Data Smearing:
An Approach to Disclosure Limitation for Tabular Data

Daniell Toth1

1- U.S. Bureau of Labor Statistics

Content represents the opinion of the authors only.

Talk Outline

The Quarterly Census of Employment and Wages (QCEW)

- Current Disclosure Limitation Method
- Wants and Needs

Synthetic Data Approachs

- Use in Disclosure Limitation
- Synthetic Data to Produce Tables

Data Smearing

- The Method
- Application to QCEW data

QCEW

Census of Establishments

- All establishments that pay Unemployment Insurance

Total monthly employment and quarterly wages

Produces quarterly tables by industry and area

NAICS	e20101	e20102	e20103	e20104	total
Series 1	2600	2899	3022	2599	11120
Sub1	1981	2256	2382	1957	8576
Sub2	32	33	37	33	135
Sub3	587	610	603	609	2409

QCEW Cell Suppression

For certain industries – few establishments in any given area = suppression

NAICS	e20101	e20102	e20103	e20104	total
Series 1	2600	2899	3022	2599	11120
Sub1	1981	2256	2382	1957	8576
Sub2		33	37	33	135
Sub3	587	610	603	609	2409

---- Primary suppression

Secondary Suppressions

The ability to publish aggregates = many secondary suppressions

NAICS	e20101	e20102	e20103	e20104	total
Series 1	2600	2899	3022	2599	11120
Sub1	1981	2256	2382	1957	8576
Sub2		33		33	135
Sub3		610		609	2409

---- Primary suppression

---- Secondary suppression

Problems and Desires

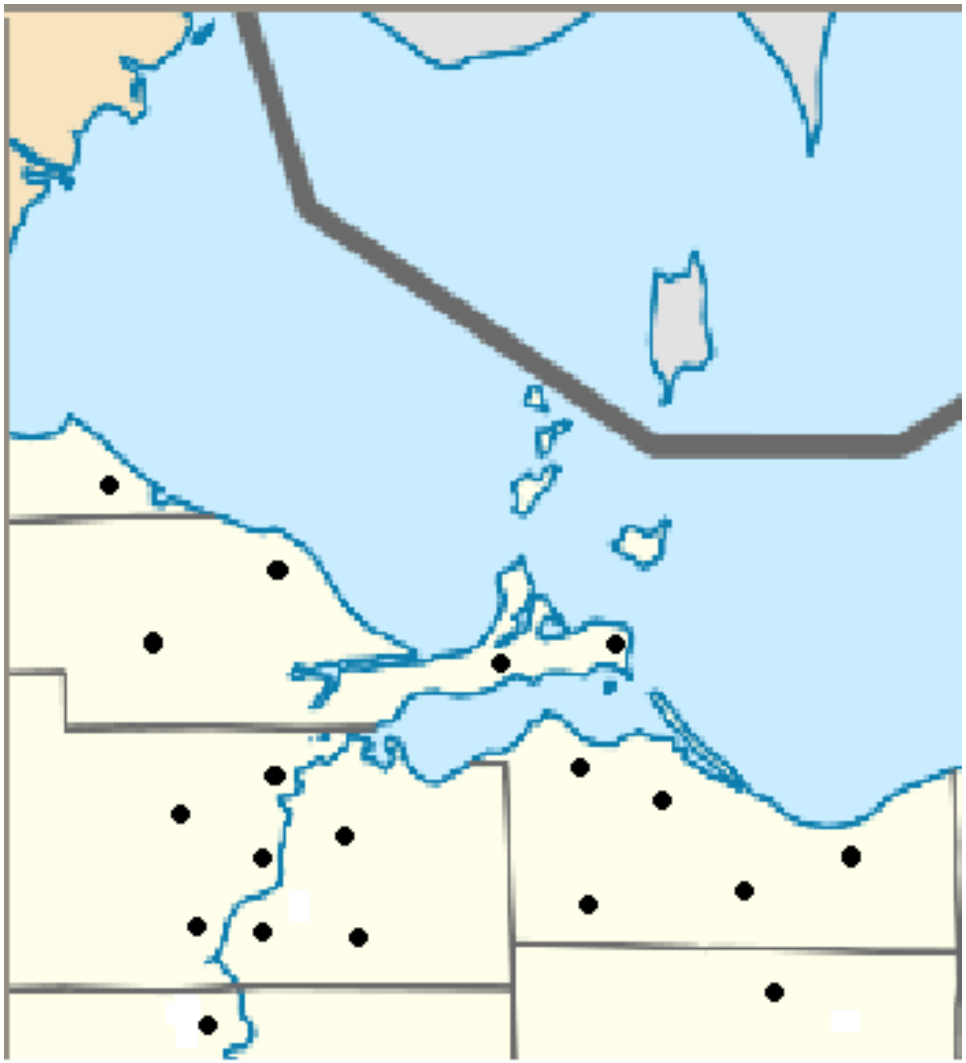
Problems:

- Way too many suppressions
- Coordinating secondary suppressions w. states
- May not protect data
  - Holan, S., Toth, D., Ferreira, M., Karr, A. (2010)

Desires:

