Reclassification Testing and Rectification for Time Series Survey Discontinuities.

Tucker McElroy[1]
U.S. Census Bureau

Joint Statistical Meetings
Portland, OR
August 4-8, 2024

[1]This presentation is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau (USCB). All time series analyzed in this presentation are from public or external data sources.

Overview

Topic: Time series survey data that exhibit discontinuities arising from changes to the

- survey (e.g., questions),
- methodology (e.g., use of synthetic data),
- collection (e.g., non-response),
- **classification** (e.g., regional definitions, **industry categories**).

Overview

Goal: test for presence of survey discontinuity arising from classification changes, and rectify discrepancies.

Outline

1. Industry Classification
2. Restatements and Survey Discontinuity
3. Concordance Testing and Rectification
4. AIES Illustration

Industry Classification

Categories: a is an industry category, and I_a is the corresponding set of establishments. The set of categories $\mathcal{A}$ is a classification; this is a partition of the set of establishments.

Example: suppose a has the label "Household Appliance Stores." When an establishment returns the survey, they may be assigned to I_a if it is known that they belonged to I_a in previous surveys; otherwise, a determination is made, which requires resources.

Industry Classification

New Categories: suppose there is an old classification $\mathcal{A}$ and a new classification $\mathcal{B}$. How do we match categories a and b of the two classification systems?

SIC and NAICS: $\mathcal{A}$ could represent 1987 SIC (Standard Industrial Classification) and $\mathcal{B}$ could represent 1997 NAICS (North American Industry Classification System). The year date refers to the definition of categories, since there can be additional categories defined at each economic census.

Industry Classification

Concordance: we have dictionaries that provide a concordance between categories. Let

$$B_a = \{b \in \mathcal{B} : I_a \cap I_b \neq \emptyset\}$$
$$A_b = \{a \in \mathcal{A} : I_a \cap I_b \neq \emptyset\}.$$

For any old a category, B_a gives all the new categories that contain any of the establishments in a.

Industry Classification

Visualization: we can visualize the relationships through a bipartite graph, where the elements of $\mathcal{A}$ and $\mathcal{B}$ are on the left and right sides respectively, and an edge is drawn from a to b if and only if $a \in A_b$ (and an edge is drawn from b to a if and only if $b \in B_a$).

Industry Classification

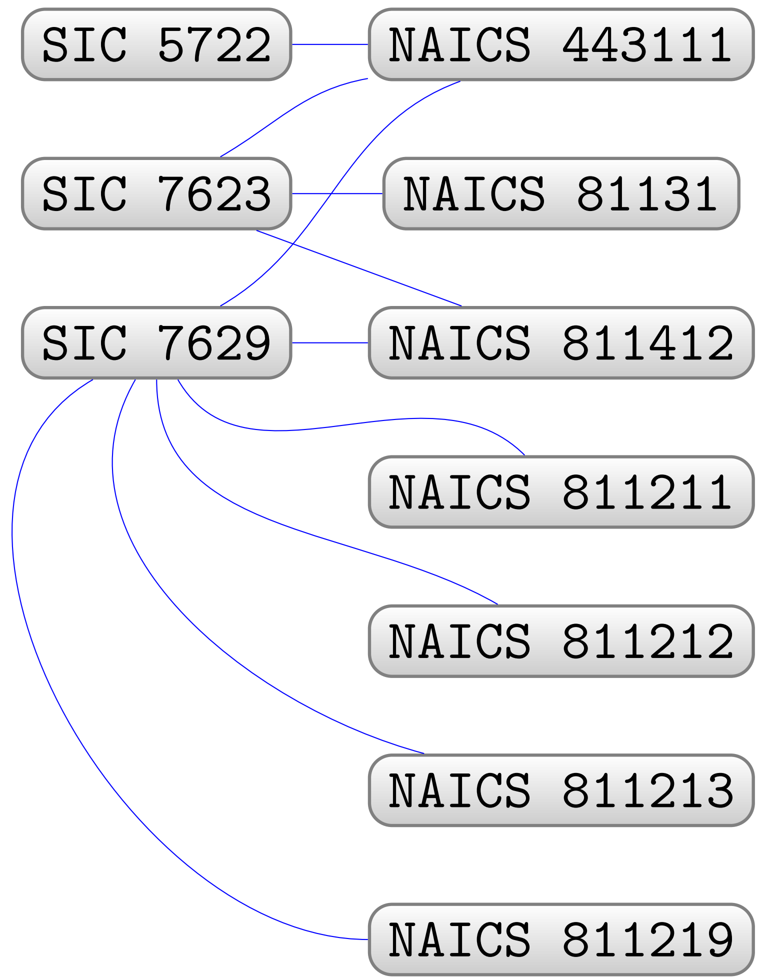
Example: "Household Appliance Stores" is the same label for both $a =$ SIC5722 and $b =$ NAICS443111. Then from the concordance[2]

