COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Change in Local Healthy Food Retail Environment by Interactions in Population, Race and Nativity:

Prediction of change in local environment with longitudinal and spatial data emphasizing model generalizability with regularization, resampling and model aggregation approaches.

David Wutchiett, Tanya Kaufman, Daniel Sheehan, Kathryn Neckerman, Kayip Kwan, Andrew Rundle, Stephen Mooney, Jeff Goldsmith, and Gina Lovasi

Population Association of America
Annual Meeting 2015

- Measure relationships between demographics, local environment and change in healthy food environment.

- Measure relationships between demographics, local environment and change in healthy food environment.

- Outline a modeling approach for prediction of change using spatial and longitudinal information; interactions.

- Measure relationships between demographics, local environment and change in healthy food environment.

- Outline a modeling approach for prediction of change using spatial and longitudinal information; interactions.

- Evaluate our modeling approaches using resampling, validation and model aggregation approaches.

Healthy Food Environment Matters

- Local characteristics and demographics are associated with presence of healthy food retail outlets (Morland et al., 2002; Moore et al., 2008; Powell et al., 2007).
- Food sociodemographic characteristics and environment have been linked to population health (Cummins and Macintyre, 2006; Lovasi et al., 2009).

Healthy food outlets = {large supermarkets, fruit & vegetable markets, natural food markets & nut stores, fish markets}

A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
  - Particularly, direction of change.
- Gain insight into processes leading to divergence in built environment, local resources, and disparities.

A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
  - Particularly, direction of change.
- Gain insight into processes leading to divergence in built environment, local resources, and disparities.

National Establishment Time Series (NETS)

Longitudinal ‘census of U.S. businesses’ based on annual snapshot of Dunn & Bradstreet data

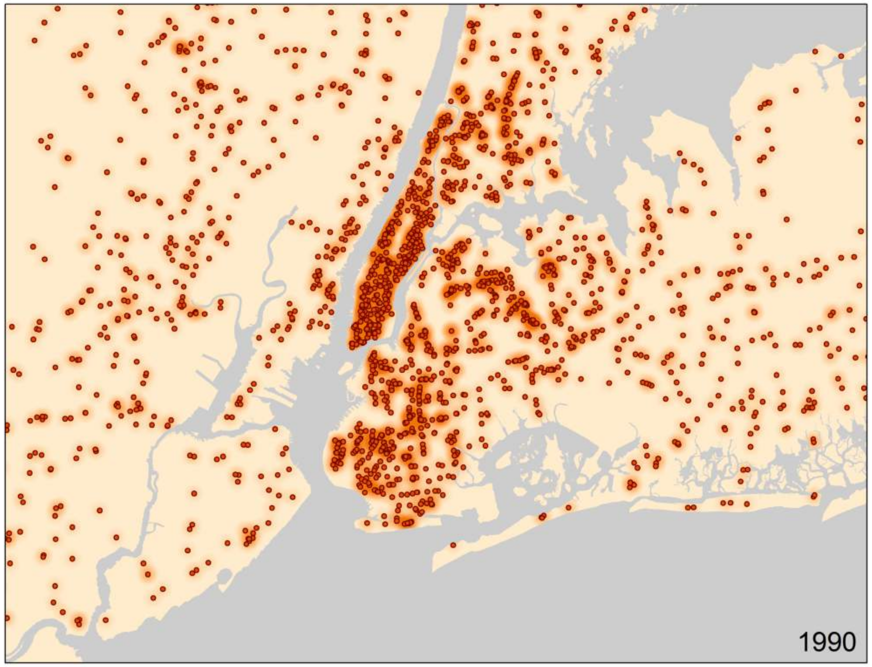
Geocoded to addresses

8 digit SIC code classifications

21 years of data

23 counties in NYC metropolitan area

Healthy Food Outlets; NYC Metro



Created by Daniel Sheehan

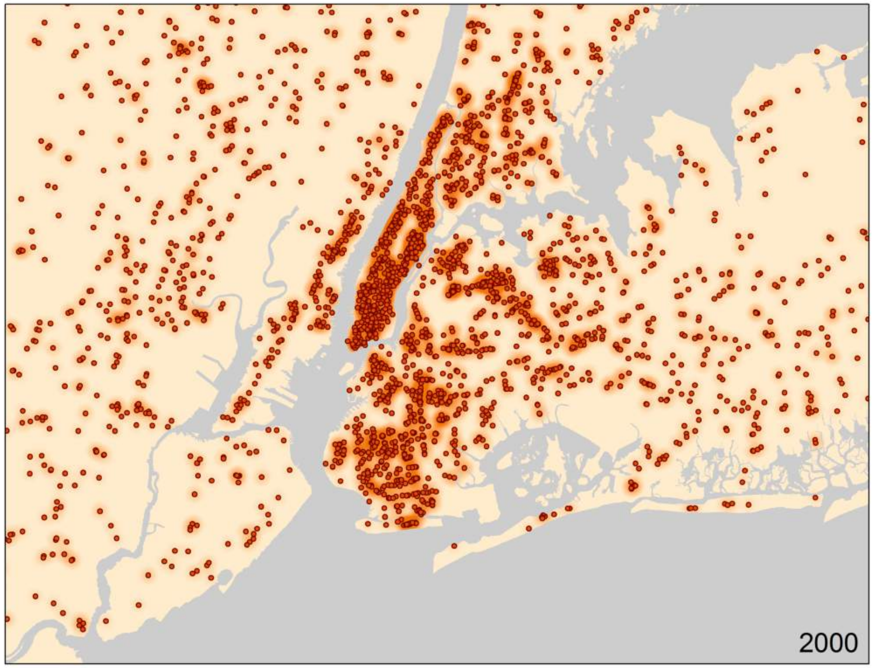
A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
  - Particularly, direction of change.
- Gain insight into processes leading to divergence in built environment, local resources, and disparities.

National Establishment Time Series (NETS)

Longitudinal 'census of U.S. businesses' based on annual snapshot of Dunn & Bradstreet data
Geocoded to addresses
8 digit SIC code classifications
21 years of data
23 counties in NYC metropolitan area

Healthy Food Outlets; NYC Metro

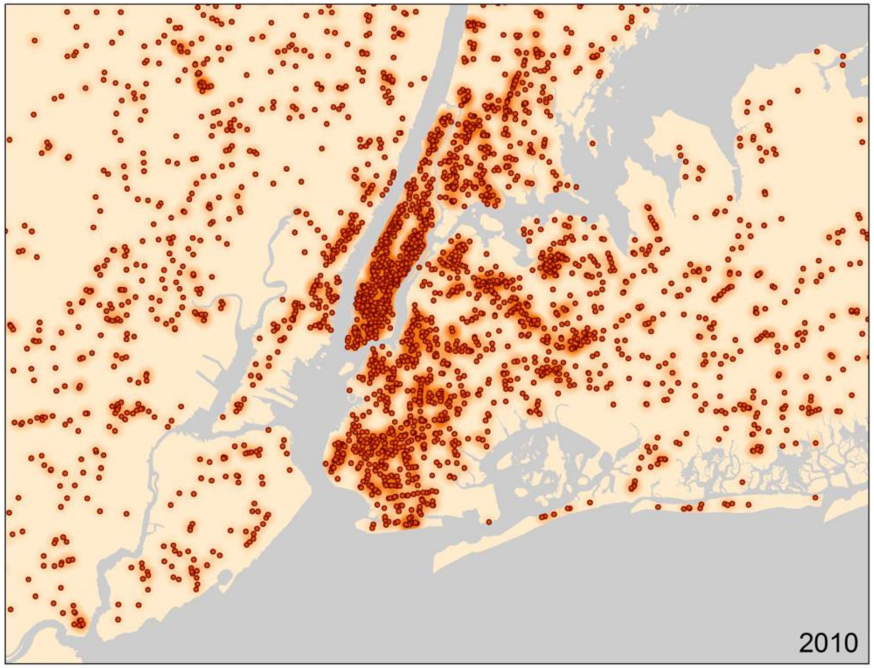


Created by Daniel Sheehan

A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
  - Particularly, direction of change.
- Gain insight into processes leading to divergence in built environment, local resources, and disparities.

Healthy Food Outlets; NYC Metro



Created by Daniel Sheehan

National Establishment Time Series (NETS)

Longitudinal ‘census of U.S. businesses’ based on annual snapshot of Dunn & Bradstreet data

Geocoded to addresses

8 digit SIC code classifications

21 years of data

23 counties in NYC metropolitan area

A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
  - Particularly, direction of change.
- Gain insight into processes leading to divergence in built environment, local resources, and disparities.