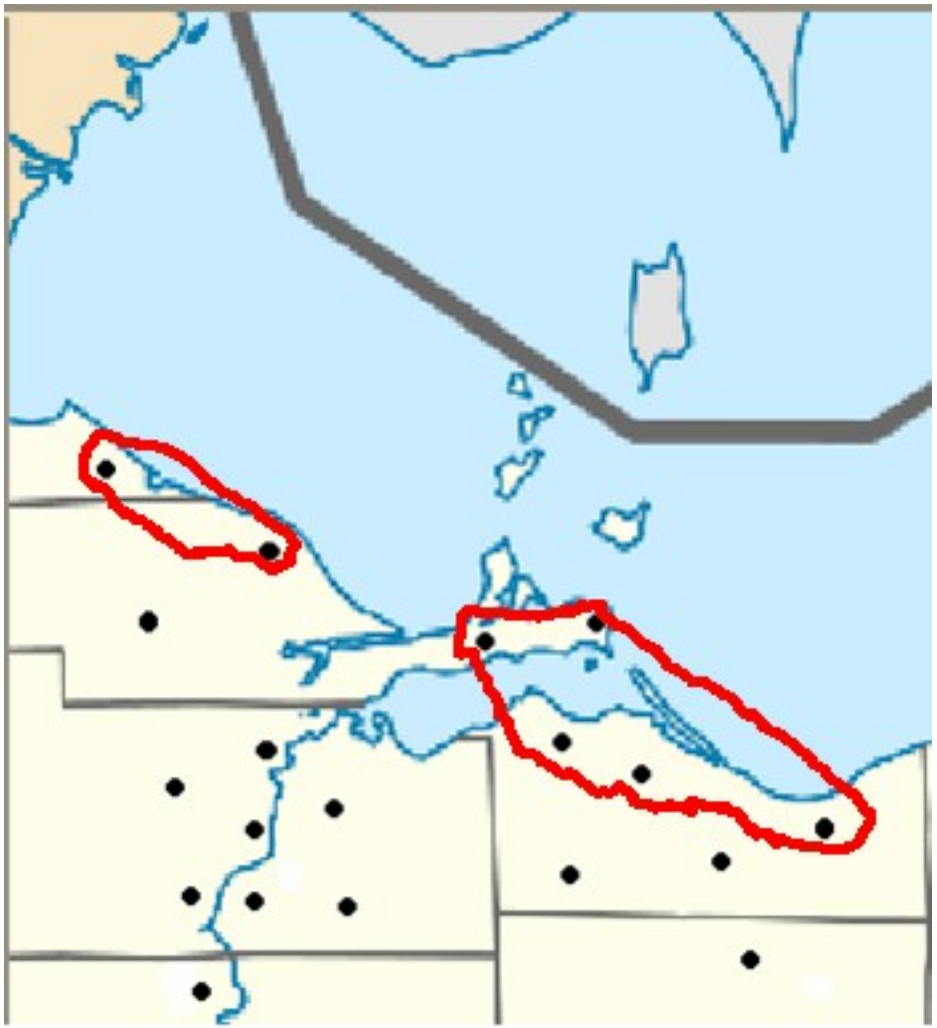
+ Less (read no) suppressions
+ Accurate high level aggregated cells
+ Produce any requested table

Limited Service Restaurants



Note: This example is completely fabricated. For illustration purposes only.

Region: Along the Lake



Note: This example is completely fabricated. For illustration purposes only.

Synthetic Data Approach

1.) Use sampled data to fit

$$f(\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_p)$$

replace y with

$$\tilde{y} \sim f(y \mid x_1, x_2, \ldots x_p)$$

$$\tilde{y}_i = y_i + \epsilon_i$$

2.) F

$$E[\epsilon_i] = 0 \text{ and } E[\epsilon_i \epsilon_j] = 0 \; i \neq j$$

Difficulties of Synthetic Data Approach

1.) Use sampled data to fit

$$f(\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_p)$$

replace y with

$$\tilde{y} \sim f(y \mid x_1, x_2, \ldots x_p)$$

QCEW is a census values can be too accurate

2.) Replace v with

$$\tilde{y}_i = y_i + \epsilon_i$$

$$E[\epsilon_i] = 0 \text{ and } E[\epsilon_i \epsilon_j] = 0 \; i \neq j$$

QCEW is highly skewed hard to choose noise factor

Data Smearing

1. Define distance between units
   based on desired domains

2. Find nearest-network of each unit

3. Synthetic value of each unit is an average of values from units in the network.

- Each unit represents an average of itself and the surrounding units

The Distance

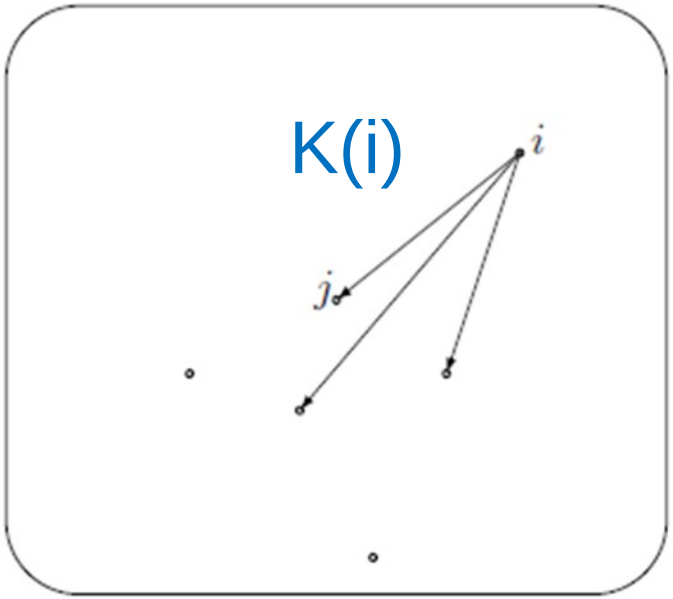
Define a distance between units

$$d(\mathbf{u}_i, \mathbf{u}_j) = ||\mathbf{u}_i - \mathbf{u}_j||$$

Example on QCEW data :

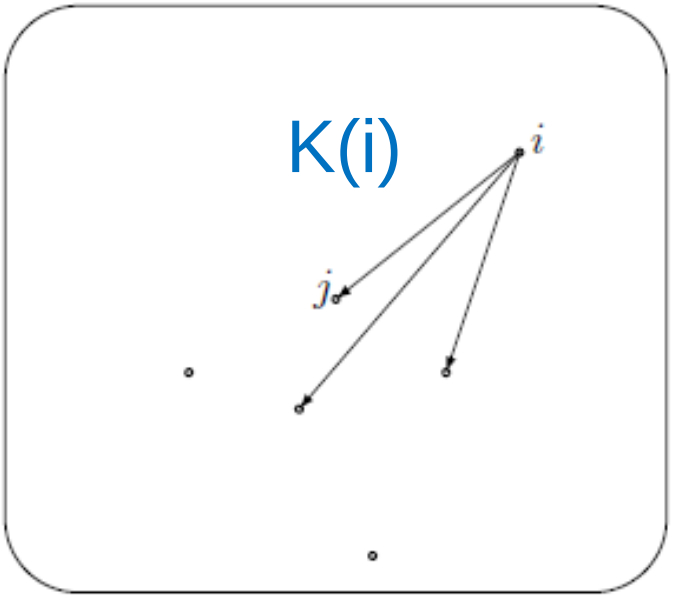
$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu \mathbb{1}_{\{ind6_i \neq ind6_j\}}$$