$$B_{\mathsf{SIC}5722} = \{\mathsf{NAICS}443111\}$$
$$A_{\mathsf{NAICS}443111} = \{\mathsf{SIC}5722, \mathsf{SIC}7623, \mathsf{SIC}7629\}.$$

We also find that

$$\begin{aligned} B_{\mathsf{SIC}7623} &= \{\mathsf{NAICS}443111, \mathsf{NAICS}81131, \mathsf{NAICS}811412\} \\ B_{\mathsf{SIC}7629} &= \{\mathsf{NAICS}443111, \mathsf{NAICS}811212, \mathsf{NAICS}811213, \\ &\quad \mathsf{NAICS}811219, \mathsf{NAICS}811412, \mathsf{NAICS}811211\}. \end{aligned}$$

[2] `https://www.census.gov/naics/?68967`


SIC 5722
SIC 7623
SIC 7629
NAICS 443111
NAICS 81131
NAICS 811412
NAICS 811211
NAICS 811212
NAICS 811213
NAICS 811219

Restatements and Survey Discontinuity

Defining Time Series: let x denote some measurement, and $\{x_t(i)\}$ denotes the time series for establishment i. The category time series is

$$x_t(a) = \sum_{i \in I_a} x_t(i)$$

for each $a \in \mathcal{A}$. Similarly, for each $b \in \mathcal{B}$

$$x_t(b) = \sum_{i \in I_b} x_t(i).$$

United States™ Census Bureau

Restatements and Survey Discontinuity

Breaks: typically the old time series $\{x_t(a)\}$ are discontinued, and only the new time series $\{x_t(b)\}$ are published. Let t_2 be the last time where the old time series has been published.

Break-point: there is a break-point where the reclassification occurs. Let t_1 denote the last time where the new time series has not been published:

$$\underbrace{\ldots, \emptyset_{t_1-1}, \emptyset_{t_1},}_{\text{unpublished}} \underbrace{x_{t_1+1}(b), x_{t_1+2}(b), \ldots}_{\text{reclassified}}$$

Restatements and Survey Discontinuity

Splicing: when a and b are in one-to-one correspondence, we can splice (or concatenate) the two time series – though there may be times of overlap. There might be a discontinuity (due to changes in instrument, response rates, questionnaire design, etc.). If $t_2 \geq t_1$ there is overlap:

$$\overbrace{\ldots,\emptyset_{t_1-1},\emptyset_{t_1},}^{\text{unpublished}}\overbrace{x_{t_1+1}(b),\ldots,x_{t_2}(b),x_{t_2+1}(b),\ldots}^{\text{reclassified}}$$

$$\underbrace{\ldots,x_{t_1-1}(a),x_{t_1}(a),x_{t_1+1}(a),\ldots,x_{t_2}(a),}_{\text{old}}\ \underbrace{\emptyset_{t_2+1},\ldots}_{\text{unpublished}}$$

Restatements and Survey Discontinuity

What do we do when there is not a one-to-one correspondence?