U.S. Census

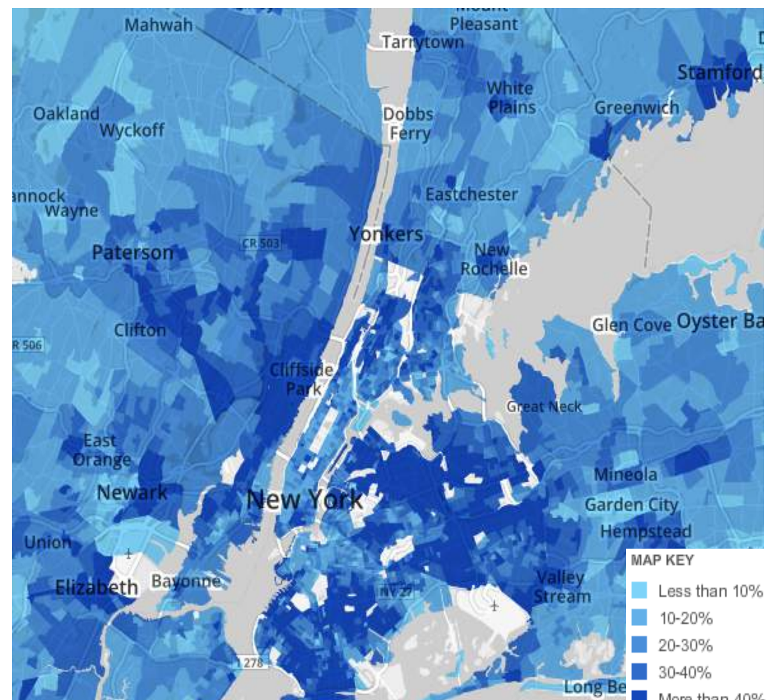
Decennial Census (1990, 2000, 2010)

American Community Survey (2007-2010)

Geographic size

Tract adjacency

% Foreign Born Population

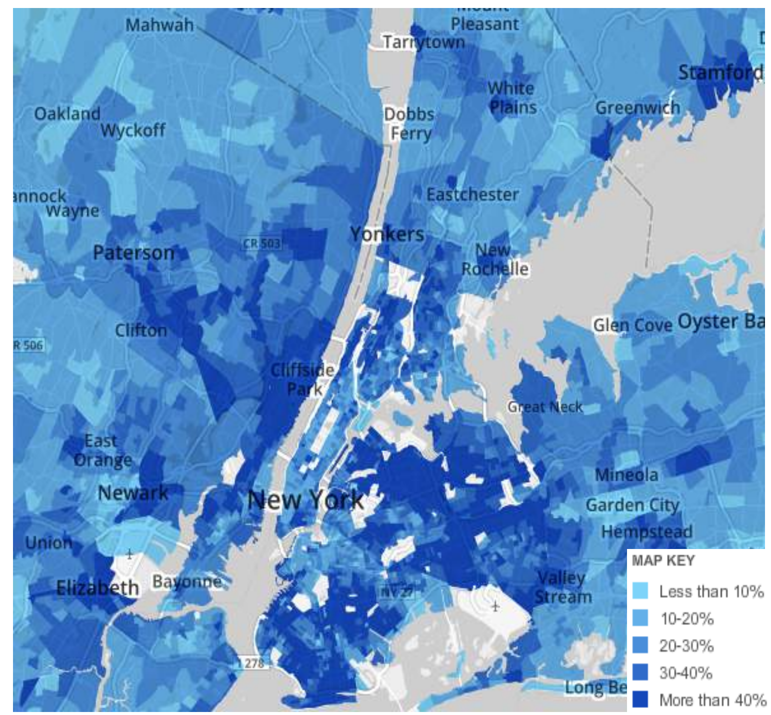


http://projects.nytimes.com/census/2010/explorer

A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
  - Particularly, direction of change.
- Gain insight into processes leading to divergence in built environment, local resources, and disparities.

% Foreign Born Population



http://projects.nytimes.com/census/2010/explorer

U.S. Census

Decennial Census (1990, 2000, 2010)

American Community Survey (2007-2010)

Geographic size

Tract adjacency

Population Meaures:

Median Income

% Poverty

% Foreign Born

% Non-Hispanic Black

% Hispanic

% Asian

Total Population

Census Tract Geographic Size

Derived Variables

Original Value

Population Meaures:

- Median Income
- % Poverty
- % Foreign Born
- % Non-Hispanic Black
- % Hispanic
- % Asian
- Total Population
- Census Tract Geographic Size

$$\%ForeignBorn_{t_n}$$

Derived Variables

Population Meaures:

- Median Income
- % Poverty
- % Foreign Born
- % Non-Hispanic Black
- % Hispanic
- % Asian
- Total Population
- Census Tract Geographic Size

Original Value

Decade Change

$$\nabla \%ForeignBorn_{t_n} = \%ForeignBorn_{t_n} - \%ForeignBorn_{t_{n-1}}$$

Derived Variables

Population Meaures:

- Median Income
- % Poverty
- % Foreign Born
- % Non-Hispanic Black
- % Hispanic
- % Asian
- Total Population
- Census Tract Geographic Size

Original Value

Decade Change

Adjacent Tracts

$$\nabla\% ForeignBorn\_Adjacent_{i,t_n} = \frac{\sum_j^{N(i)} \nabla\% \; Poverty_{\,j,t_n} * Population_{j,t_n}}{\sum_j^{N(i)} Population_{j,t_n}}$$

Derived Variables

<u>Population Meaures:</u>

- Median Income
- % Poverty
- % Foreign Born
- % Non-Hispanic Black
- % Hispanic
- % Asian
- Total Population
- Census Tract Geographic Size

Original Value

Decade Change

Adjacent Tracts

Change in Adjacent Tracts

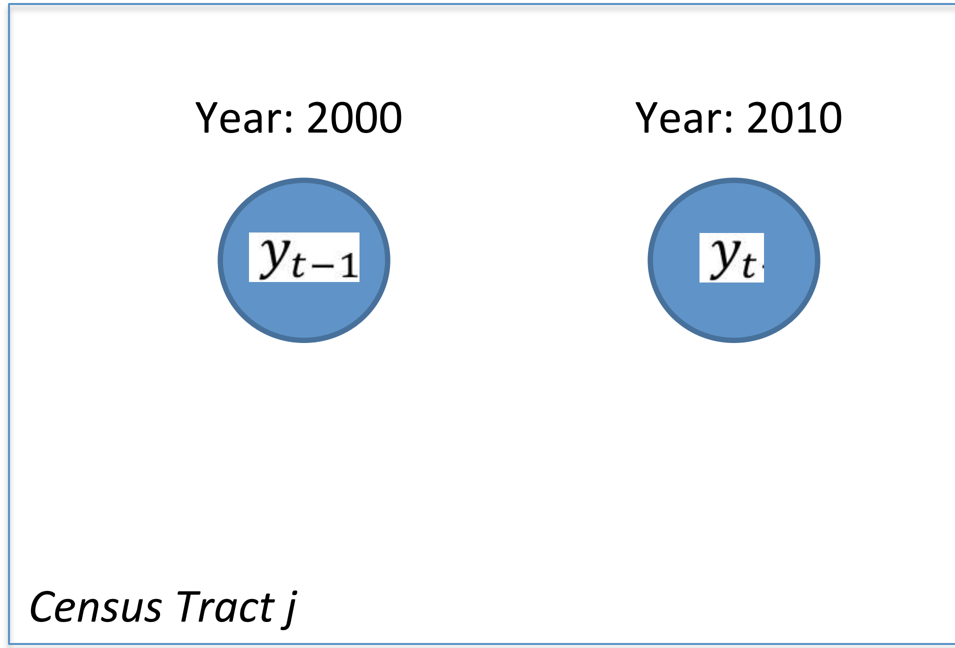
$$\nabla\%ForeignBorn\_Adjacent_{i,t_n} = \frac{\sum_j^{N(i)} \nabla\%\ Poverty_{j,t_n} * Population_{j,t_n}}{\sum_j^{N(i)} Population_{j,t_n}}$$