k-Nearest Neighbor



k = 3

Isolated Units



k = 3

j in K(i), but unit i is no other unit's neighborhood

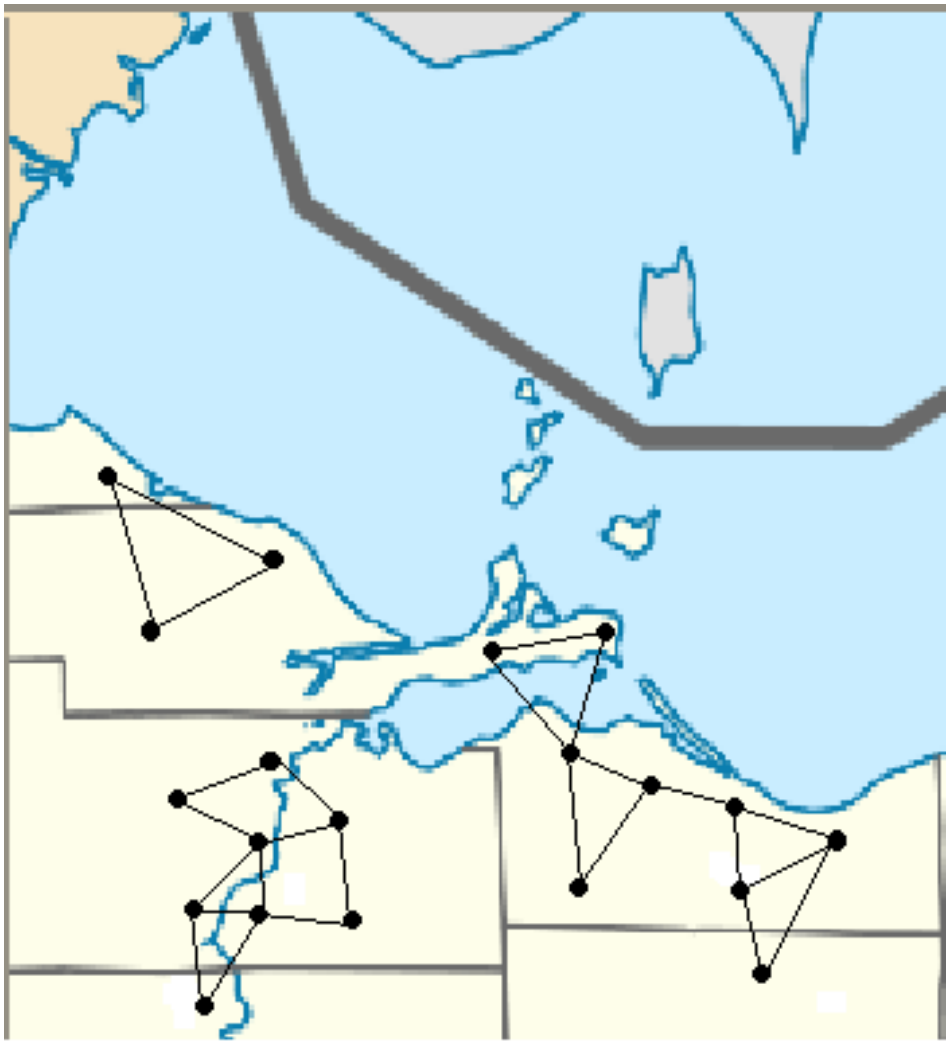
k-network

To make sure each unit gets spread out among other units we expand the k-nearest neighborhood

$$\overline{K(i)} = K(i) \cup \left\{ j \;\middle|\; i \in K(j) \right\}$$

If unit i is in j's neighborhood than j is in i's neighborhood

Nearest Network: k=2



Note: This example is completely fabricated. For illustration purposes only.

Sample from Network

Draw SRS without replacement of size n ≤ k from $\overline{K(i)}$

Selected units will be used to produce synthetic values.

- harder to identify which units are used to produce synthetic values

Let $\delta_j(i)$ be the indicator function (=1 if unit j is selected).

The Synthetic Values

Replace micro-data values with

$$\tilde{\mathbf{Y}}_i = w_i \mathbf{Y}_i + \sum_{j \in \overline{K(i)}} w_j \delta_j(i) \mathbf{Y}_j$$

$\delta_j(i)$ Sample indicator function for unit j

w_i Weights to compute average

This can be repeated to produce multiple tables.