Restatement: to find past values (for $t \leq t_1$) of $x_t(b)$, we need to find I_b at earlier times; however, this can be expensive to determine. Instead, we know I_{A_b}, the establishments in the old categories that map to b. So the restatement of $x_t(b)$ is defined as

$$y_t(b) = \sum_{i \in I_{A_b}} x_t(i) = \sum_{a \in A_b} x_t(a).$$

Restatements and Survey Discontinuity

Terminology: we say the time series has been restated on the $\mathcal{A}$ basis.

- $x_t(b)$ is the "reclassified" time series
- $y_t(b)$ is the "restated" time series

These are available at different times, typically.

Error: the restatement *does* include all establishment that are in b, but *can* include some establishments not in b.

Restatements and Survey Discontinuity

We can compute the restatement up to our last values for the old classification: $t \leq t_2$.

$$\overbrace{\ldots, \emptyset_{t_1-1}, \emptyset_{t_1},}^{\text{unpublished}} \overbrace{x_{t_1+1}(b), \ldots, x_{t_2}(b), x_{t_2+1}(b), \ldots}^{\text{reclassified}}$$

$$\underbrace{\ldots, y_{t_1-1}(b), y_{t_1}(b), y_{t_1+1}(b), \ldots, y_{t_2}(b),}_{\text{restatement}} \; \underbrace{\emptyset_{t_2+1}, \ldots}_{\text{unpublished}}$$

Restatements and Survey Discontinuity

Example: with b corresponding to NAICS443111, the restatement is

$$y_t(b) = x_t(a_1) + x_t(a_2)$$

with $a_1 =$ SIC5722 and $a_2 =$ SIC7623. Although every establishment in a_1 is also in b, there are some establishments in a_2 that are not in b, because SIC7623 also shares establishments with NAICS81131 and NAICS811412.

- $a_1 =$ SIC5722 is publicly available
- $a_2 =$ SIC7623 is not

Restatements and Survey Discontinuity

Linking: due to error an adjustment should be made to the restatement, such that $y_t(b)$ for $t \leq t_1$ "matches" the unknown $x_t(b)$. Assume the unknown $x_t(b)$ for $t \leq t_1$ differs from $y_t(b)$ by a scalar factor τ: then $linking$ estimates τ and modifies the restatement to

$$\widehat{\tau} \cdot y_t(b).$$

We can estimate τ from the overlap period $t_1 + 1, \ldots, t_2$ (say, of length $K = t_2 - t_1$), where both $x_t(b)$ and $y_t(b)$ are available. The geometric mean is one estimate of τ:

$$\widehat{\tau} = \left(\prod_{h=1}^{K} x_{t_1+h}(b) / y_{t_1+h}(b) \right)^{1/K}.$$

Restatements and Survey Discontinuity

Example: in Shimberg et al. (2002), availability of the SIC data up through March 2001 allowed setting $K = 2$, with t_1 corresponding to January 2001 and t_2 corresponding to March 2001. (Additional steps were applied in Shimberg, such as benchmarking.)

Only internal calculations are possible for this case; $a_2 = \text{SIC}7623$ is not publicly available.

Concordance Testing and Rectification

Challenges: How to test whether a restatement is close to the reclassified time series? If it is not close, how do we rectify?

Notation: let

- $\{X_t\}$ be the reclassified process, observed for $t > t_1$
- $\{Y_t\}$ be the restated process, observed for $t \leq t_2$

Lower case letters for sample paths, and underline for vectors.

Concordance Testing and Rectification

Concordance Testing: test $x_t = y_t$, with rejection indicating the processes are $discordant$.

Rectification: alter x_t or y_t so that the restatement and reclassification are no longer discordant.