Conceptualizing the Model

Census Tract j

Conceptualizing the Model

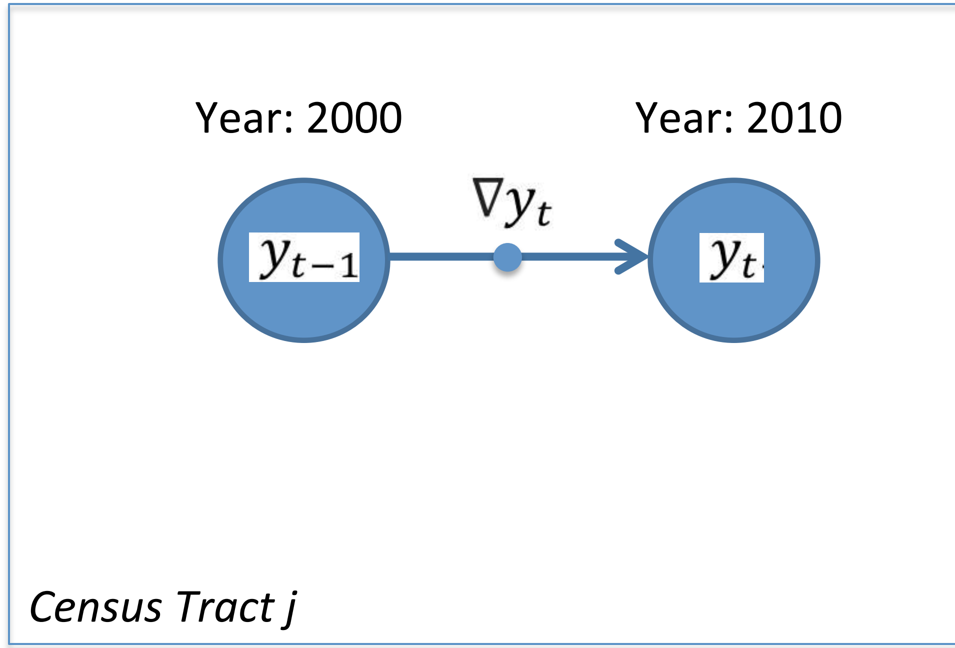
Total healthy food outlets: Y



Conceptualizing the Model

Total healthy food outlets: Y

Change in healthy food outlets: ∇Y

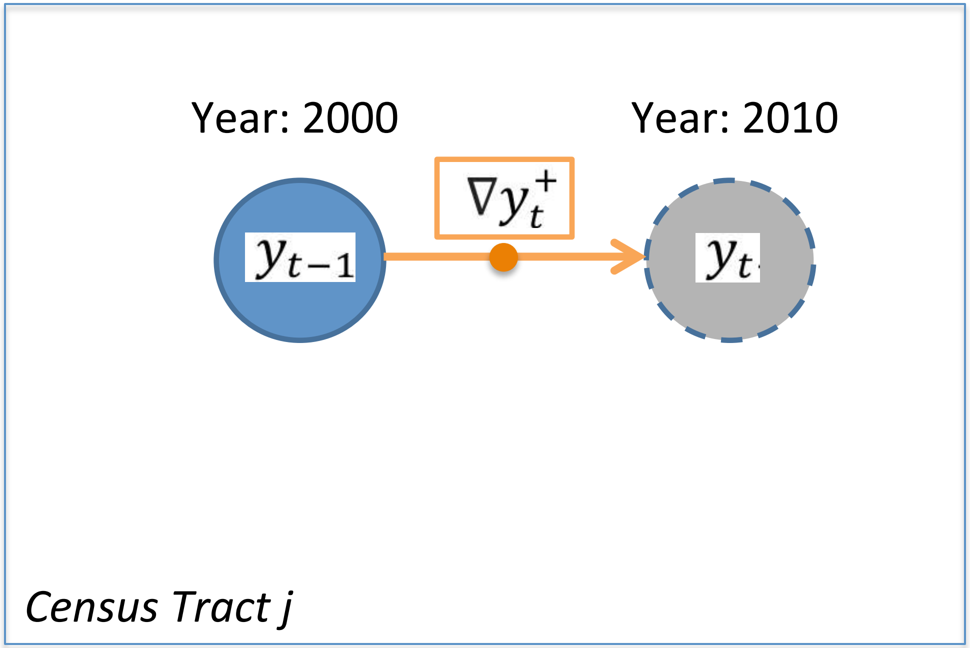


Conceptualizing the Model

Total healthy food outlets: Y

Change in healthy food outlets: ∇Y

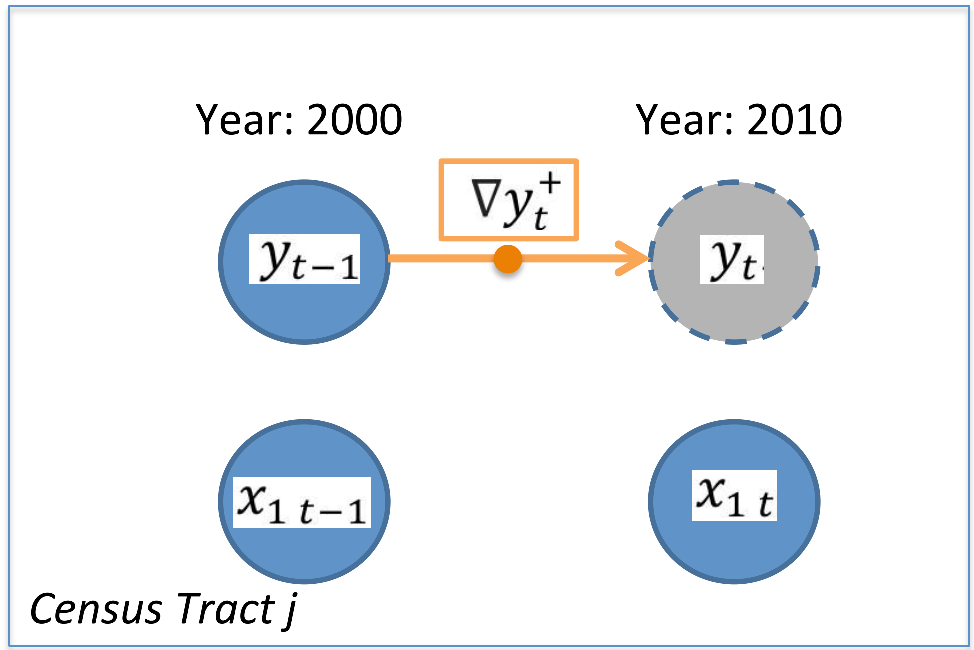
Derived indicator variable for whether positive change was observed