Closed Areas

Any subset of population $C \subseteq U$ is a “closed area” if

$$C = \bigcup_{i \in C} \overline{K(i)}$$

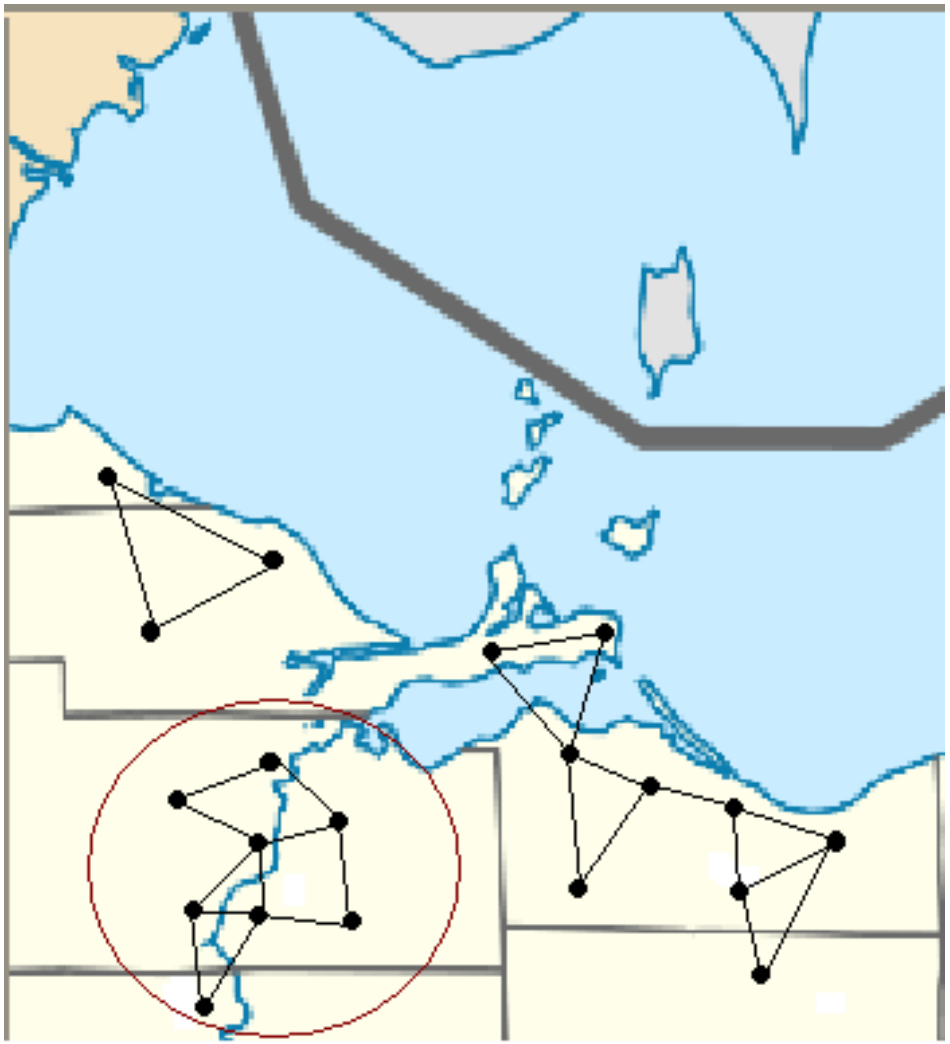
Closed areas contain all units contributing data to the estimate.

-easy to evaluate properties for these areas

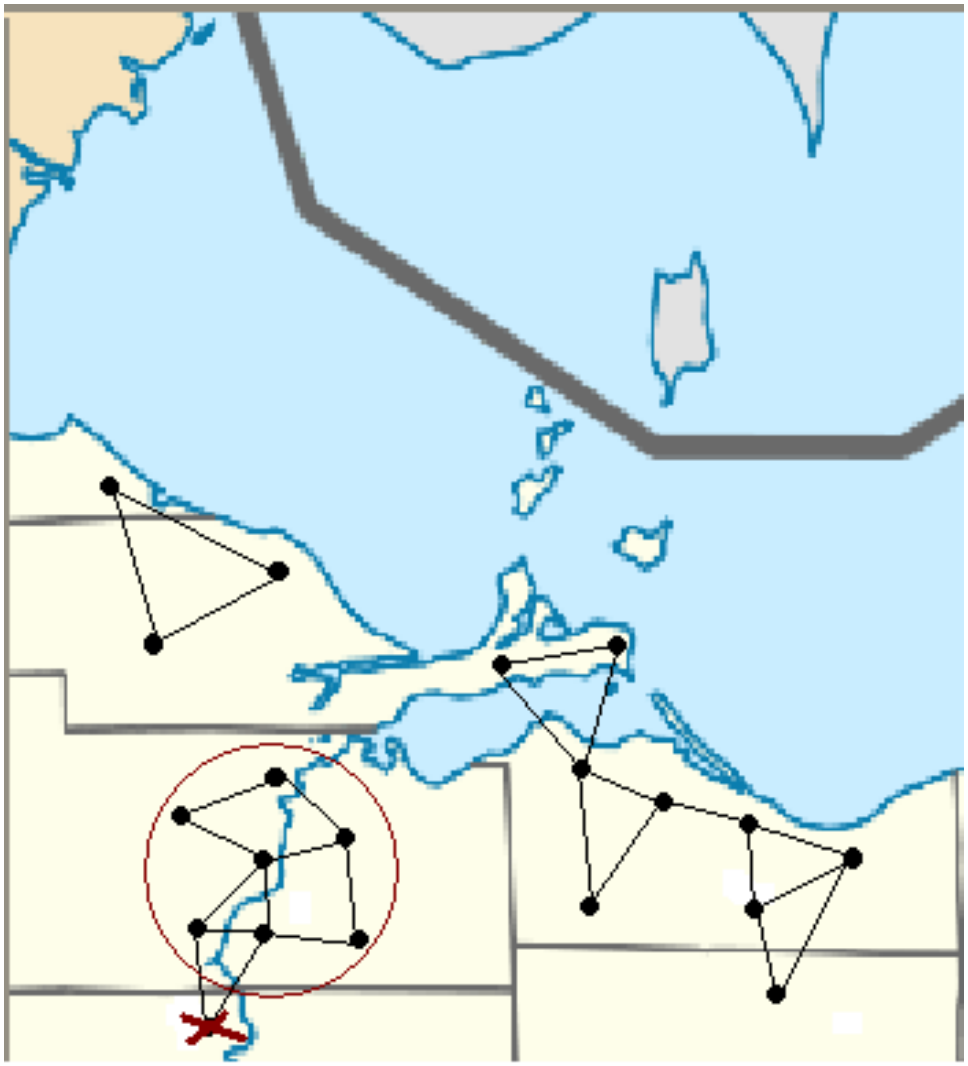
For any subset C, there exists a closed area that contains C.

Denote: $\overline{C}$ as the smallest closed area that contains C.

