013	0%
OLS, source 2	0.313 (-37.40%)	0.028	0%	0.539 (7.86%)	0.014	8%	0.571 (14.04%)	0.011	0%
OLS, average	0.382 (-23.61%)	0.031	2%	0.462 (20.31%)	0.012	5%	0.468 (-6.35%)	0.014	14%
Proposed model									
when X distribution unknown									
Working credibility	Average			Average			Average		
	Posterior	SD	CP	Posterior	SD	CP	Posterior	SD	CP
	Mean			Mean			Mean		
	(%Bias)			(%Bias)			(%Bias)		
XP	0.700 (40.08%)	0.037	0%	0.617 (23.45%)	0.014	0%	0.612 (22.50%)	0.022	0%
P	0.602 (20.46%)	0.044	34%	0.554 (10.84%)	0.013	3%	0.547 (9.48%)	0.020	28%
R	0.506 (1.27%)	0.041	88%	0.520 (3.99%)	0.011	64%	0.513 (2.64%)	0.016	86%
O	0.442 (-11.53%)	0.032	60%	0.502 (0.308%)	0.013	99%	0.499 (-0.20%)	0.013	91%
XO	0.394 (-21.19%)	0.033	6%	0.487 (-2.53%)	0.011	80%	0.489 (-2.19%)	0.014	81%

B.8 Graphical posterior predictive check of Section 3.6

Figure B.2: Distribution of observed exposures W_1 (NETS) and W_2 (InfoUSA), and density of 500 replicated W s from a fitted model (lightblue lines). Subfigures (a-d) show exposures to convenience stores for 5th and 7th-grade schools; (e-h) show exposures to grocery stores. Subfigures (a, b, e, f) are from our proposed model with the reference values and (c, d, g, h) are from our proposed model with reference values $\times 1.25$. See Section 3.6 and Figure 3.3 for reference values of agreement statistics.



(Figure B.2 continued) Distribution of observed exposures W_1 (NETS) and W_2 (InfoUSA), and density of 500 replicated W s from a fitted model (lightblue lines). Subfigures (a-d) show exposures to convenience stores for 5th and 7th-grade schools; (e-h) show exposures to grocery stores. Subfigures (a, b, e, f) are from our proposed model with the reference values and (c, d, g, h) are from our proposed model with reference values $\times 1.25$. See Section 3.6 and Figure 3.3 for reference values of agreement statistics.



APPENDIX C

Chapter IV Supplementary Materials

C.1 Proof of marginal distribution of X_{it}

In this proof, the subscript i as a subject indicator is suppressed. Let φ_S denote the characteristic function of a random variable S . If $S = p \circ X_1$, where X_1 is a nonnegative interger-valued random variable, then $\varphi_S(t) = \mathbb{E}(pe^{it} + 1 - p)^{X_1}$ (Bourguignon et al., 2016). If $X_1 \sim Pois(\lambda_1)$, then $\varphi_S(u) = \exp\{p\lambda_1(e^{iu} - 1)\}$. By the uniqueness of the characteristic function and the assumption $\epsilon_t \sim Pois(\mu)$ for $t > 1$, $\varphi_{X_2} = \varphi_{p \circ X_1} \cdot \varphi_{\epsilon_2} = \exp\{p\lambda_1(e^{iu} - 1)\} \exp\{\mu(e^{iu} - 1)\} = \exp\{(p\lambda_1 + \mu)(e^{iu} - 1)\}$. Thus, $X_2 \sim Pois(p\lambda_1 + \mu)$. With the similar arguments, $X_{it} \sim Pois(p^{t-1}\lambda_1 + \frac{1-p^{t-1}}{1-p}\mu)$ for $t > 1$.

C.2 The Markov chain Monte Carlo sampler

In Bayesian measurement error model, we treat latent true exposure as missing values and fill in $\{X_{it}\}$ given observed data $\{Y_{it}, W_{it}^A, W_{it}^B\}$ and current values of the parameters. Then, we sample other random variables by Gibbs sampler. The sketch of the estimation process is as follows.

Prior distributions:

$$\begin{aligned}
\pi(\boldsymbol{\beta}) &\propto 1, \pi(\log(\sigma)) \sim N(0, 1), \pi(\rho) \sim Unif(-1, 1) \\
\pi(p) &\sim Beta(1, 1) \\
\pi(\theta_{11t}^A, \theta_{12t}^A, \theta_{21t}^A) &\sim Dirichlet(\gamma_1^A, \gamma_2^A, \gamma_3^A) \\
\pi(\theta_{11t}^B, \theta_{12t}^B, \theta_{21t}^B) &\sim Dirichlet(\gamma_1^B, \gamma_2^B, \gamma_3^B) \\
\lambda_{i1} &\sim G, G \sim DP(\alpha, G_0), G_0 \sim Gamma(a, b) \tag{C.1} \\
\mu_i &\sim G', G' \sim DP(\alpha', G'_0), G'_0 \sim Gamma(a', b') \tag{C.2} \\
\alpha &\sim Gamma(a_0, b_0), \alpha' \sim Gamma(a'_0, b'_0) \tag{C.3}
\end{aligned}$$

Initial values:

Initial values for regression parameters are set to $\boldsymbol{\beta}^{(0)} \sim N(\mathbf{0}, I_2)$, $\sigma^{(0)} = 0.5$, $\rho^{(0)} = 0.7$. We also initialize the true exposure values with $\hat{x}_{it} = \max(\lceil ppv_1 W_{1i} \rceil, \lceil ppv_2 W_{2i} \rceil)$ for all i , where $\lceil a \rceil$ defines the smallest integer that is not smaller than a . Initial cluster means for baseline exposure are set to $\lambda_{i1}^{(0)} = \frac{W_{i1}^A + W_{i1}^B}{2}$, the total mean of observed exposures from two sources. Let Z_i be a group indicator for λ_{i1} and Z'_i be a group indicator for μ_i . Initial cluster means for arrival rate $\lambda_{i1}^{(0)}$ is set to 1. The base measures are $G = Gamma(1, 1)$ and $G' = Gamma(1, 5)$. We start with a single cluster for λ_{i1} and μ_i so that $z_i = z'_i = 1$ for all i . Hyperparameter for priors for concentration parameters, α and α' , are Prior of concentration parameters α and α' are specified as $Gamma(1, 1)$ and $Gamma(1, 10)$ as described in Section 4.3.2.