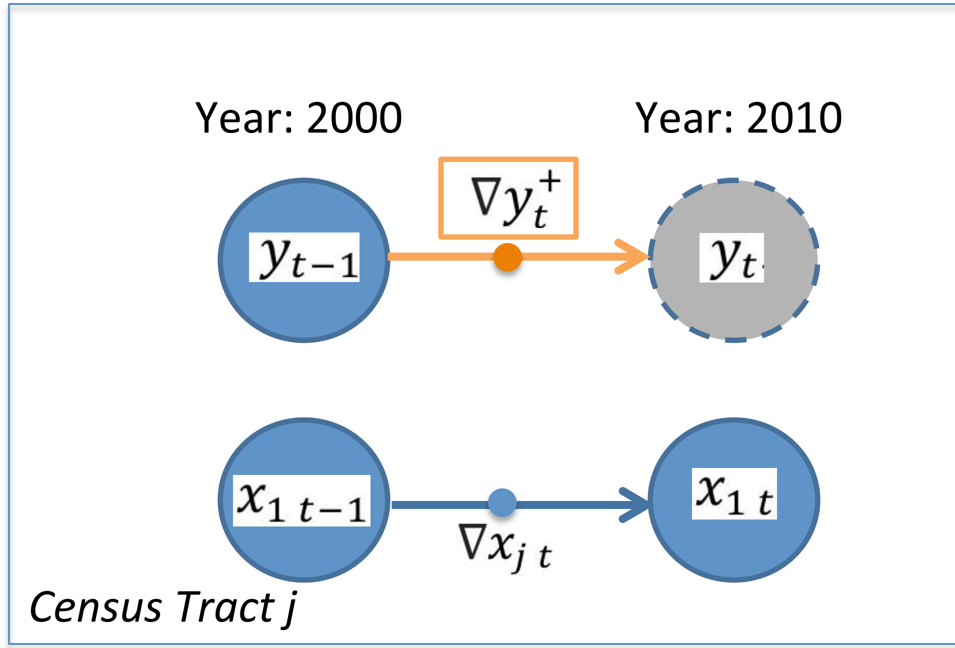
Conceptualizing the Model

Tract population characteristics by time: $x_{j\ t}$

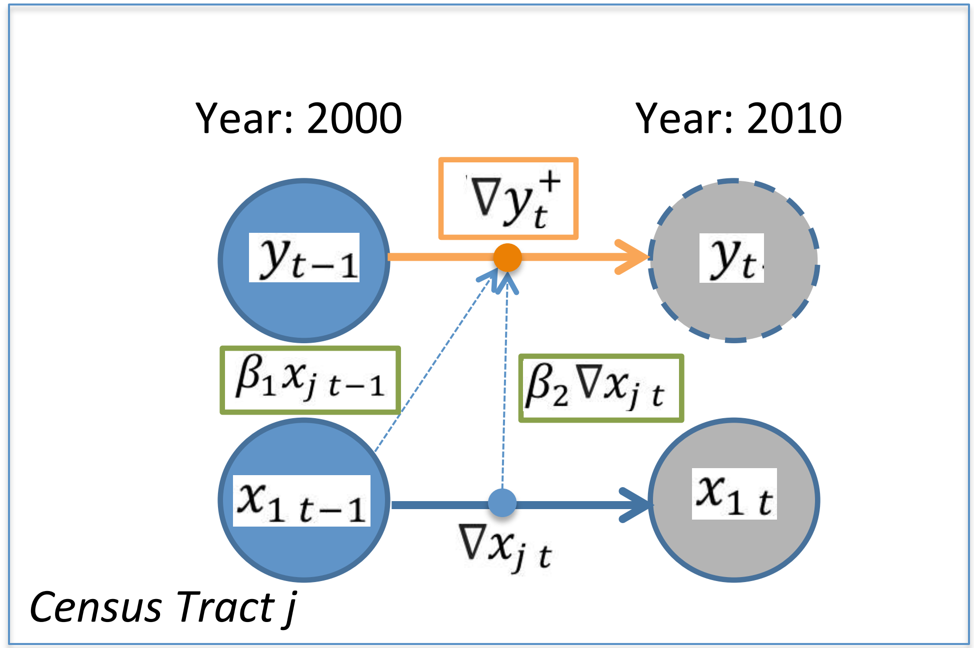
e.g., % Foreign Born



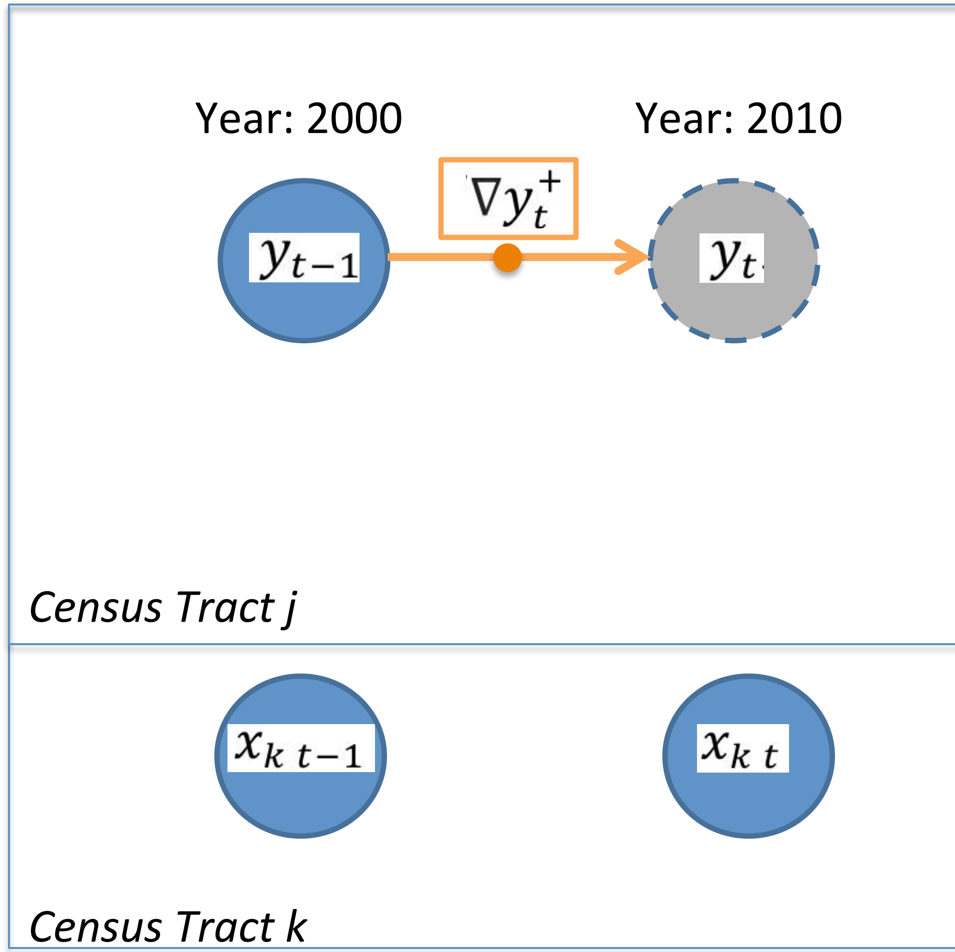
Conceptualizing the Model



Conceptualizing the Model



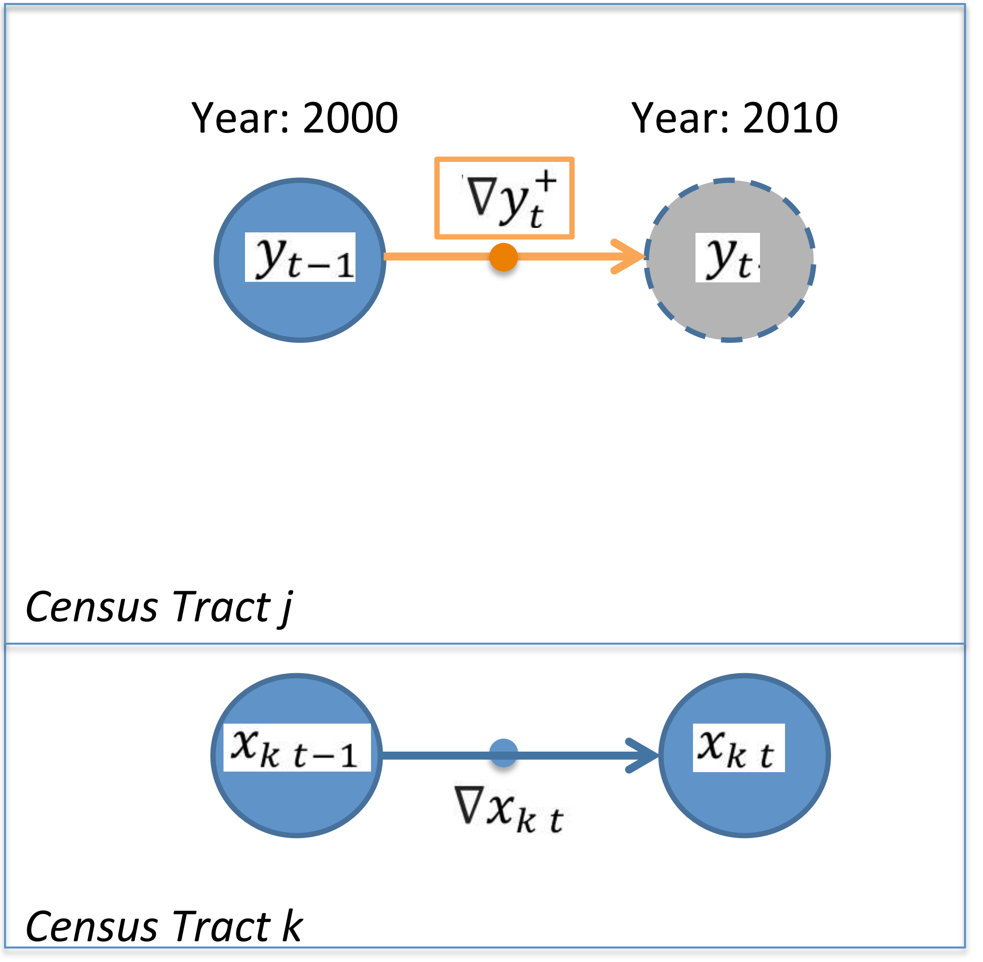
Conceptualizing the Model



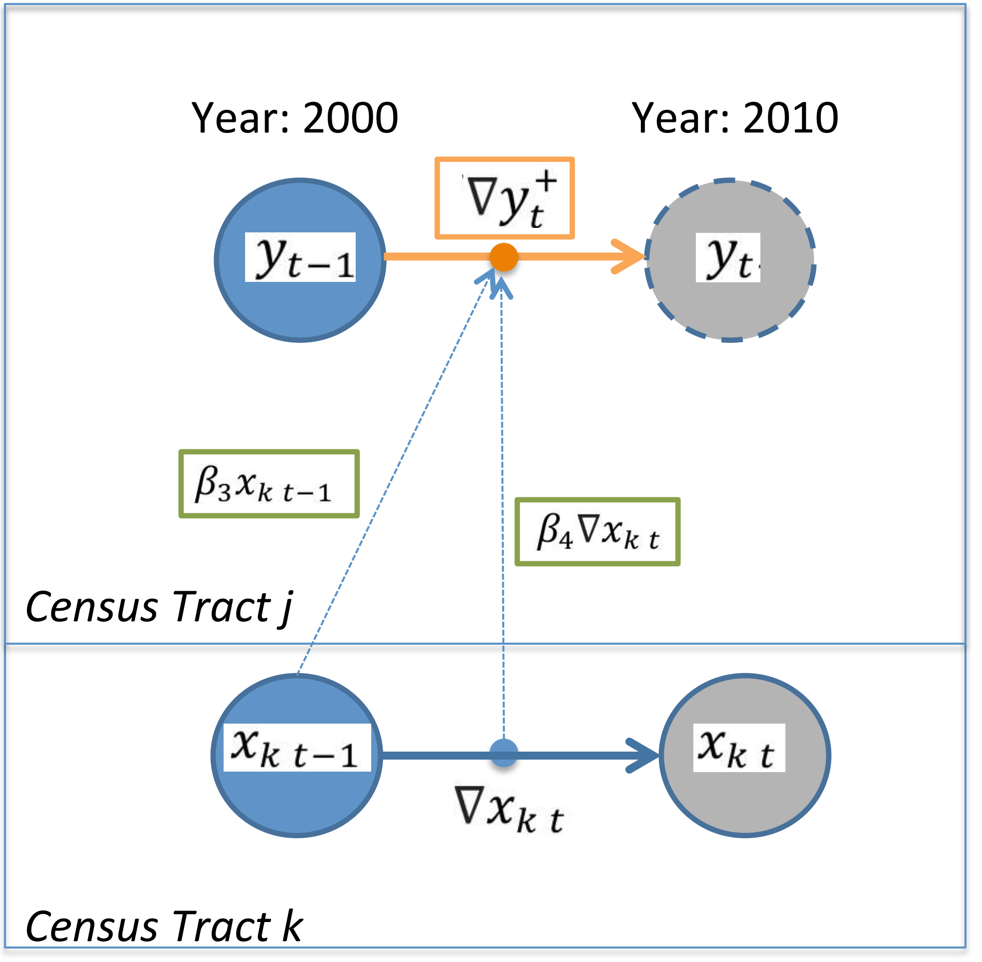
Adjacent tract population characteristics by time:

e.g., % Foreign Born

Conceptualizing the Model



Conceptualizing the Model



Expansions & Considerations

Interactions in Explanatory Variables:

38 main effects
703 interactions
9056 observations.

Risks:

- With many parameters there is risk of **overfitting relationships.**
- **Multicollinearity** can lead to **erratic estimates**. Reason to believe population characteristics will be correlated.
- High dimensional models suffer in terms of **interpretability.**
- Limitations of interpretations of estimate probability based on p-values (Gelman 2013).

Expansions & Considerations

Interactions in Explanatory Variables:

38 main effects
703 interactions
9056 observations.

Risks:

- With many parameters there is risk of **overfitting relationships.**
- **Multicollinearity** can lead to **erratic estimates**. Reason to believe population characteristics will be correlated.
- High dimensional models suffer in terms of **interpretability.**
- Limitations of interpretations of estimate probability based on p-values (Gelman 2013).

Approaches:

Bootstrapping – estimate consistency of estimates and prediction based on resampling of empirical distribution (Efron 1987).

Expansions & Considerations

Interactions in Explanatory Variables:
38 main effects
703 interactions
9056 observations.

Risks:

- With many parameters there is risk of **overfitting relationships.**
- **Multicollinearity** can lead to **erratic estimates**. Reason to believe population characteristics will be correlated.
- High dimensional models suffer in terms of **interpretability.**
- Limitations of interpretations of estimate probability based on p-values (Gelman 2013).

Approaches:
Bootstrapping – estimate consistency of estimates and prediction based on resampling of empirical distribution (Efron 1987).
Cross-Validation – select and evaluate models based on generalizability to withheld data.

Expansions & Considerations

Interactions in Explanatory Variables:
38 main effects
703 interactions
9056 observations.

Risks:

- With many parameters there is risk of **overfitting relationships.**
- **Multicollinearity** can lead to **erratic estimates**. Reason to believe population characteristics will be correlated.
- High dimensional models suffer in terms of **interpretability.**
- Limitations of interpretations of estimate probability based on p-values (Gelman 2013).

Approaches:
Bootstrapping – estimate consistency of estimates and prediction based on resampling of empirical distribution (Efron 1987).
Cross-Validation – select and evaluate models based on generalizability to withheld data.
Lasso Regularization – penalize coefficient estimates based on magnitude; selects subset of explanatory variables; can improve prediction. (Tibshirani 1996).

Expansions & Considerations

Interactions in Explanatory Variables:
38 main effects
703 interactions
9056 observations.