Area of Interest is Closed



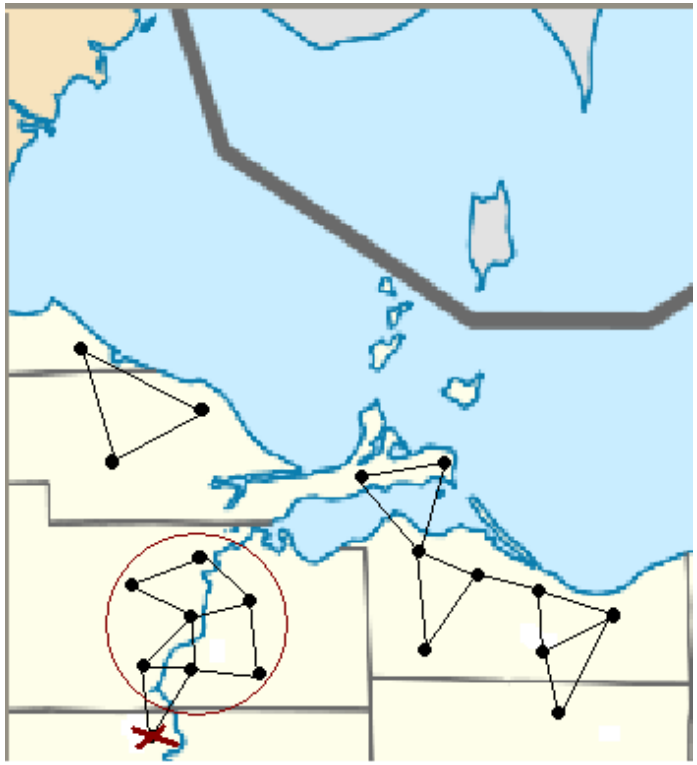
Smaller Area: Not Closed



Note: This example is completely fabricated. For illustration purposes only.

An Area's Boundary

Define the boundary of area C



as the set $\partial(C) = \overline{C} - C.$

Correct Cell Estimates "On Average"

Lemma 2.1 *If a cell C is a closed area*

and $$w_i = \left(1 + n \sum_{j \in \overline{K(i)}} 1/\left|\overline{K(j)}\right|\right)^{-1},$$

then

$$E\left[\sum_{i \in C} \tilde{\mathbf{Y}}_i\right] = \sum_{i \in C} \mathbf{Y}_i.$$

Consistent Estimates

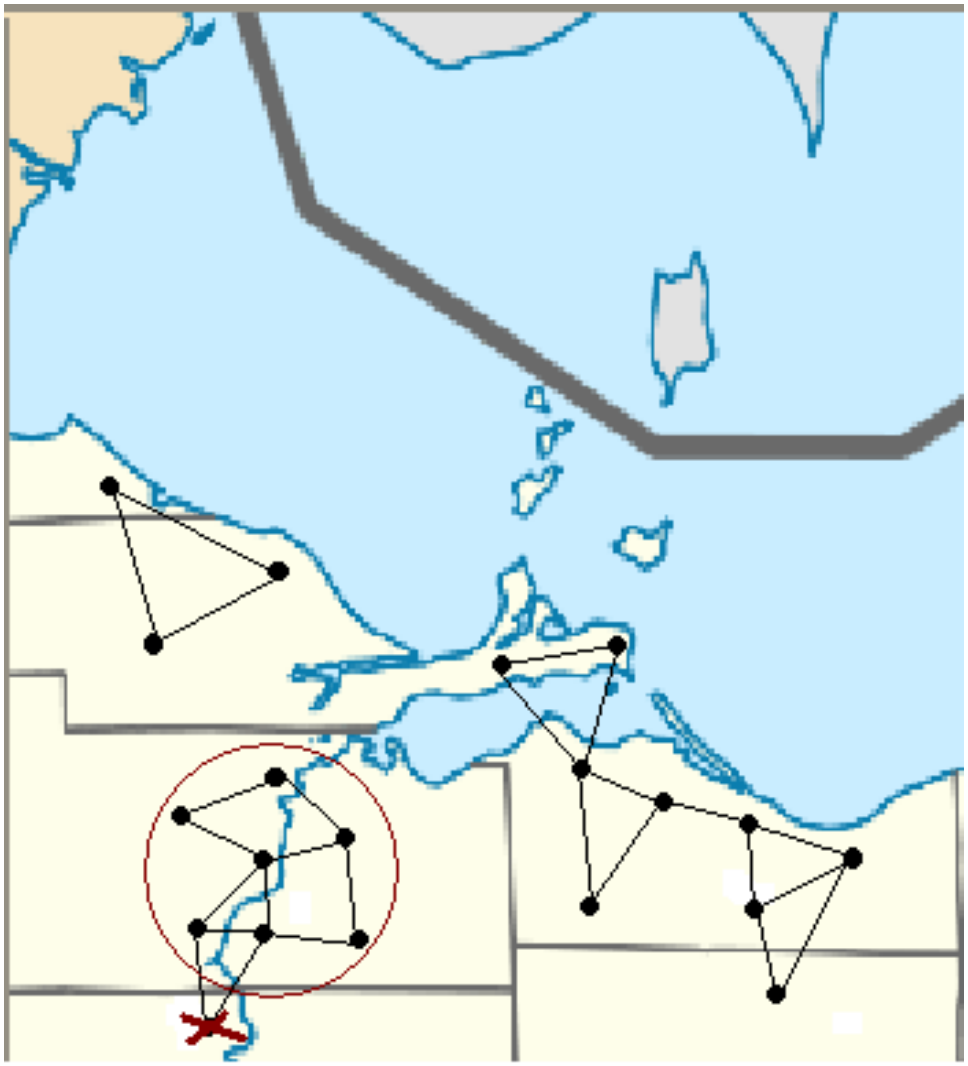
Property 2.1 *Assume* $|y_i - E[\tilde{y}_i]| < M < \infty$ *for all* i.