1. For $i = 1, \dots, n$ and $t = 1, \dots, T$, sample X_{it} given the current values of other variables. For the possible range of support, we compute the complete likelihood (4.3.2) and draw a value for X_{it} by histogram sampling. For computational purpose, instead of using the range $[0 : \infty)$, we define the support of X_{it} to be $[0 : \max(\max(W_{1it}, W_{2it}) * 2, 10)]$. Note that this support handles a degenerate case when $W_{1it} = W_{2it} = 0$. We made an assumption that $[0:10]$ covers the possible values of the true exposures when observed exposures have low counts < 5 .

2. Sample $\{\lambda_{i1}\}$ and α given current $\{X_{i1}\}$. The Chinese restaurant process (CRP) iteratively samples the subject-specific λ_{i1} as follows.

- (a) Draw a sample for the cluster index z_i for all i with

$$p(z_i = k, k \leq \widehat{K}) \propto n_{k,-i} \text{Pois}(X_{i1} | \lambda_k) \text{ and}$$

$$p(z_i = \widehat{K} + 1) \propto \alpha \frac{b^a}{(b+1)^{(a+X_{i1})}} \frac{1}{X_{i1}!} \frac{\Gamma(a+X_{i1})}{\Gamma(a)}.$$

If a new cluster is formed, $\widehat{K} = \widehat{K} + 1$.

- (b) Draw a sample for the cluster mean λ_k from $k=1$ to maximum number of active clusters $\widehat{K}$ clusters is as below where n_k denotes the size of the k th cluster,

$$\lambda_k | \{X_{i1}\}, z_i = k \sim \text{Gamma}(a + \sum_{i=1}^{n_k} X_{i1}, b + n_k).$$

- (c) The concentration parameter α is sampled according to Escobar and West.

3. For $t = 2, \dots, T$ and $i = 1, \dots, n$, sample innovation e_{it} from its conditional distribution

$$p(e_{it} | X_{it}, X_{it-1}, p, \mu_i, z'_i) \propto \frac{1}{e_{it}! (X_{it} - e_{it})! (X_{it-1} - (X_{it} - e_{it}))!} \left(\frac{\mu_{z'_i} p}{(1-p)} \right)^{e_{it}} \times \mathbb{1}_{[\max(0, X_{it}-X_{it-1}), X_{it}]}(e_{it}).$$

Although innovations are not of our interest and collapsed in other steps, the sampled innovations are used to draw μ_i and p in the following steps.

4. Sample $\{\mu_i\}$ and α' given other variables using the Chinese restaurant process (CRP). Detailed derivation is given by Aldor-Noiman et al..

- (a) Draw a sample for the cluster index z'_i for all i given $\{e_{it}\}$ with

$$p(z'_i = k', k' \leq \widehat{K}') \propto n_{k',-i} \prod_{t=2}^T \text{Pois}(e_{it} | \mu_{k'}) \text{ and}$$

$$p(z'_i = \widehat{K}' + 1) \propto \alpha \frac{b'^{a'}}{(b')^{a' + \sum_{t=2}^T e_{it}}} \prod_{t=2}^T \left(\frac{1}{e_{it}!} \right) \frac{\Gamma(a' + \sum_{t=2}^T e_{it})}{\Gamma(a')}.$$

If a new cluster is formed, $\widehat{K}' = \widehat{K}' + 1$.

- (b) Draw a sample for the cluster mean μ_k from $k'=1$ to $\widehat{K}'$ clusters as below

$$\mu_k | \{e_{it}\}, z'_i = k' \sim \text{Gamma}(a' + \sum_{i \in \{i: z'_i = k'\}} \sum_{t=2}^T e_{it}, b' + n'_k(T-1)).$$

(c) The concentration parameter α' is sampled according to Escobar and West.

5. Sample the thinning parameter p from

$$p|rest \propto Beta(\sum_{i=1}^n \sum_{t=2}^T (X_{it} - e_{it}) + 1, \sum_{i=1}^n \sum_{t=2}^T (X_{it-1} - (X_{it} - e_{it})) + 1).$$

6. Sample σ, ρ using the separation strategy that separates out the standard deviations and correlations from the covariance (Barnard et al., 2000). Since $\Sigma(\sigma, \rho) = \text{diag}(\sigma) R(\rho) \text{diag}(\sigma)$, where $\Sigma(\sigma, \rho)$ is the matrix obtained by the values of σ and ρ , we draw each component σ and ρ one at time in a Gibbs sampler. We assume log normal prior distribution for standard deviation and sample σ^{cand} from Metropolis-Hastings algorithm using $N(\log(\sigma^{curr}), 1)$ as a proposal distribution where σ^{curr} denotes current value of σ . We accept this candidate with the probability $\frac{\prod_i f(\mathbf{y}_i; \mathbf{x}_i, \beta, \sigma^{cand}, \rho) g(\sigma^{curr} | \sigma^{cand})}{\prod_i f(\mathbf{y}_i; \mathbf{x}_i, \beta, \sigma^{curr}, \rho) g(\sigma^{cand} | \sigma^{curr})}$ where $g(\cdot)$ is a probability density function of proposal distribution.

Then, given updated value for σ , we need to sample ρ with a requirement of the positive definite R . Since $|R(\rho)| > 0$ is a sufficient condition for $R(\rho)$ to be positive definite, we sample a candidate ρ^* from $Unif(-1, 1)$ that satisfies $|R(\rho)| > 0$. We accept this candidate with the probability $\frac{\prod_i f(\mathbf{y}_i; \mathbf{x}_i, \beta, \sigma, \rho^{cand})}{\prod_i f(\mathbf{y}_i; \mathbf{x}_i, \beta, \sigma, \rho^{curr})}$.