Risks:

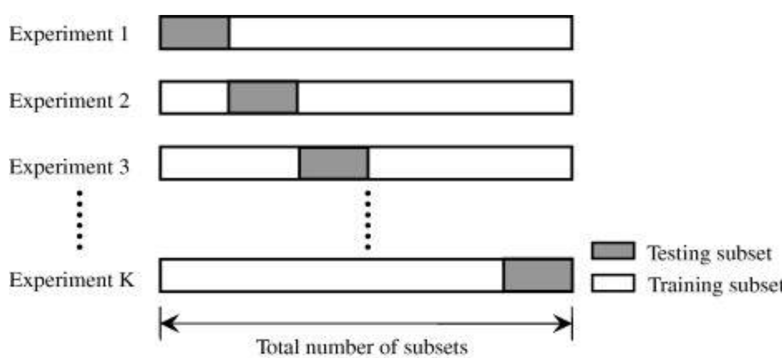
- With many parameters there is risk of **overfitting relationships.**
- **Multicollinearity** can lead to **erratic estimates**. Reason to believe population characteristics will be correlated.
- High dimensional models suffer in terms of **interpretability.**
- Limitations of interpretations of estimate probability based on p-values (Gelman 2013).

Approaches:
Bootstrapping – estimate consistency of estimates and prediction based on resampling of empirical distribution (Efron 1987).
Cross-Validation – select and evaluate models based on generalizability to withheld data.
Lasso Regularization – penalize coefficient estimates based on magnitude; selects subset of explanatory variables; can improve prediction. (Tibshirani 1996; Lim & Hastie 2013).
Model Averaging – combine many models estimated on resampled values and subsets. Can improve prediction and reduce variance (Breiman 1996; Hoeting et al., 1999).

Generalization & Model Validation

1: Partition (k-fold CV)

K-fold cross-validation (CV)

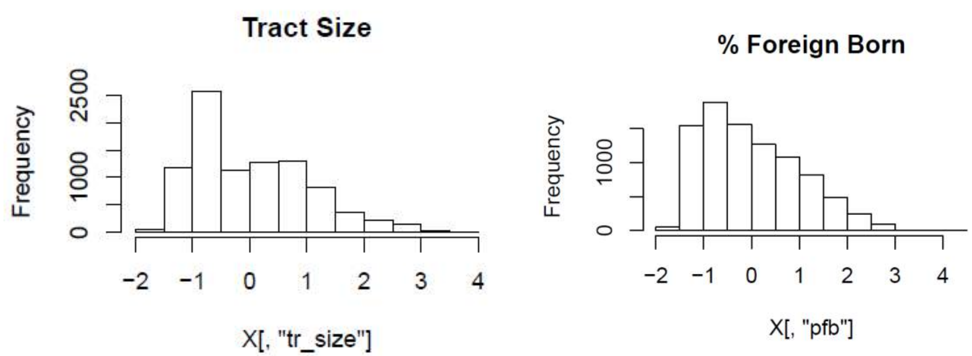


Generalization & Model Validation

1: Partition (k-fold CV)

2: Resample

Non-parametric bootstrap



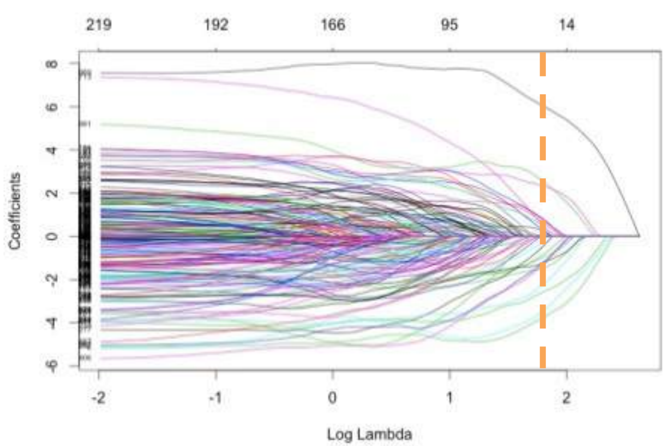
Generalization & Model Validation

1: Partition (k-fold CV)

2: Resample

3: Fit and Evaluate Models

Lasso Regularization



$$\hat{\beta}^{lasso} = argmin_{\beta}\left\{\boxed{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2} + \boxed{\lambda\sum_{j=1}^{p}|\beta_j|}\right\}$$

Generalization & Model Validation

1: Partition (k-fold CV)

2: Resample

3: Fit and Evaluate Models

4: Model averaging

Bayesian Model Averaging

$$\hat{\theta}_{\mathrm{BMA}} = \sum_{k=1}^{K} \hat{\theta}_k p\left(M_k \,|\, \boldsymbol{Z}\right)$$

Bootstrap Aggregation

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Generalization & Model Validation

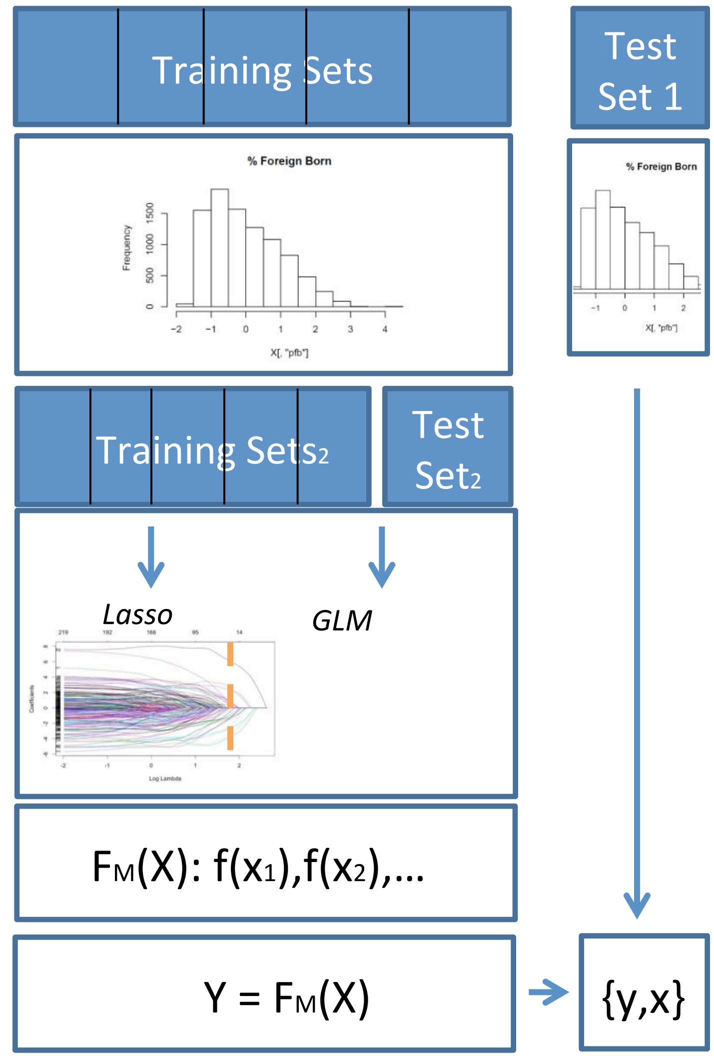
1: Partition (k-fold CV)

2: Resample

3: Fit and Evaluate Models

4: Model averaging

5: Evaluate prediction with resampled test set



Generalization & Model Validation

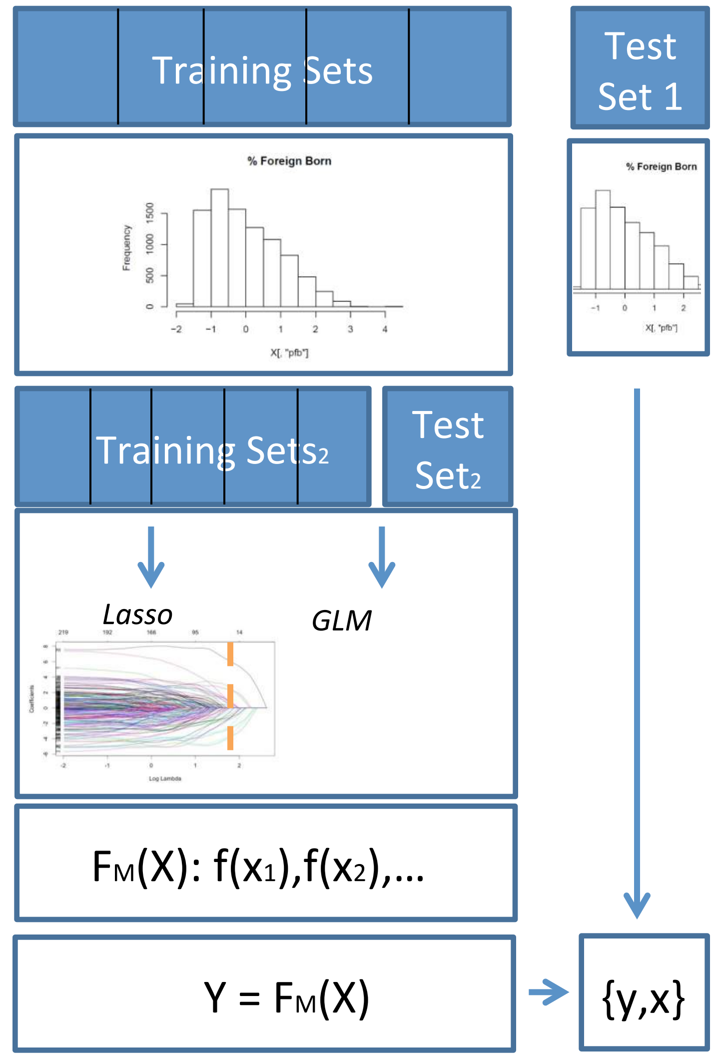
1: Partition (k-fold CV)