$$If\ |\partial(C)| = o\left(\sum_{i \in C} \boldsymbol{Y}_i\right)$$

$$then$$

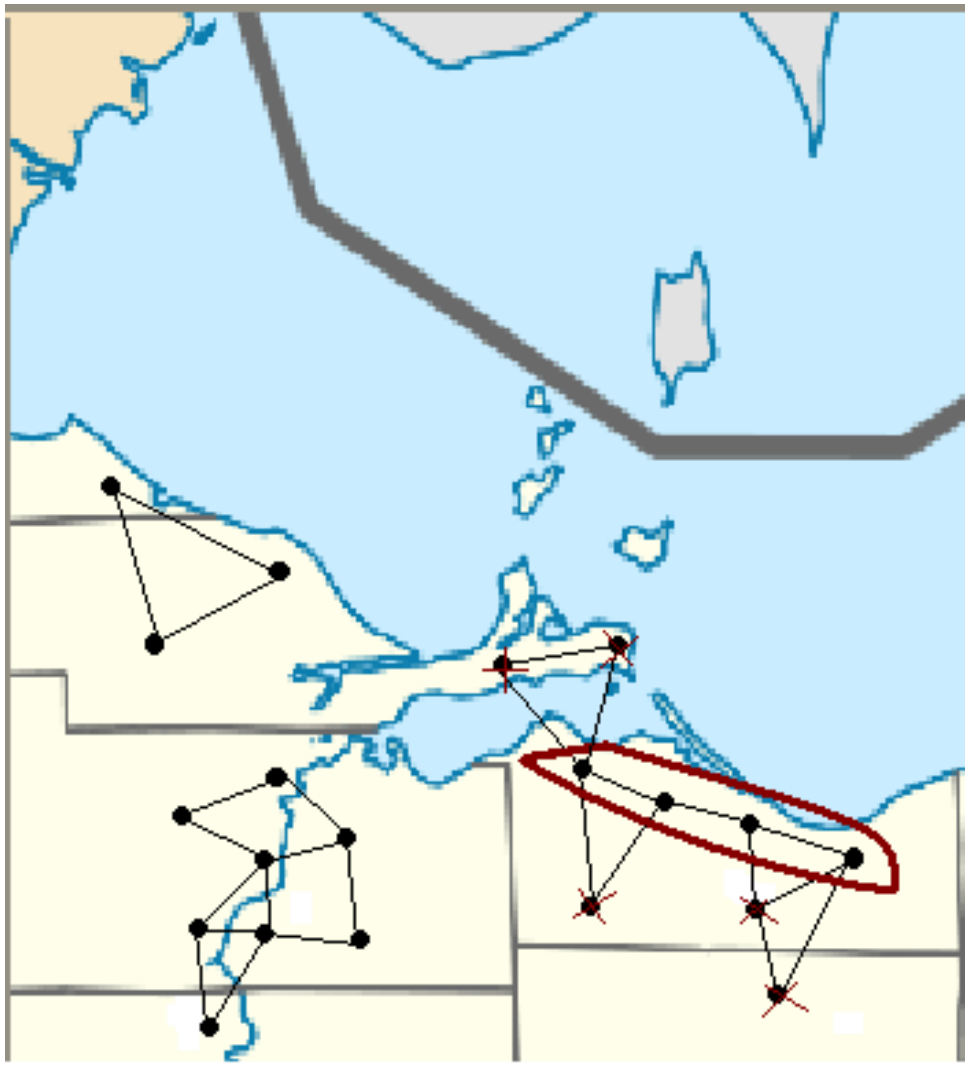
$$\lim_{|C| \to \infty} \left(\sum_{i \in C} \boldsymbol{Y}_i\right)^{-1} E\left[\sum_{i \in C} \tilde{\boldsymbol{Y}}_i\right] = 1$$

Probably OK



Note: This example is completely fabricated. For illustration purposes only.

Probably NOT Ok



Note: This example is completely fabricated. For illustration purposes only.

Application to QCEW Data

Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	2600	2899	3022	2599	11120
Sub1	1981	2256	2382	1957	8576
Sub2	32	33	37	33	135
Sub3	587	610	603	609	2409

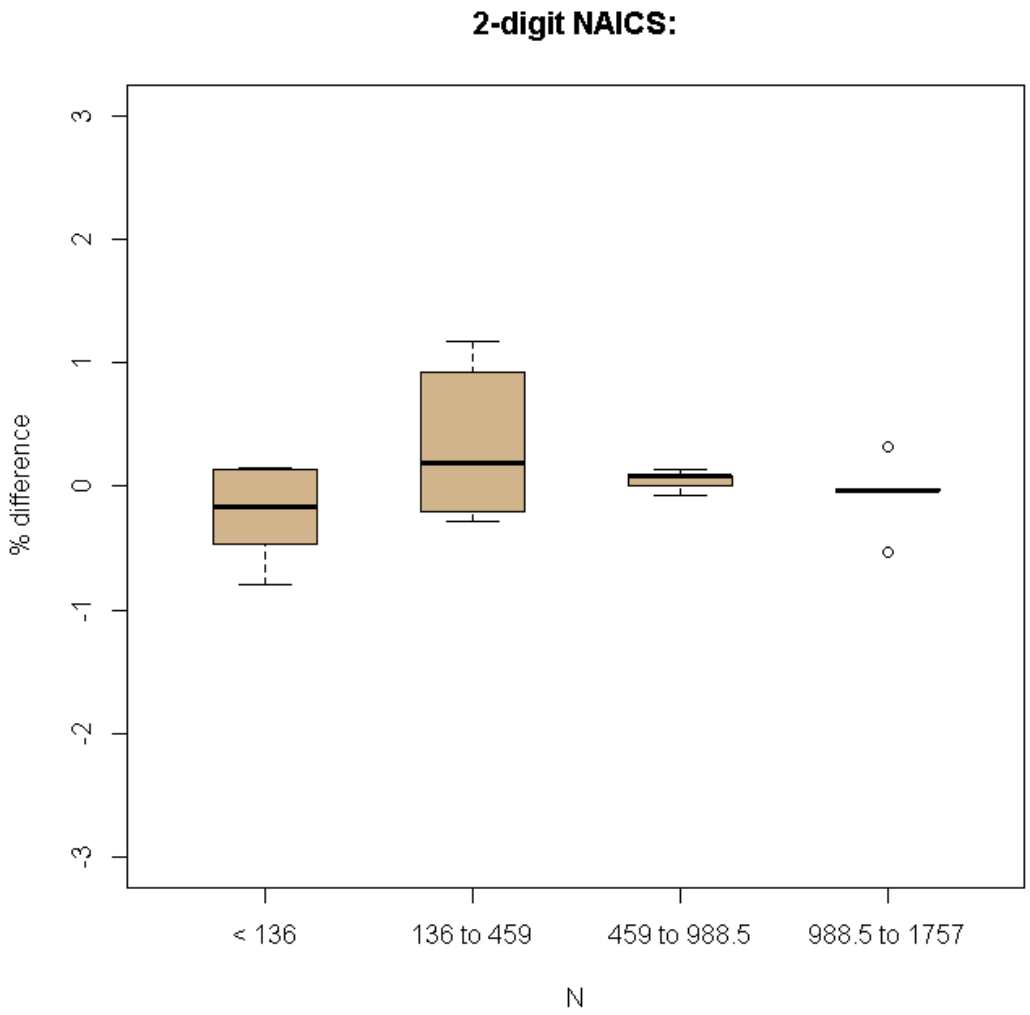
Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	2572	2912	3040	2609	11133
Sub1	1951	2273	2405	1966	8595
Sub2	37	23	42	26	128
Sub3	584	616	593	617	2410