7. Sample β from

$$\beta|\{y_{it}\}, \{X_{it}\}, \Sigma(\sigma, \rho) \sim N_2 \left((\sum_{i=1}^n \mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i)^{-1} \sum_{i=1}^n \mathbf{x}_i^T \Sigma^{-1} \mathbf{y}_i, (\sum_{i=1}^n \mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i)^{-1} \right).$$

8. For $t = 1, \dots, T$, sample $(sen_t^A, ppv_t^A, sen_t^B, ppv_t^B)$. We simultaneously sample sensitivity and PPV values of two sources by Metropolist-Hastings algorithm. We draw candidates $(sen_t^{A*}, ppv_t^{A*}, sen_t^{B*}, ppv_t^{B*})$ each from truncated normal distribution with lower truncation at 0, upper truncation at 1, current values as mean, and standard deviation 0.1. The truncation is to bound the values $sen_t^s, ppv_t^s \in (0, 1)$. Then, the straightforward transformation yields $(\theta_{11t}^{A*}, \theta_{12t}^{A*}, \theta_{21t}^{A*}, \theta_{11t}^{B*}, \theta_{12t}^{B*}, \theta_{21t}^{B*})$. We accept the set of candidates with the probability

$$\frac{\prod_i f(W_{it}^A; \hat{x}_{it}, sen_t^{A*}, ppv_t^{A*}) f(W_{it}^B; \hat{x}_{it}, sen_t^{B*}, ppv_t^{B*}) \pi(\theta_t^{A*}; \gamma^A) \pi(\theta_t^{B*}; \gamma^B)}{\prod_i f(W_{it}^A; \hat{x}_{it}, sen_t^{1(r)}, ppv_t^{1(r)}) f(W_{it}^B; \hat{x}_{it}, sen_t^{B(r)}, ppv_t^{B(r)}) \pi(\theta_t^{1(r)}; \gamma^A) \pi(\theta_t^{B(r)}; \gamma^B)} \times$$

$$\frac{h(\text{sen}_t^{1(r)}, \text{ppv}_t^{1(r)}, \text{sen}_t^{B(r)}, \text{ppv}_t^{B(r)}; \text{sen}_t^{A*}, \text{ppv}_t^{A*}, \text{sen}_t^{B*}, \text{ppv}_t^{B*})}{h(\text{sen}_t^{A*}, \text{ppv}_t^{A*}, \text{sen}_t^{B*}, \text{ppv}_t^{B*}; \text{sen}_t^{1(r)}, \text{ppv}_t^{1(r)}, \text{sen}_t^{B(r)}, \text{ppv}_t^{B(r)})}$$

where θ_t denotes $(\theta_{11t}, \theta_{12t}, \theta_{21t})$ and $h(\cdot)$ denotes the density of proposal distribution.

C.3 Simulation details

Figure C.1: Simulated true exposure $\{X_{it}\}$ for $t = 1, \dots, 8$ and $N = 300$ with different scenarios of p . The value of individual expected baseline exposure is either $\lambda_{i1} = 5$ or $\lambda_{i1} = 15$ with the probability 0.9 and 0.1, respectively. Exposures of subjects with $\lambda_{i1} = 5$ are light gray curves and exposures of subjects with $\lambda_{i1} = 15$ are colored in dark gray. Arrival rate μ_i takes a value of 0.5 or 1.5, with equal probabilities. Magenta line shows the curve of annual sample mean of X_{it} by the combination of arrival rate and expected baseline exposure.

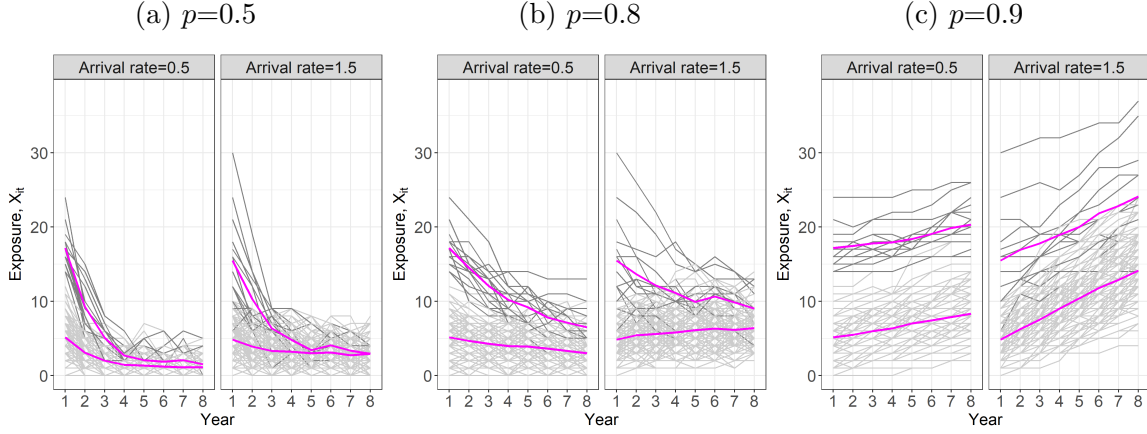
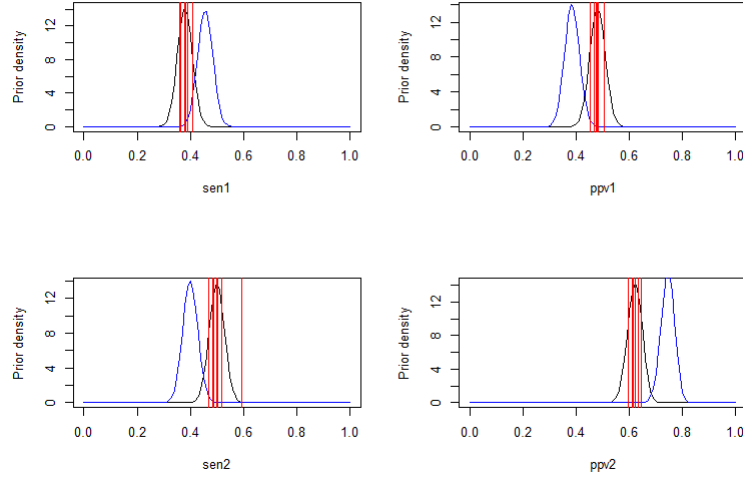
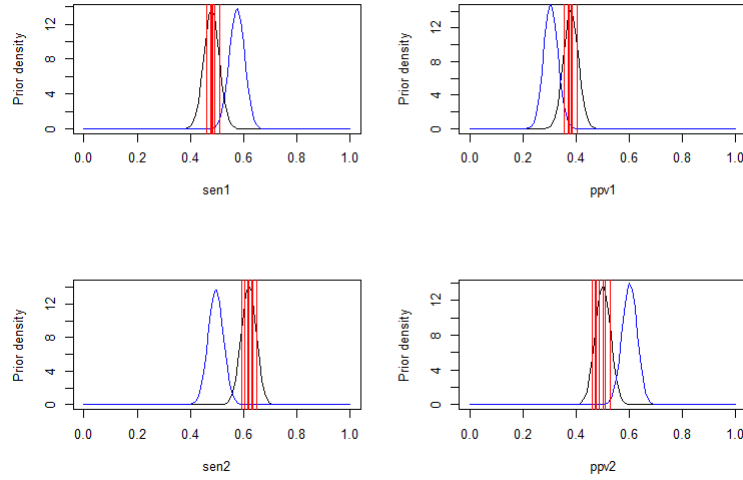