2: Resample

3: Fit and Evaluate Models

4: Model averaging

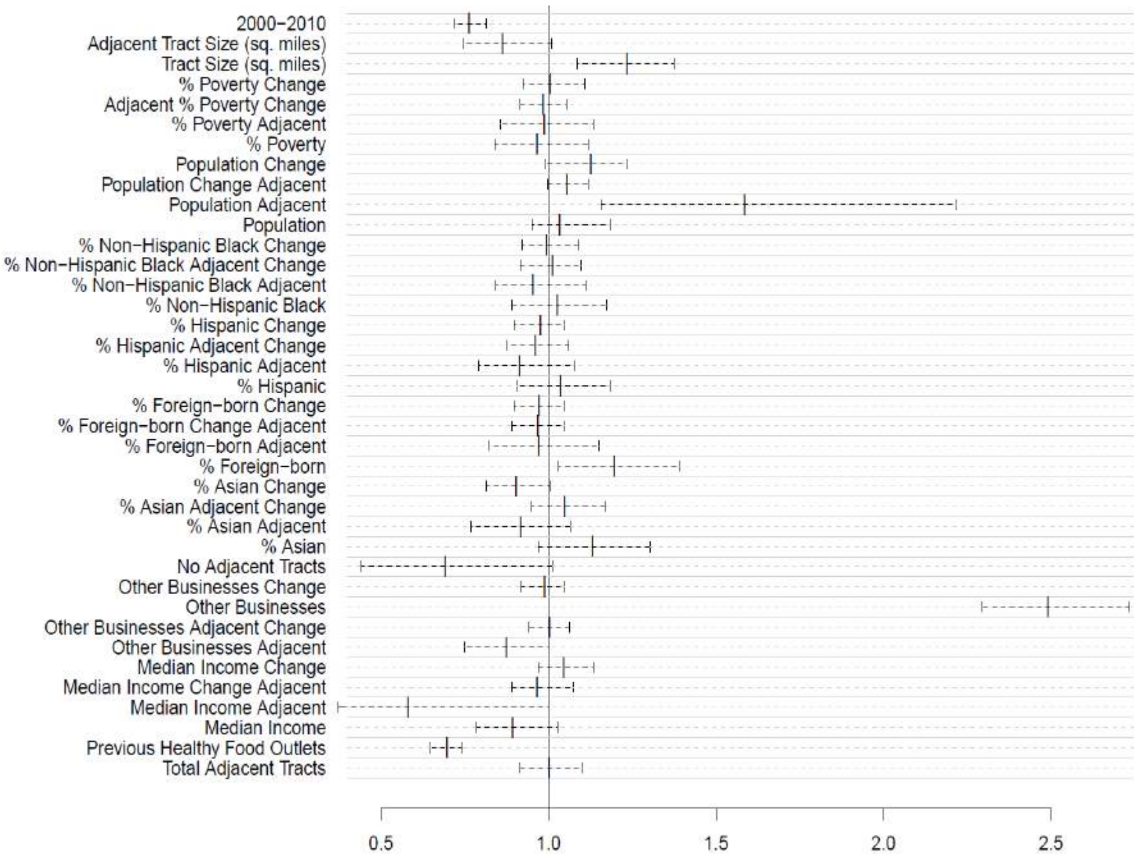
5: Evaluate prediction with resampled test set



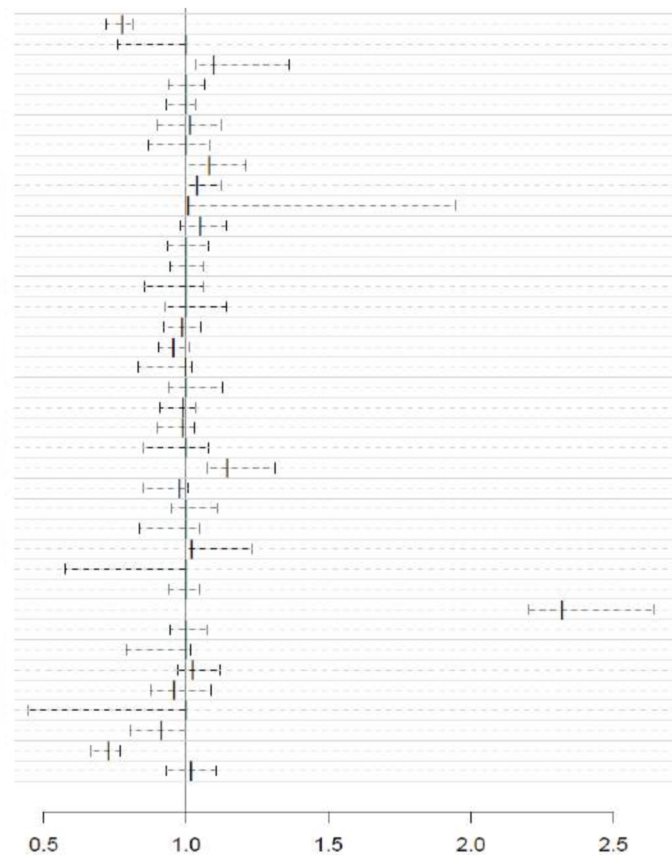
Bootstrap Prediction:	<u>Lasso</u>		<u>GLM</u>	
	<u>Bag</u>	<u>BMA</u>	<u>Bag</u>	<u>BMA</u>
AIC (sum)	19469	19926	19881	19476
Deviance (mean):	961	964	937	939
Misclassified:	24.1%	23.6%	23.9%	23.9%