Example 2: Specific MSA (Even Smaller Cells)

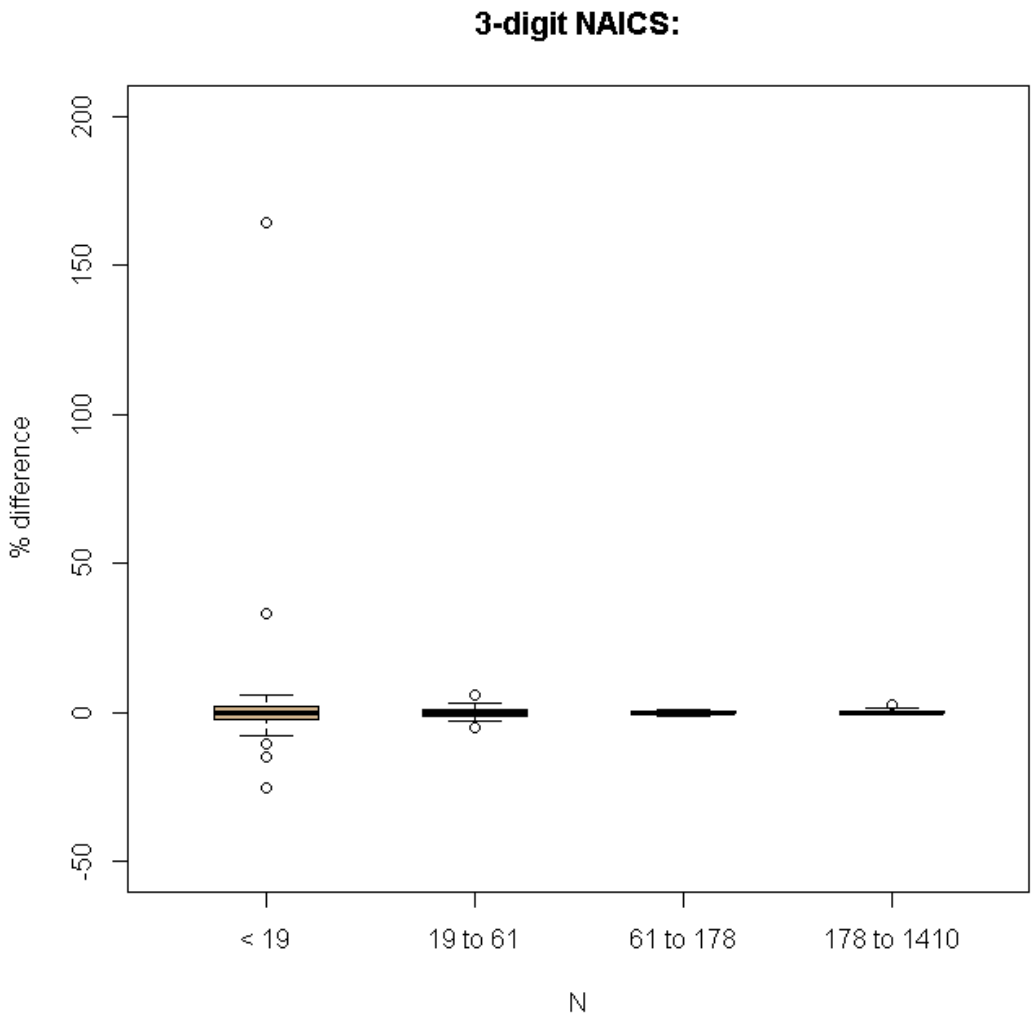
Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	721	754	706	722	2903
Sub1	573	608	566	580	2327
Sub2	51	52	50	48	201
Sub3	97	94	90	94	375

Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	717	732	716	722	2887
Sub1	589	606	572	565	2332
Sub2	27	32	51	51	161
Sub3	101	94	93	106	394