Figure C.2: True sensitivities and PPVs (red vertical lines) when $T = 8$. Prior with correct information ($sen_T^A, ppv_T^A, sen_T^B, ppv_T^B$) matches the data generating distribution (black curve). Misspecified prior (blue curve) shows incorrect prior beliefs about source qualities.

$$(a) \ sen_t^A < ppv_t^A, \ sen_t^B < ppv_t^B$$



$$(b) \ sen_t^A > ppv_t^A, \ sen_t^B > ppv_t^B$$



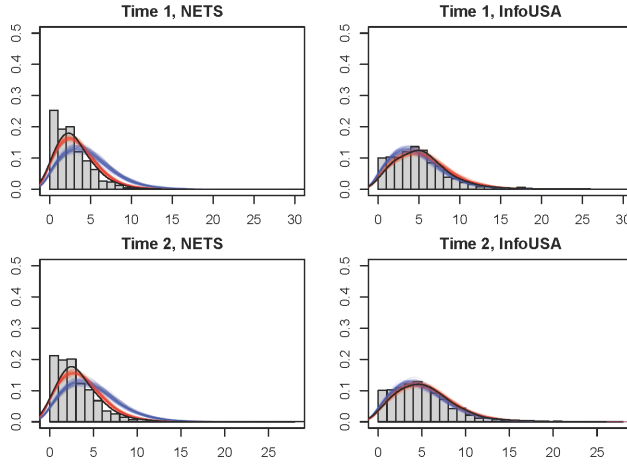
C.4 Application results

Table C.1: Posterior summaries of variables from Childhood BMI study in Section 4.5.

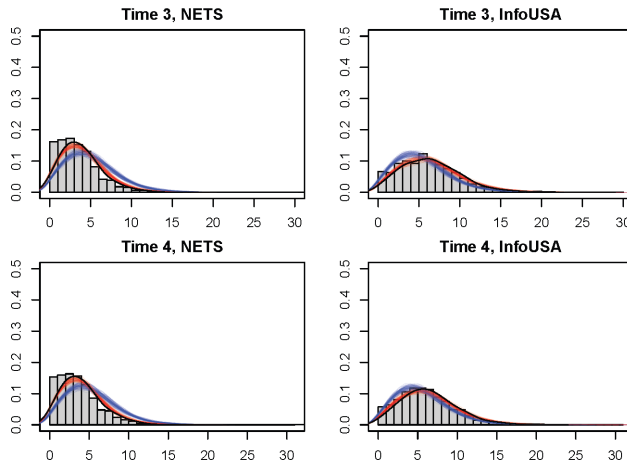
	Grade 5, Female				Grade 5, Male				Grade 5, Female				Grade 5, Female			
	Mean	SD	2.50%	97.50%	Mean	SD	2.50%	97.50%	Mean	SD	2.50%	97.50%	Mean	SD	2.50%	97.50%
p	0.998	0.001	0.996	0.999	0.998	0.001	0.997	0.999	0.994	0.002	0.990	0.997	0.994	0.002	0.990	0.997
Intercept	0.349	0.028	0.295	0.402	0.478	0.026	0.428	0.528	0.390	0.034	0.324	0.456	0.387	0.030	0.329	0.448
$\log_2(\exp+1)$	0.025	0.004	0.017	0.033	0.024	0.004	0.016	0.031	0.016	0.006	0.005	0.028	0.018	0.005	0.008	0.028
Year	0.008	0.002	0.004	0.013	0.015	0.002	0.011	0.019	0.012	0.003	0.007	0.017	0.015	0.003	0.010	0.020
Spline	-0.017	0.004	-0.026	-0.009	-0.022	0.004	-0.030	-0.013	-0.023	0.005	-0.033	-0.012	-0.028	0.005	-0.038	-0.017
% Others	0.420	0.054	0.313	0.524	0.561	0.050	0.462	0.659	0.214	0.071	0.076	0.352	0.227	0.063	0.105	0.350
% Asian	-0.266	0.028	-0.320	-0.213	-0.030	0.026	-0.080	0.020	-0.223	0.038	-0.297	-0.150	-0.032	0.034	-0.099	0.037
% Black non-Hispanic	0.522	0.027	0.468	0.575	0.266	0.026	0.215	0.317	0.690	0.040	0.613	0.769	0.413	0.036	0.345	0.483
% Hispanic	0.664	0.021	0.623	0.705	0.684	0.020	0.645	0.723	0.584	0.027	0.530	0.638	0.609	0.024	0.561	0.656
School Size	-0.006	0.001	-0.008	-0.004	-0.003	0.001	-0.005	-0.002	-0.005	0.001	-0.006	-0.003	-0.002	0.001	-0.004	-0.001
Census tract median household income	-0.001	0.000	-0.001	0.000	-0.001	0.000	-0.002	-0.001	0.000	0.000	-0.001	0.001	0.000	0.000	-0.001	0.001
Census tract education level	-0.177	0.037	-0.249	-0.106	-0.055	0.035	-0.124	0.015	-0.186	0.047	-0.277	-0.093	-0.160	0.041	-0.240	-0.081
ρ	0.252	0.010	0.232	0.271	0.231	0.010	0.211	0.251	0.457	0.017	0.423	0.489	0.374	0.017	0.342	0.407
σ	1.634	0.012	1.612	1.657	1.627	0.011	1.605	1.649	2.060	0.031	1.998	2.121	1.997	0.029	1.938	2.058

Figure C.3: Annual distribution of *observed* exposures to convenience stores from 2001 to 2008 using NETS and InfoUSA, for 5th grade female students who attend urban schools in California. Black curve is an empirical density of *observed* exposures, which approximates the true density function using `density()` in R. Red curves are empirical density of posterior predictive W 's the proposed model that we estimate sen_t^s and ppv_t^s for $t = 1, \dots, T$, $s = A, B$ as described in Section 4.3. Blue curves represent the empirical density of posterior predictive W 's from the model that fixed sen_t^s and ppv_t^s to the reference values without any updates so that $(sen_t^{\text{NETS}}, ppv_t^{\text{NETS}}) = (0.3769, 0.4812)$ and $(sen_t^{\text{InfoUSA}}, ppv_t^{\text{InfoUSA}}) = (0.5047, 0.6182)$ for all t . This figure shows the need of estimating source credibility for historical data.