Bootstrap Confidence Intervals

GLM Logistic Odds Ratios & 95%-CI

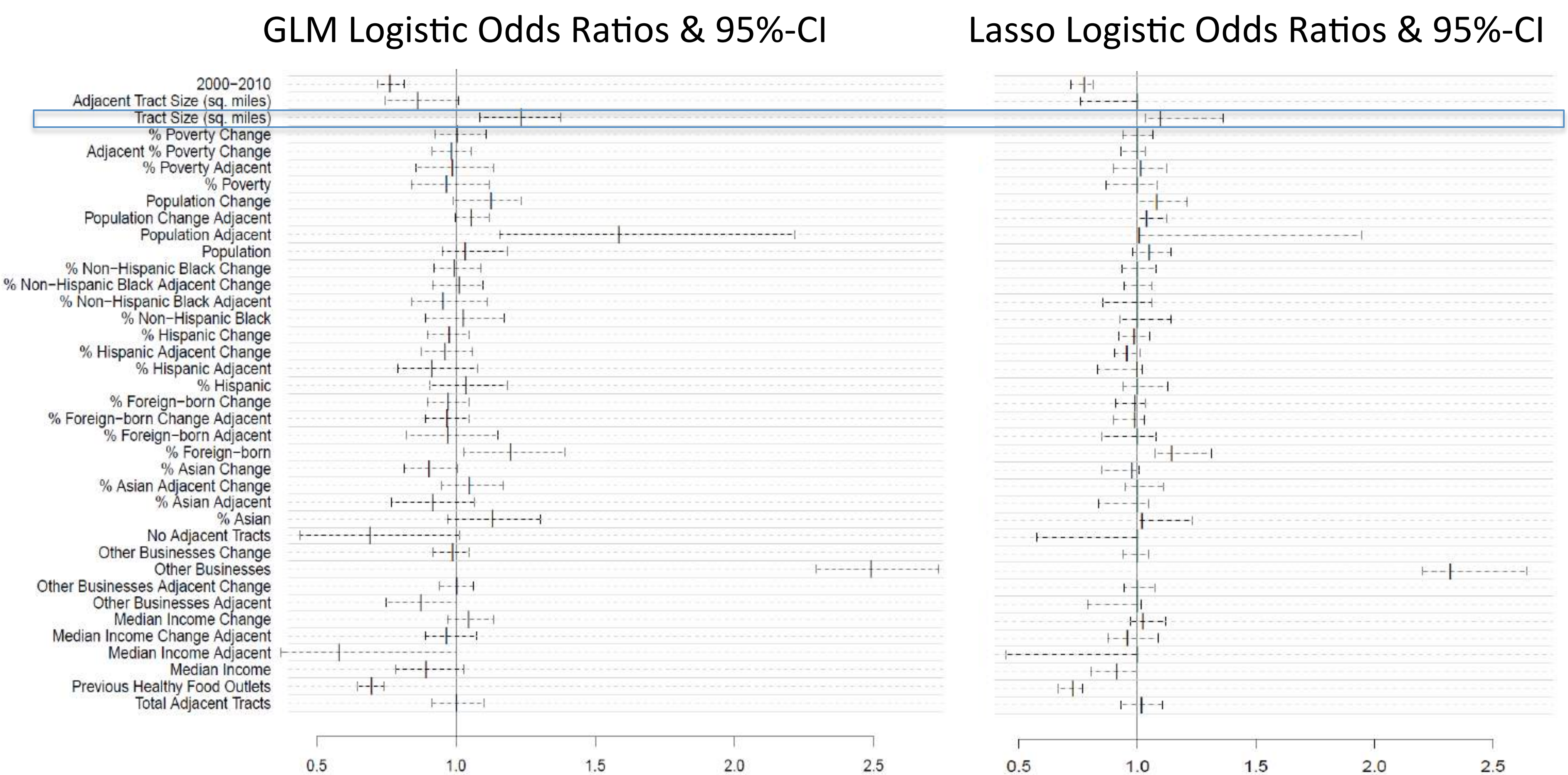


Lasso Logistic Odds Ratios & 95%-CI



COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

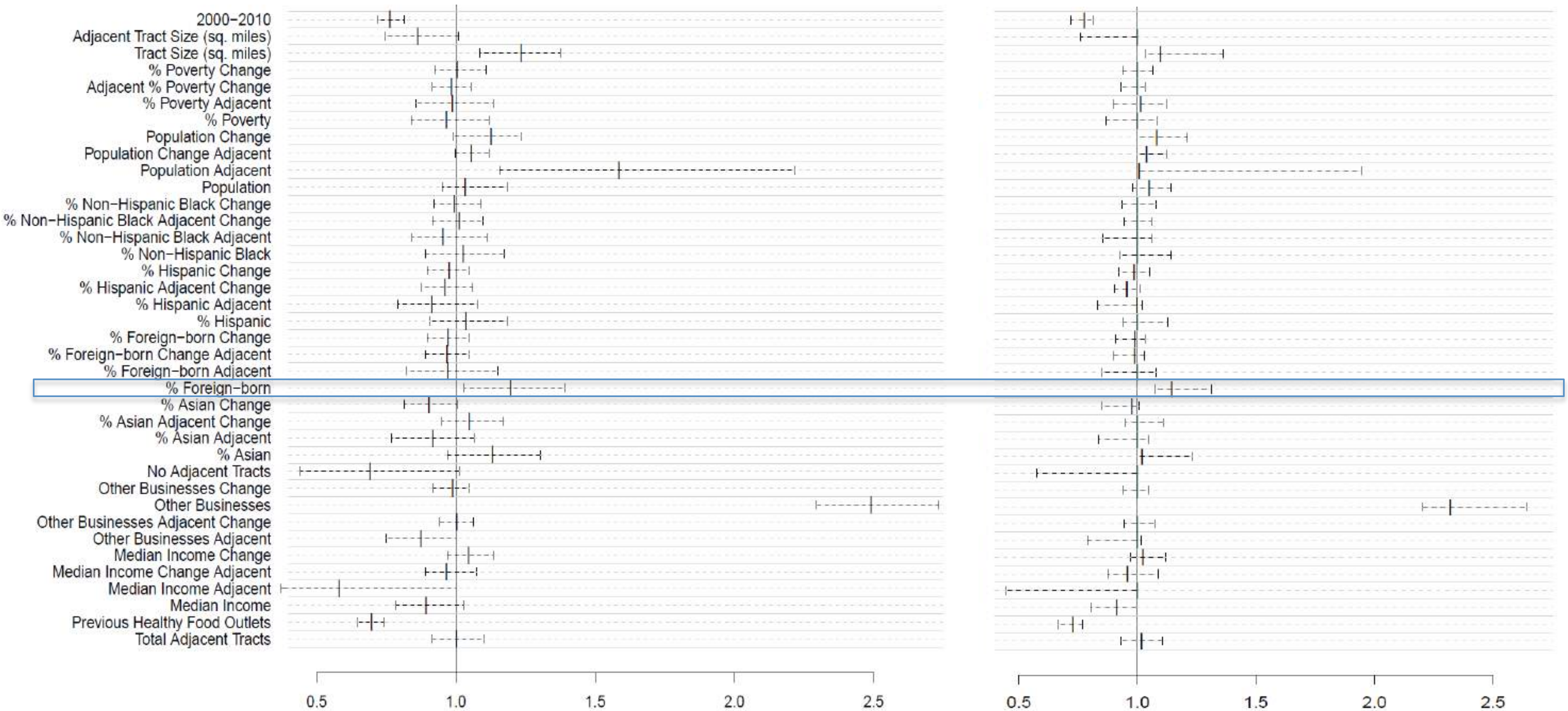
Bootstrap Confidence Intervals



Bootstrap Confidence Intervals

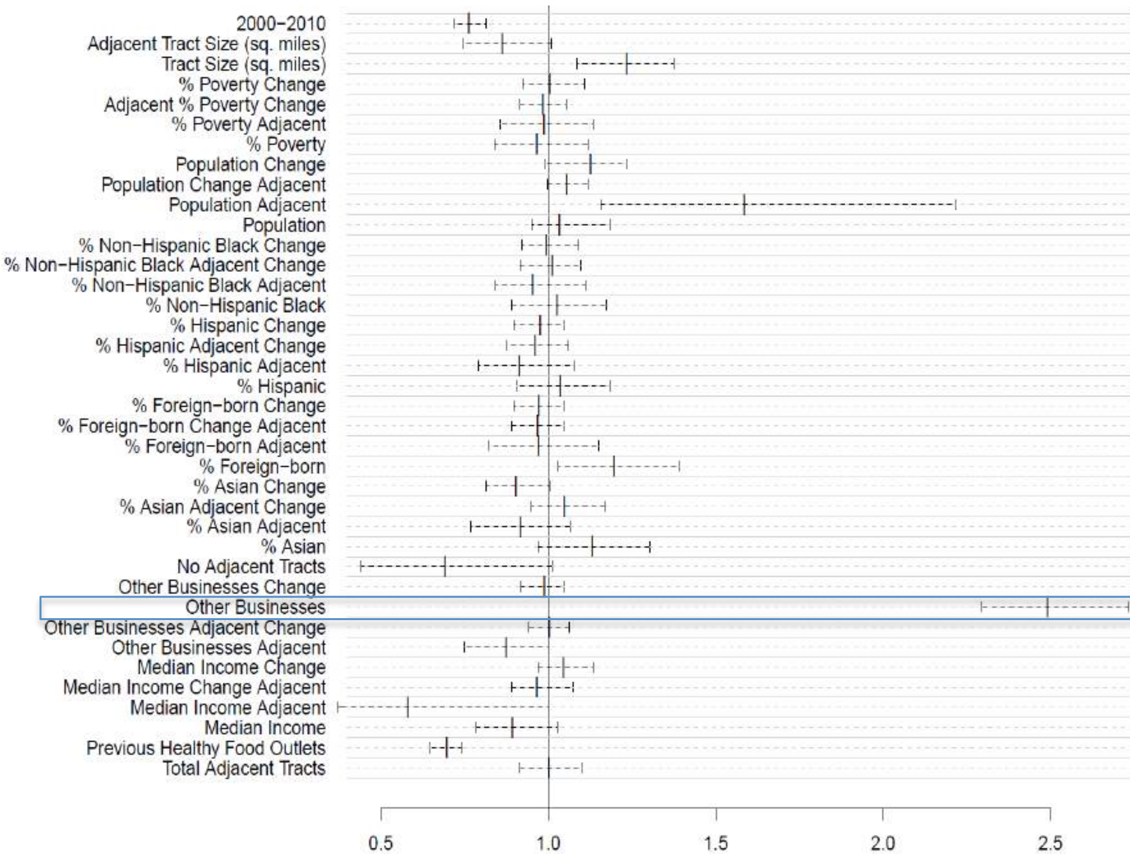
GLM Logistic Odds Ratios & 95%-CI

Lasso Logistic Odds Ratios & 95%-CI

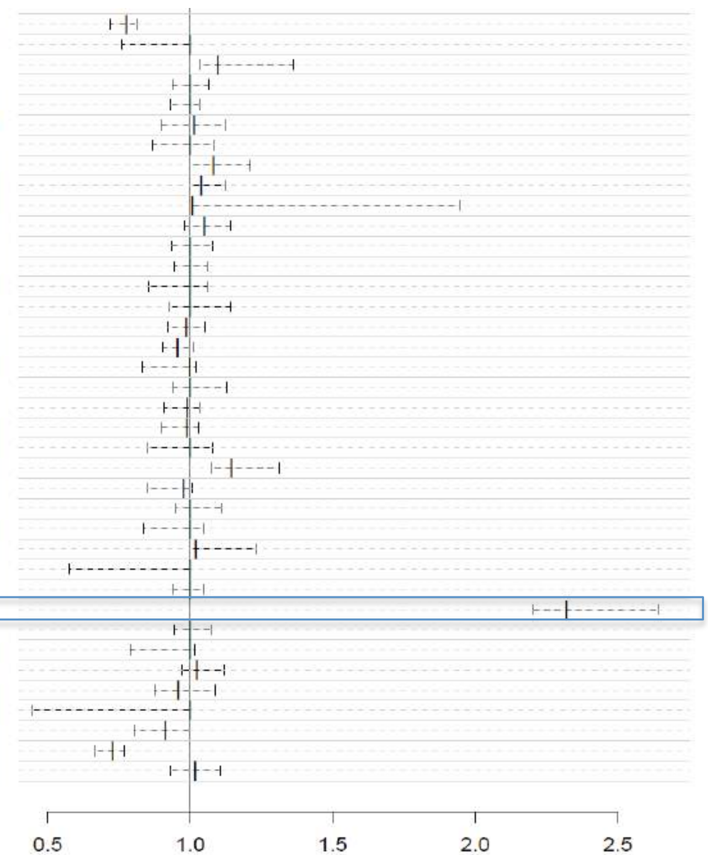


Bootstrap Confidence Intervals

GLM Logistic Odds Ratios & 95%-CI



Lasso Logistic Odds Ratios & 95%-CI

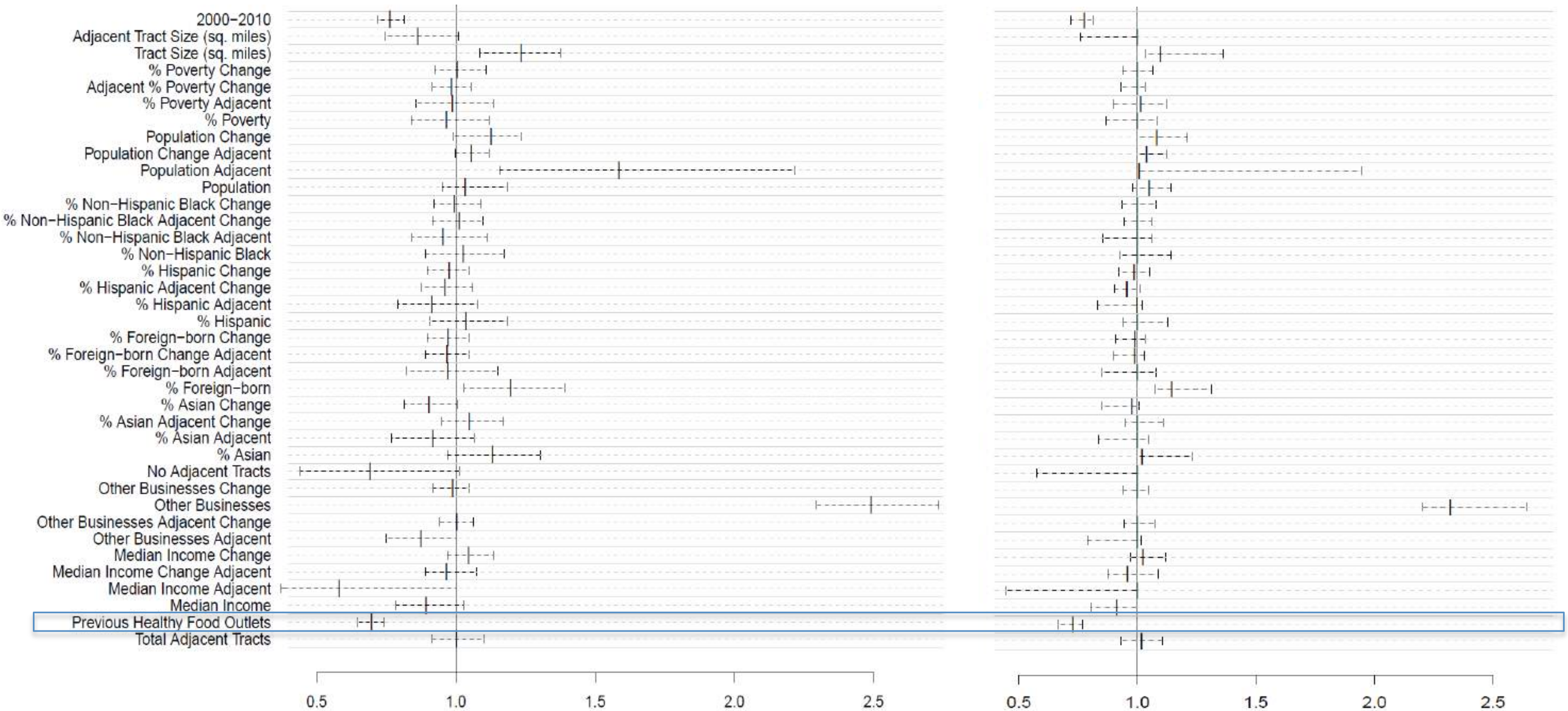


COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Bootstrap Confidence Intervals

GLM Logistic Odds Ratios & 95%-CI

Lasso Logistic Odds Ratios & 95%-CI



Conclusions

- Local characteristics including foreign born population, total businesses and tract size show associations with positive change in healthy food outlets.
- Greater number of healthy food outlets in the previous decade was associated with lower odds of increase.
- Longitudinal and spatial demographic data can be used to develop models that examine change across time and space.
- Averaging of models fit to bootstrap samples using lasso penalization, with weights based on relative likelihood, improved prediction in resampled test-set data.

Future Directions

- Alternate approaches to penalization and variable selection in interactions models – less naive approaches to this case of hierarchical modeling.
- Specify as fully Bayesian approach with evaluation of prediction (e.g., WAIC).
- Carry out simulations to examine the role of correlation and noise in prediction given relationships in longitudinal and spatial data in context of model validation.
- Additional outcomes: different business types and continuous measures of change.
- Expand inputs; additional transformations, basis expansions, variable selection steps.
- Vary and extend time intervals included.
- Add additional localities to further assess generalizability.

References

Breiman L. Bagging Predictors. Machine Learning. 1996. 24(2):123-140.

Cummins S, Macintyre S. Food environments and obesity—neighbourhood or nation? International journal of epidemiology. 2006. 35(1):100-4.