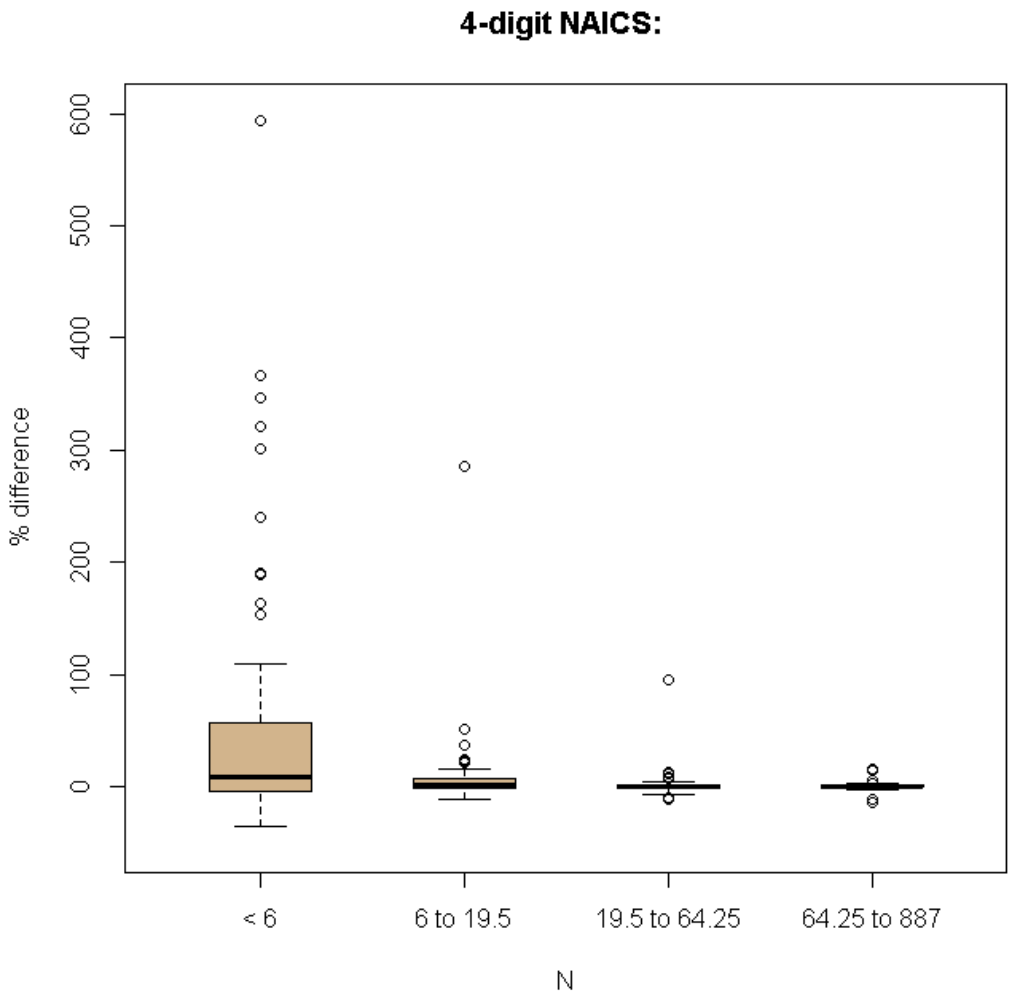
Distribution of % Difference in Values of Cells



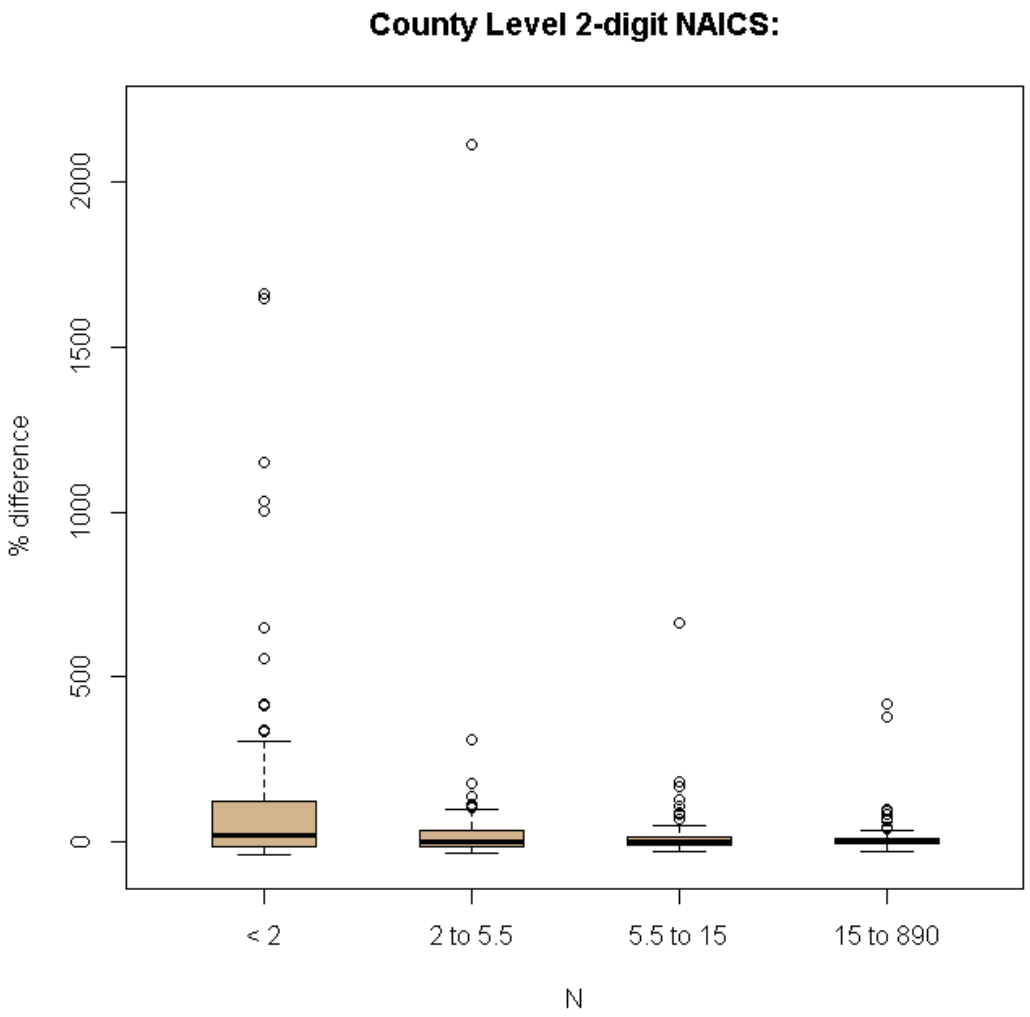
Smaller Cells



Even Smaller Cells



Cells Based on Variable not in Distance Function



Conclusions of Empirical Results

- Allows releasing of “micro-dataset” for use in producing aggregated tables.
- Method seems to offer adequate protection to small and large establishments
- As N gets larger the cell using the synthetic data get closer to true value (relative difference).
- Accuracy of cell depends strongly on distance used.

Thank You

toth.daniell@bls.gov