(a) Year: 2001, 2002

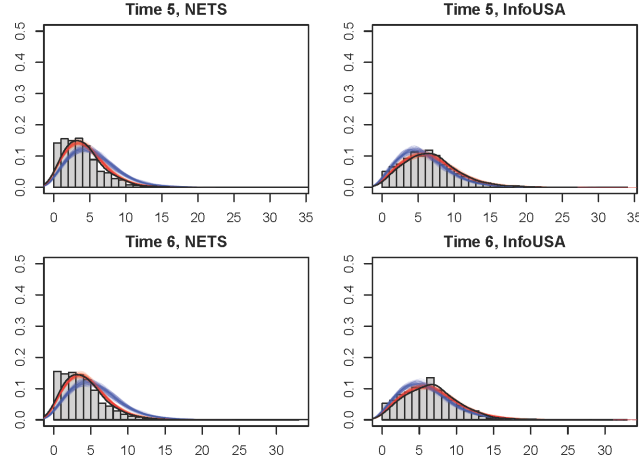


(b) Year: 2003, 2004

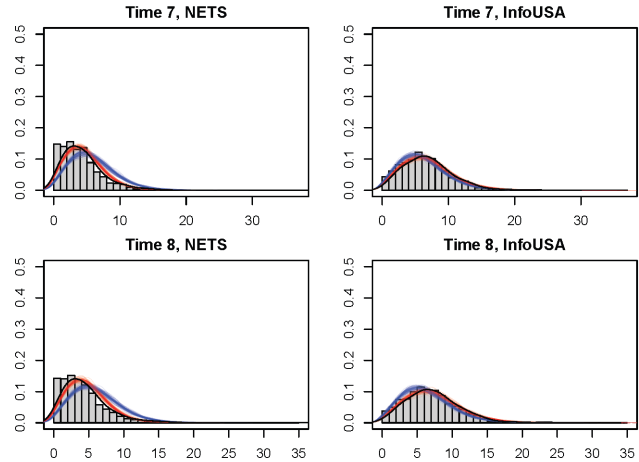


(Figure C.3 continued) Annual distribution of *observed* exposures to convenience stores from 2001 to 2008 using NETS and InfoUSA, for 5th grade female students who attend urban schools in California. Black curve is an empirical density of *observed* exposures, which approximates the true density function using `density()` in R. Red curves are empirical density of posterior predictive W 's the proposed model that we estimate sen_t^s and ppv_t^s for $t = 1, \dots, T$, $s = A, B$ as described in Section 4.3. Blue curves represent the empirical density of posterior predictive W 's from the model that fixed sen_t^s and ppv_t^s to the reference values without any updates so that $(sen_t^{\text{NETS}}, ppv_t^{\text{NETS}}) = (0.3769, 0.4812)$ and $(sen_t^{\text{InfoUSA}}, ppv_t^{\text{InfoUSA}}) = (0.5047, 0.6182)$ for all t . This figure shows the need of estimating source credibility for historical data.

(c) Year: 2005, 2006



(d) Year: 2007, 2008



BIBLIOGRAPHY

BIBLIOGRAPHY

- Aldor-Noiman, S., Brown, L. D., Fox, E. B., and Stine, R. A. (2016). Spatio-temporal low count processes with application to violent crime events. *Statistica Sinica* pages 1587–1610.
- Aldous, D. J. (1985). Exchangeability and related topics. In *École d’Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.
- Alexeeff, S. E., Carroll, R. J., and Coull, B. (2016). Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics* **17**, 377–389.
- Alzaid, A. and Al-Osh, M. (1988). First-order integer-valued autoregressive (INAR (1)) process: distributional and regression properties. *Statistica Neerlandica* **42**, 53–61.
- Apanasovich, T. V., Carroll, R. J., and Maity, A. (2009). SIMEX and standard error estimation in semiparametric measurement error models. *Electronic Journal of Statistics* **3**, 318–348.
- Armstrong, B. (1985). Measurement error in the generalised linear model. *Communications in Statistics-Simulation and Computation* **14**, 529–544.
- Athens, J. K., Duncan, D. T., and Elbel, B. (2016). Proximity to fast-food outlets and supermarkets as predictors of fast-food dining frequency. *Journal of the Academy of Nutrition and Dietetics* **116**, 1266–1275.
- Auchincloss, A. H., Moore, K. A., Moore, L. V., and Diez Roux, A. V. (2012). Improving retrospective characterization of the food environment for a large region in the united states during a historic time period. *Health Place* **18**, 1341–1347.
- Bader, M. D., Ailshire, J. A., Morenoff, J. D., and House, J. S. (2010). Measurement of the local food environment: a comparison of existing data sources. *American journal of epidemiology* **171**, 609–617.
- Barker, R. J., Schofield, M. R., Link, W. A., and Sauer, J. R. (2018). On the reliability of n-mixture models for count data. *Biometrics* **74**, 369–377.
- Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

- Barnatchez, K., Crane, L. D., and Decker, R. (2017). An assessment of the national establishment time series (NETS) database, Finance and Economics Discussion Series 2017-110, Washington: Board of Governors of the Federal Reserve System.
- Bartolucci, F., Pandolfi, S., and Pennoni, F. (2022). Discrete latent variable models. *Annual Review of Statistics and Its Application* **9**, 425–452.
- Berger, N., Kaufman, T. K., Bader, M. D., Rundle, A. G., Mooney, S. J., Neckerman, K. M., and Lovasi, G. S. (2019). Disparities in trajectories of changes in the unhealthy food environment in new york city: a latent class growth analysis, 1990–2010. *Social Science & Medicine* **234**, 112362.
- Berke, E. M., Koepsell, T. D., Moudon, A. V., Hoskins, R. E., and Larson, E. B. (2007). Association of the built environment with physical activity and obesity in older persons. *American Journal of Public Health* **97**, 486–492.
- Böckenholt, U. (1998). Mixed INAR(1) Poisson regression models: Analyzing heterogeneity and serial dependencies in longitudinal count data. *Journal of Econometrics* **89**, 317–338.
- Boone, J. E., Gordon-Larsen, P., Stewart, J. D., and Popkin, B. M. (2008). Validation of a gis facilities database: quantification and implications of error. *Annals of Epidemiology* **18**, 371–377.
- Bourguignon, M., LP Vasconcellos, K., Reisen, V. A., and Ispány, M. (2016). A Poisson INAR(1) process with a seasonal structure. *Journal of Statistical Computation and Simulation* **86**, 373–387.
- Brandt, P. T. and Williams, J. T. (2001). A linear Poisson autoregressive model: The poisson AR(p) model. *Political Analysis* **9**, 164–184.
- Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC press.
- Buonaccorsi, J. P., Laake, P., and Veierød, M. B. (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics* **61**, 831–836.
- Butz, W. P. and Torrey, B. B. (2006). Some frontiers in social science. *Science* **312**, 1898–1900.
- California Department of Education (2019). California Department of Education, Physical Fitness Testing (PFT). <http://www.cde.ca.gov/ta/tg/pf/>, Accessed: 2019-06-05.
- Carroll, R. J. (1998). Measurement error in epidemiologic studies. *Encyclopedia of biostatistics* **3**, 2491–2519.
- Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association* **88**, 185–199.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective, 2nd edition*. Chapman and Hall, FL.

- Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* **85**, 652–663.
- Caspi, C. E. and Friebur, R. (2016). Modified ground-truthing: an accurate and cost-effective food environment validation method for town and rural areas. *International Journal of Behavioral Nutrition and Physical Activity* **13**, 1–8.
- Cheng, Y., Chakraborty, A., and Datta, G. (2015). Hierarchical Bayesian methods for combining surveys. In *Proceedings of the Survey Research Methods Section*, pages 4099–4111. American Statistical Association Alexandria, VA.
- Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association* **89**, 1314–1328.
- Cummins, S. and Macintyre, S. (2009). Are secondary data sources on the neighbourhood food environment accurate? case-study in glasgow, uk. *Preventive medicine* **49**, 527–528.
- Dail, D. and Madsen, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics* **67**, 577–587.
- Denes, F. V., Silveira, L. F., and Beissinger, S. R. (2015). Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution* **6**, 543–556.
- Diana, A., Griffin, J., and Matechou, E. (2018). A polya tree based model for unmarked individuals in an open wildlife population. In *International Conference on Bayesian Statistics in Action*, pages 3–11. Springer.
- Diana, A., Matechou, E., Griffin, J., and Johnston, A. (2020). A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK. *The Annals of Applied Statistics* **14**, 473–493.
- Diez Roux, A. V., Mujahid, M. S., Hirsch, J. A., Moore, K., and Moore, L. V. (2016). The impact of neighborhoods on cardiovascular risk: the MESA neighborhood study. *Global Heart* **11**, 353–363.
- Dong, X. L., Berti-Equille, L., and Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment* **2**, 550–561.
- Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., and Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64**, 635–644.
- Dun and Bradstreet (2021). <https://www.dnb.com/duns-number.html>, Accessed: 2020-03-19.
- D’Angelo, H., Fleischhacker, S., Rose, S. W., and Ribisl, K. M. (2014). Field validation of secondary data sources for enumerating retail tobacco outlets in a state without tobacco outlet licensing. *Health & place* **28**, 38–44.

- Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics* **53**, 262–272.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association* **90**, 577–588.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.
- Fienberg, S. E. (1972a). The analysis of incomplete multi-way contingency tables. *Biometrics* **28**, 177–202.
- Fienberg, S. E. (1972b). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika* **59**, 591–603.
- Flegal, K. M., Wei, R., Ogden, C. L., Freedman, D. S., Johnson, C. L., and Curtin, L. R. (2009). Characterizing extreme values of body mass index–for–age by using the 2000 centers for disease control and prevention growth charts. *The American journal of clinical nutrition* **90**, 1314–1320.
- Fleischhacker, S. E., Evenson, K. R., Sharkey, J., Pitts, S. B. J., and Rodriguez, D. A. (2013). Validity of secondary retail food outlet data: a systematic review. *American journal of preventive medicine* **45**, 462–473.
- Fleischhacker, S. E., Rodriguez, D. A., Evenson, K. R., Henley, A., Gizlice, Z., Soto, D., and Ramachandran, G. (2012). Evidence for validity of five secondary data sources for enumerating retail food outlets in seven american indian communities in north carolina. *International Journal of Behavioral Nutrition and Physical Activity* **9**, 1–14.
- Fletcher-Lartey, S. M. and Caprarelli, G. (2016). Application of GIS technology in public health: successes and challenges. *Parasitology* **143**, 401.
- Fokianos, K. and Tjøstheim, D. (2011). Log-linear poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578.
- Gelman, A. and Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. *Bayesian statistics* **4**, 625–631.
- Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics* **28**, 35.
- Goldberg, D. W. (2008). A Geocoding Best Practices Guide. North American Association of Central Cancer Registries, Springfield, IL.
- Grafova, I. B. (2008). Overweight children: assessing the contribution of the built environment. *Preventive medicine* **47**, 304–308.
- Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American journal of epidemiology* **112**, 564–569.
- Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International journal of epidemiology* **29**, 722–729.

- Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical science* **20**, 111–140.
- Gustafson, P., Le, N. D., and Saskin, R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57**, 598–609.
- Gustafson, P. and Le Nhu, D. (2002). Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics* **58**, 878–887.
- Healy, M. A. and Gilliland, J. A. (2012). Quantifying the magnitude of environmental exposure quadification when using imprecise address proxies in public health research. *Spatial and Spatio-temporal Epidemiology* **3**, 55–67.
- Held, L., Hofmann, M., Höhle, M., and Schmid, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics* **7**, 422–437.
- Hickson, D. A., Diez Roux, A. V., Smith, A. E., Tucker, K. L., Gore, L. D., Zhang, L., and Wyatt, S. B. (2011). Associations of fast food restaurant availability with dietary intake and weight among african americans in the jackson heart study, 2000–2004. *American Journal of Public Health* **101**, S301–S309.
- Hirsch, J. A., Meyer, K. A., Peterson, M., Rodriguez, D. A., Song, Y., Peng, K., Huh, J., and Gordon-Larsen, P. (2016). Obtaining longitudinal built environment data retrospectively across 25 years in four us cities. *Frontiers in Public Health* **4**, 65.
- Hirsch, J. A., Moore, K. A., Cahill, J., Quinn, J., Zhao, Y., Bayer, F. J., Rundle, A., and Lovasi, G. S. (2020). Business data categorization and refinement for application in longitudinal neighborhood health research: a methodology. *Journal of Urban Health* **98**, 1–14.
- Hougaard, P., Lee, M.-L. T., and Whitmore, G. (1997). Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. *Biometrics* **53**, 1225–1238.
- Howard, P. H., Fitzpatrick, M., and Fulfroost, B. (2011). Proximity of food retailers to schools and rates of overweight ninth grade students: an ecological study in california. *BMC public health* **11**, 1–8.
- Huang, Y. and Wang, C. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association* **95**, 1209–1219.
- Imai, K. and Yamamoto, T. (2010). Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science* **54**, 543–560.
- InfoUSA (2021). Infousa. infousa business listing description. <https://www.infousa.com/Product/Business-Lists/>, Accessed: 2021-02-05.
- Ingram, D. D. and Franco, S. J. (2014). 2013 NCHS urban–rural classification scheme for counties.