Two Problems:

1. Extend the reclassification backwards ($t \leq t_1$) by using the restatement

2. If the restatement is viewed as more reliable, modify the reclassification forwards ($t > t_2$) to match

Concordance Testing and Rectification

Suppose $t_2 \geq t_1$. Setting $H \geq 0$, let

- $\flat$ denote $\{t_1 - H, \ldots, t_1\}$
- $\natural$ denote $\{t_1 + 1, \ldots, t_2\}$
- $\sharp$ denote $\{t_2 + 1, \ldots T\}$

$$\underbrace{t_1 - H, \ldots, t_1,}_{\flat} \underbrace{t_1 + 1, \ldots, t_2,}_{\natural} \underbrace{t_2 + 1, \ldots, T}_{\sharp}$$

Concordance Testing and Rectification

Data Structure: $\underline{X}^\flat$ and $\underline{Y}^\sharp$ are missing.

$$\begin{array}{ccc} \emptyset & \underline{X}^\natural & \underline{X}^\sharp \\ \underline{Y}^\flat & \underline{Y}^\natural & \emptyset \end{array}$$

Concordance Testing and Rectification

Extending Backwards the Reclassification: the optimal Mean Squared Error (MSE) estimator of $\underline{X}^\flat$ given the available information in the reclassification is

$$\widehat{\underline{X}^\flat} = \mathbb{E}[\underline{X}^\flat | \underline{X}^\natural, \underline{X}^\sharp],$$

and $V_\flat = \mathbb{E}[(\widehat{\underline{X}^\flat} - \underline{X}^\flat)(\widehat{\underline{X}^\flat} - \underline{X}^\flat)^{'}]$ is the MSE matrix. This estimate ignores information in the restatement; we could include this if it were feasible to model X and Y as a bivariate system (which would required $K = t_2 - t_1$ to be large).

Concordance Testing and Rectification

Concordance Test: the null hypothesis is

$$H_0 : \underline{x}^\flat = \underline{y}^\flat.$$

The error $\widehat{\underline{X}^\flat} - \underline{X}^\flat$ has mean zero and variance matrix $V_\flat$. The test statistic is

$$\mathcal{T}^\flat = (\widehat{\underline{x}^\flat} - \underline{y}^\flat)' V_\flat^{-1} (\widehat{\underline{x}^\flat} - \underline{y}^\flat),$$

which has a χ^2 distribution on $H+1$ degrees of freedom if the processes are Gaussian.

Concordance Testing and Rectification

Rectification: here we view the restatement $\underline{y}^\flat$ as the backwards extension of the reclassification; if it deviates too much from the backcasts, we modify the restatement.

Modification: modify the restatement from $\underline{y}^\flat$ to $\underline{\widehat{x}}^\flat$ such that H_0 is no longer rejected. We may want other properties, such as having growth rates of $\underline{\widehat{x}}^\flat$ match those of $\underline{y}^\flat$.

Concordance Testing and Rectification

Forward Modifying the Reclassification: the optimal Mean Squared Error (MSE) estimator of $\underline{Y}^\sharp$ given the available information in the restatement is

$$\widehat{\underline{Y}^\sharp} = \mathbb{E}[\underline{Y}^\sharp | \underline{Y}^\flat, \underline{Y}^\natural],$$

and $V_\sharp = \mathbb{E}[(\widehat{\underline{Y}^\sharp} - \underline{Y}^\sharp)(\widehat{\underline{Y}^\sharp} - \underline{Y}^\sharp)^{'}]$ is the MSE matrix. This estimate ignores information in the reclassification; we could include this if it were feasible to model X and Y as bivariate system (which would require $K = t_2 - t_1$ to be large).

Concordance Testing and Rectification

Concordance Test: the null hypothesis is

$$H_0 : \underline{y}^\sharp = \underline{x}^\sharp$$

The error $\widehat{\underline{Y}^\sharp} - \underline{Y}^\sharp$ has mean zero and variance matrix $V_\sharp$. The test statistic is

$$\mathcal{T}^\sharp = (\widehat{\underline{y}^\sharp} - \underline{x}^\sharp)' V_\sharp^{-1} (\widehat{\underline{y}^\sharp} - \underline{x}^\sharp),$$

which has a χ^2 distribution on $T - t_2$ degrees of freedom if the processes are Gaussian.