Efron B. Better Bootstrap Confidence Intervals. Journal of the American Statistical Association. 1987. 82(397):171-185.

Gelman A. P values and statistical practice. Epidemiology. 2013. 24(1):69-72.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian Model Averaging: A Tutorial. 1999. Statistical Science. 14(4):382-401.

Kaufman TK, Sheehan DM, Rundle A, Neckerman KM, Bader MD, Jack D, Lovasi GS. Measuring Health-Relevant Businesses Using the National Establishment Time-Series (NETS) Database: A Dynamic Longitudinal Assessment Over 21 Years.

Lim M, Hastie T. Learning interactions through hierarchical group-lasso regularization. 2013. arXiv preprint arXiv:13082719.

Lovasi GS, Hutson MA, Guerra M, Neckerman KM. Built environments and obesity in disadvantaged populations. Epidemiological Review. 2009. 3(1): 7-20.

Morland K, Wing S, Diez Roux A, Poole C. Neighborhood characteristics associated with the location of food stores and food service places. American journal of preventive medicine. 2002; 22(1):23-9.

Moore LV, Roux AVD, Nettleton JA, Jacobs DR. Associations of the Local Food Environment with Diet Quality—A Comparison of Assessments based on Surveys and Geographic Information Systems The Multi-Ethnic Study of Atherosclerosis. American journal of epidemiology. 2008; 167(8):917-24.

Powell LM, Slater S, Mirtcheva D, Bao Y, Chaloupka FJ. Food store availability and neighborhood characteristics in the United States. Preventive medicine. 2007;44(3):189-95.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B. 1996. 58(1): 267-288.

Thank you

Acknowledgements:

Eunice Kennedy Shriver NICHD
National Institute of Child Health & Human Development

Columbia University
Built Environment & Health Project
www.beh.columbia.edu

Software

- R
  - boot, parallel, reshape2, stringr, mice, maptools, PCIT
  - glmnet, glinternet
  - qplot

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Interactions & Penalization

Subset Selection:

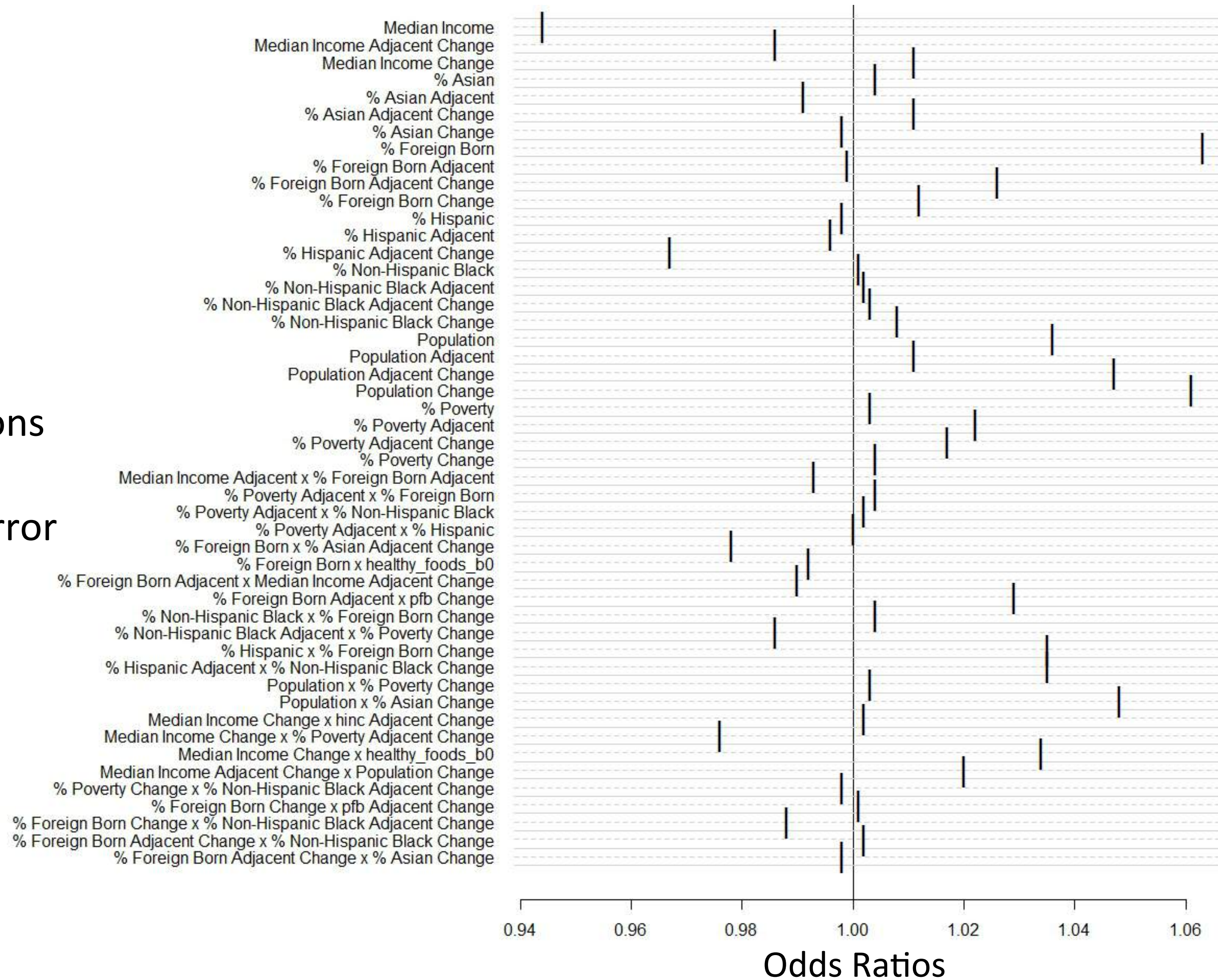
38 main effects
703 interactions

reduced to:

34 Main Effects
48 Two-Way Interactions

Based on 10-fold CV error minimization

Demographic Explanatory Variables



Interactions & Penalization

<u>Interactions Lasso:</u>
AIC: 9531
BIC: 10115
Deviance: 9367

<u>Main Effects Lasso:</u>
AIC: 9323
BIC: 9479
Deviance: 9272

<u>GLM:</u>
AIC: 9254
BIC: 9531
Deviance: 9176