- Jacquez, G. M. (2012). A research agenda: does geocoding positional error matter in health gis studies? *Spatial and spatio-temporal epidemiology* **3**, 7–16.
- Joe, H. (2014). *Dependence modeling with copulas*. CRC press.
- Joe, H. and Zhu, R. (2005). Generalized poisson distribution: the property of mixture of poisson and comparison with negative binomial distribution. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **47**, 219–229.
- Jones, K. K., Zenk, S. N., Tarlov, E., Powell, L. M., Matthews, S. A., and Horoi, I. (2017). A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments. *BMC research notes* **10**, 1–12.
- Kaufman, T. K., Sheehan, D. M., Rundle, A., Neckerman, K. M., Bader, M. D., Jack, D., and Lovasi, G. S. (2015a). Measuring health-relevant businesses over 21 years: refining the National Establishment Time-Series (NETS), a dynamic longitudinal data set. *BMC Res Notes* **8**, 507.
- Kaufman, T. K., Sheehan, D. M., Rundle, A., Neckerman, K. M., Bader, M. D., Jack, D., and Lovasi, G. S. (2015b). Measuring health-relevant businesses over 21 years: refining the national establishment time-series (NETS), a dynamic longitudinal data set. *BMC research notes* **8**, 1–13.
- Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V., et al. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment. *Statistics in medicine* **39**, 2197–2231.
- Kery, M., Dorazio, R. M., Soldaat, L., Van Strien, A., Zuiderwijk, A., and Royle, J. A. (2009). Trend estimation in populations with imperfect detection. *Journal of Applied Ecology* **46**, 1163–1172.
- Küchenhoff, H., Lederer, W., and Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics & Data Analysis* **51**, 6197–6211.
- Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* **62**, 85–96.
- Kukush, A., Schneeweis, H., and Wolf, R. (2004). Three estimators for the poisson regression model with measurement errors. *Statistical Papers* **45**, 351–368.
- Larson, N. I., Story, M. T., and Nelson, M. C. (2009). Neighborhood environments: disparities in access to healthy foods in the U.S. *Am J Prev Med* **36**, 74–81.
- Lebel, A., Daepf, M. I., Block, J. P., Walker, R., Lalonde, B., Kestens, Y., and Subramanian, S. (2017). Quantifying the foodscape: a systematic review and meta-analysis of the validity of commercially available business data. *PLoS One* **12**, e0174417.
- Lee, H. (2012). The role of local food availability in explaining obesity risk among young school-aged children. *Social science & medicine* **74**, 1193–1203.

- Levin, B. (1981). A representation for multinomial cumulative distribution functions. *The Annals of Statistics* **9**, 1123–1126.
- Li, L., Palta, M., and Shao, J. (2004). A measurement error model with a poisson distributed surrogate. *Statistics in medicine* **23**, 2527–2536.
- Liao, X., Zucker, D. M., Li, Y., and Spiegelman, D. (2011). Survival analysis with error-prone time-varying covariates: A risk set calibration approach. *Biometrics* **67**, 50–58.
- Liese, A. D., Barnes, T. L., Lamichhane, A. P., Hibbert, J. D., Colabianchi, N., and Lawson, A. B. (2013). Characterizing the food retail environment: impact of count, type, and geospatial error in 2 secondary data sources. *Journal of nutrition education and behavior* **45**, 435–442.
- Liese, A. D., Colabianchi, N., Lamichhane, A. P., Barnes, T. L., Hibbert, J. D., Porter, D. E., Nichols, M. D., and Lawson, A. B. (2010). Validation of 3 food outlet databases: completeness and geospatial accuracy in rural and urban food environments. *American journal of epidemiology* **172**, 1324–1333.
- Lockwood, J. R. and McCaffrey, D. F. (2017). Simulation-Extrapolation with Latent Heteroskedastic Error Variance. *Psychometrika* .
- Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science* **32**, 293–312.
- Lovasi, G. S. (2015). Communities designed to support cardiovascular health for older adults. <https://grantome.com/grant/NIH/R01-AG049970-01A1>.
- Lucan, S. C., Maroko, A. R., Bumol, J., Torrens, L., Varona, M., and Berke, E. M. (2013). Business list vs ground observation for measuring a food environment: saving time or waste of time (or worse)? *Journal of the Academy of Nutrition and Dietetics* **113**, 1332–1339.
- Manrique-Vallier, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* **72**, 1246–1254.
- Mao, G., Wei, Y., and Liu, Y. (2017). SIMEX method for censored quantile regression with measurement error. *Communications in Statistics-Simulation and Computation* **46**, 7552–7560.
- Martin, J., Royle, J. A., Mackenzie, D. I., Edwards, H. H., Kery, M., and Gardner, B. (2011). Accounting for non-independent detection when estimating abundance of organisms with a Bayesian approach. *Methods in Ecology and Evolution* **2**, 595–601.
- Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters* **8**, 1235–1246.
- Miller, J. W. and Harrison, M. T. (2014). Inconsistency of pitman-yor process mixtures for the number of components. *The Journal of Machine Learning Research* **15**, 3333–3370.

- Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**, 340–356.
- Must, A. and Anderson, S. (2006). Body mass index in children and adolescents: considerations for population-based applications. *International journal of obesity* **30**, 590–594.
- NAICS (2019). NAICS association, SIC Codes and Counts by Division. <https://www.naics.com/sic-codes-counts-division/?div=G>, Accessed: 2019-01-21.
- Narayanan, A. (1990). Computer generation of Dirichlet random vectors. *Journal of Statistical Computation and Simulation* **36**, 19–30.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9**, 249–265.
- Neath, A. A. and Samaniego, F. J. (1997). On the efficacy of bayesian inference for non-identifiable models. *The American Statistician* **51**, 225–232.
- Nelder, J. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society: Series B (Methodological)* **54**, 273–284.
- NETS (2021). Business Dynamics Research Consortium, National Establishment Time-series (NETS) Database: Database description. <http://exceptionalgrowth.org>, Accessed: 2021-02-05.
- Ortiz (2003). An act to amend section 49431 of, and to add section 49431.5 to, the education code, relating to schools.
- Osthus, D., Weaver, B., Casleton, E., Hamada, M., and Steiner, S. (2021). On gauge repeatability and reproducibility studies for counts. *Quality Engineering* **33**, 641–654.
- Parveen, N., Moodie, E., and Brenner, B. (2015). The non-zero mean SIMEX: Improving estimation in the face of measurement error. *Observational Studies* **1**, 91–123.
- Parveen, N., Moodie, E., and Brenner, B. (2017). Correcting covariate-dependent measurement error with non-zero mean. *Statistics in Medicine* **36**, 2786–2800.
- Pearce, J., Witten, K., and Bartie, P. (2006). Neighbourhoods and health: a gis approach to measuring community resource accessibility. *Journal of Epidemiology & Community Health* **60**, 389–395.
- Pietrobelli, A., Faith, M. S., Allison, D. B., Gallagher, D., Chiumello, G., and Heymsfield, S. B. (1998). Body mass index as a measure of adiposity among children and adolescents: a validation study. *The Journal of Pediatrics* .
- Pollock, K. H. and Otto, M. C. (1983). Robust estimation of population size in closed animal populations from capture-recapture experiments. *Biometrics* **39**, 1035–1049.
- Powell, L. M., Han, E., Zenk, S. N., Khan, T., Quinn, C. M., Gibbs, K. P., Pugach, O., Barker, D. C., Resnick, E. A., Myllyluoma, J., and Chaloupka, F. J. (2011). Field validation of secondary commercial data sources on the retail food outlet environment in the U.S. *Health Place* **17**, 1122–1131.

- Puig, P. and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli* **13**, 544–555.
- Richardson, S. and Gilks, W. R. (1993a). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology* **138**, 430–442.
- Richardson, S. and Gilks, W. R. (1993b). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine* **12**, 1703–1722.
- Rosner, B., Willett, W., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in medicine* **8**, 1051–1069.
- Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60**, 108–115.
- Rutt, C. D. and Coleman, K. J. (2005). Examining the relationships among built environment, physical activity, and body mass index in El Paso, TX. *Preventive medicine* **40**, 831–841.
- Sánchez, B. N., Sanchez-Vaznaugh, E. V., Uscilka, A., Baek, J., and Zhang, L. (2012). Differential associations between the food environment near schools and childhood overweight across race/ethnicity, gender, and grade. *American journal of epidemiology* **175**, 1284–1293.
- Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics* **8**, 341–377.
- Shaw, P. A., Deffner, V., Keogh, R. H., Tooze, J. A., Dodd, K. W., Küchenhoff, H., Kipnis, V., Freedman, L. S., Error, M., of the STRATOS, M. T. G. T., et al. (2018). Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. *Annals of epidemiology* **28**, 821–828.
- Skellam, J. G. (1946). The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society. Series A (General)* **109**, 296–296.
- Sonderman, J. S., Mumma, M. T., Cohen, S. S., Cope, E. L., Blot, W. J., and Signorello, L. B. (2012). A multi-stage approach to maximizing geocoding success in a large population-based cohort study through automated and interactive processes. *Geospatial health* **6**, 273.
- Song, X. and Huang, Y. (2006). A corrected pseudo-score approach for additive hazards model with longitudinal covariates measured with error. *Lifetime data analysis* **12**, 97–110.
- Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association* **95**, 51–61.

- Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72**, 583–592.
- Swartz, T. B., Haitovsky, Y., Vexler, A., and Yang, T. Y. (2004). Bayesian identifiability and misclassification in multinomial data. *Canadian Journal of Statistics* **32**, 285–302.
- Teh, Y. W. et al. (2010). Dirichlet process. *Encyclopedia of machine learning* **1063**, 280–287.
- Thamerus, M. (1997). Modelling count data with heteroscedastic measurement error in the covariates. Discussion Paper 58, SFB 386. Ludwig-Maximilians-Universität München.
- Thomas, D. C. (2009). *Statistical methods in environmental epidemiology*. Oxford University Press, USA.
- Thornton, L. E., Pearce, J. R., and Kavanagh, A. M. (2011). Using geographic information systems (gis) to assess the role of the built environment in influencing obesity: a glossary. *International Journal of Behavioral Nutrition and Physical Activity* **8**, 1–9.
- van Smeden, M., Lash, T. L., and Groenwold, R. H. (2020). Reflection on modern methods: five myths about measurement error in epidemiological research. *International journal of epidemiology* **49**, 338–347.
- Wang, N., Carroll, R., and Liang, K.-Y. (1996). Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics* **52**, 401–411.
- Wang, N., Carroll, R. J., Lin, X., and Gutierrez, R. G. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association* **93**, 249–261.
- Wang, Y., Jia, P., Cheng, X., and Xue, H. (2019). Improvement in food environments may help prevent childhood obesity: evidence from a 9-year cohort study. *Pediatric obesity* **14**, e12536.
- Wenger, S. J. and Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology* **89**, 2953–2959.
- Wong, M. S., Chan, K. S., Jones-Smith, J. C., Colantuoni, E., Thorpe Jr, R. J., and Bleich, S. N. (2018). The neighborhood environment and obesity: Understanding variation by race/ethnicity. *Preventive medicine* **111**, 371–377.
- Yi, G. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer Science and Business Media, LLC.
- Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health* **7**, 37.
- Zandbergen, P. A., Hart, T. C., Lenzer, K. E., and Camponovo, M. E. (2012). Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets. *Spatial and Spatio-temporal Epidemiology* **3**, 69–82.

- Zenk, S. N., Tarlov, E., Wing, C., Matthews, S. A., Tong, H., Jones, K. K., and Powell, L. M. (2018). Long-term weight loss effects of a behavioral weight management program: Does the community food environment matter? *International journal of environmental research and public health* **15**, 211.
- Zhao, B., Rubinstein, B. I. P., Gemmell, J., and Han, J. (2012). A Bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.* **5**, 550–561.
- Zhou, Q., Zhao, L., Zhang, L., Xiao, Q., Wu, T., Visscher, T., Zhao, J., Xin, J., Yu, X., Xue, H., et al. (2021). Neighborhood supermarket access and childhood obesity: a systematic review. *Obesity Reviews* **22**, e12937.
- Zimmerman, D. L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics* **64**, 262–270.
- Zimmerman, D. L. and Li, J. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *International Journal of Health Geographics* **9**, 1–11.
- Zimmerman, D. L., Li, J., and Fang, X. (2010). Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in medicine* **29**, 1025–1036.