Concordance Testing and Rectification

Rectification: here we view the reclassification $\underline{x}^\sharp$ as less reliable than the restatement; if it deviates too much from the forecasts $\widehat{\underline{y}^\sharp}$, we modify the reclassification from $\underline{x}^\sharp$ to $\widetilde{\underline{x}}^\sharp$ such that H_0 is no longer rejected.

Modification: suppose we want the growth rates of $\widetilde{\underline{x}}^\sharp$ to match those of $\underline{x}^\sharp$. Letting r denote the first component of $\widetilde{\underline{x}}^\sharp$, we require

$$\widetilde{\underline{x}}^\sharp = (r - x_{t_2+1})\iota + \underline{x}^\sharp,$$

where ι is a vector of ones. Select r to be as close as possible to x_{t_2+1} and such that H_0 is not rejected; there is a formula for such r.

AIES Illustration

Data Integration: the U.S. Census Bureau (USCB) is integrating seven current annual economic surveys into the Annual Integrated Economic Survey (AIES)[3].

Reclassification: the older data categories are in one-to-one correspondence with new AIES data categories; but restatement (of older date) can differ from AIES data due to methodology changes.

[3] https://www.census.gov/programs-surveys/aies/about.html

AIES Illustration

AIES Challenge: let $\{X_t\}$ be the AIES process (for some particular variable or category) and $\{Y_t\}$ be the restatement based upon data compiled from relevant surveys; is AIES greatly divergent from the supporting surveys? If so, we want to *forward modify the reclassification* (application of our second problem)[4].

Data: whereas $\{Y_t\}$ is public, $\{X_t\}$ has not yet been produced, so we will use fictitious numbers.

Specification: consider "Automotive parts, access, and tire stores" (4413), or Auto for short. This annual time series $\{Y_t\}$ is available from 1992 through 2022, and is plotted in log scale in Figure 1.

[4] Disclaimer: this is research, not an official method of USCB.

AIES Illustration

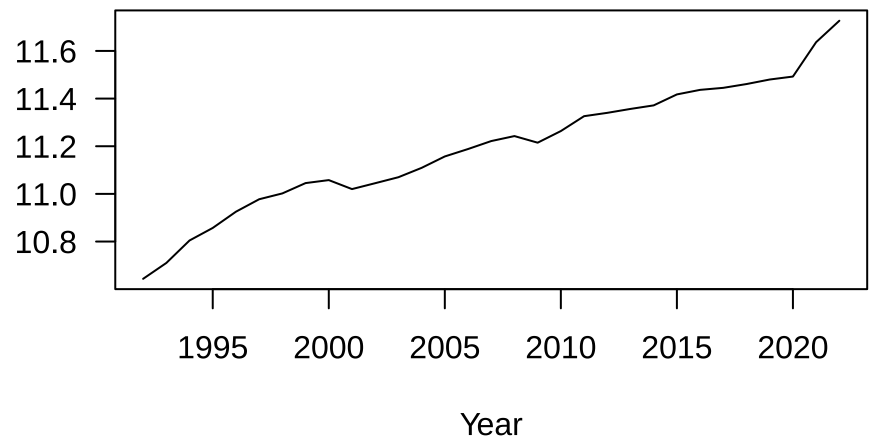


Figure 1: Plot of logged "Automotive parts, access, and tire stores", series 4413 of ARTS, for years 1992 through 2022.

AIES Illustration

Modeling: to generate forecasts, we model logged $\{Y_t\}$ via an ARIMA(0,1,1) with trend constant and level shift.

Synthetic AIES: for years 2023 and 2024, we create synthetic values by multiplying the 2022 Auto value by 1.5 and 1.7 respectively.

Forecasting: we forecast the Auto series 2 steps ahead (Figure 2), with backcasts for comparison, and shading to denote pointwise confidence intervals.

AIES Illustration

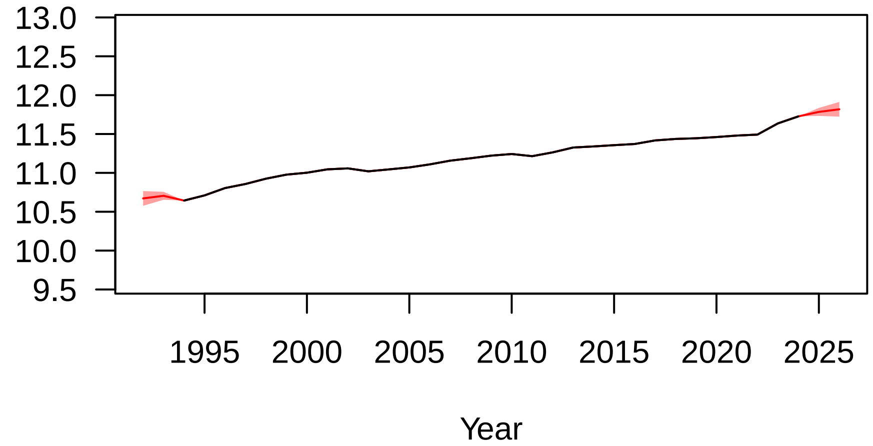


Figure 2: Plot of logged “Automotive parts, access, and tire stores”, series 4413 of ARTS, for years 1992 through 2022. Series in black, with forecasts and forecast intervals in red.

AIES Illustration

Concordance: the Wald statistic is 200.37, with critical value 5.99 for $\alpha = .05$; so null is rejected, indicating discordance.

Rectification: rectification lowers the Wald statistic to 9.72, which corresponds to $\alpha = .0078$; this is the best possible with the constraint of preserving growth rates.

AIES Illustration

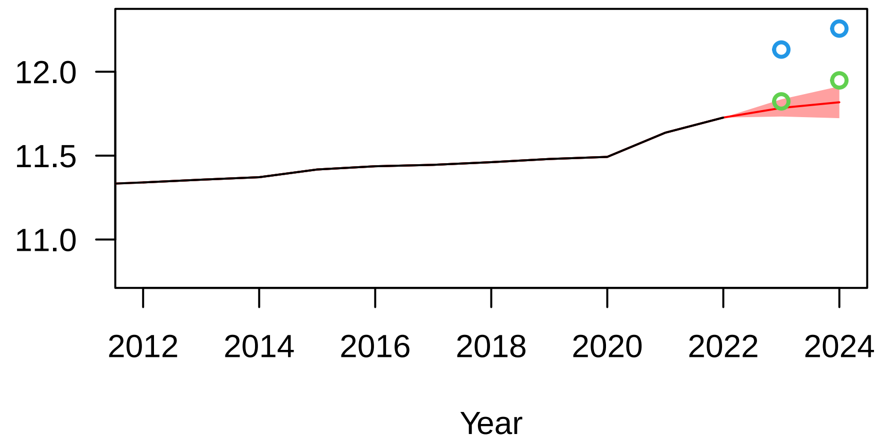


Figure 3: Plot of logged "Automotive parts, access, and tire stores", series 4413 of ARTS, for years 1992 through 2022. Series in black, with forecasts and forecast intervals in red. Blue dots are the synthetic AIES values, and green dots are their rectifications.

Summary

1. We provide a formal framework for reclassification and restatements
2. We develop concordance testing frameworks for the problems of *backwards reclassification extension* and *forward reclassification modification*
3. We develop rectification strategies
4. Illustrated on synthetic AIES data

Future Work

1. Develop SIC to NAICS reclassification illustrations for monthly time series

2. Extend AIES illustration (eventually replace synthetic with real data subject to privacy constraints)

Contact: tucker.s.mcelroy@census.gov