Machine Learning is (not!) all you need

Impact of classification-induced error on price indices using scanner data

William Spackman, Serge Goussev, Mackenzie Wall, Greg DeVilliers, David Chiumera

Presented at 2024 Ottawa Group, 2024-05-15

Delivering insight through data for a better Canada

Outline

RESEARCH OVERVIEW AND OBJECTIVES

OPEN DATA USED

EXPERIMENT DESCRIPTION AND METHODS

RESULTS

CONCLUSIONS

Research Objectives & Problem Statement

Context

Machine Learning is increasingly applied in price statistics

Scale of Alternative Data Sources (ADS) makes validation of 100% of records infeasible

Problem statement

Misclassification is known to cause measurement error

- False positives or false negatives could potentially affect the average movement

The relationship between misclassification and bias in the price index is not clearly understood

Authors are unaware of a public research paper on the topic

Objective

Study misclassification on scanner data

Evaluate how misclassification could impact the elementary indices (bilateral & multilaterals)

Evaluate mitigating strategies

Use open data and code to enable extension of the research and better peer-review

Research Questions

- RQ1: Does misclassification affect a Törnqvist price index for one period?
  - Inject various levels of random misclassification into the data to see if an elementary price index for a single product category could be affected in one reporting period

- RQ2: Does misclassification affect a GEKS-Törnqvist (i.e. CCDI) over a long period of time?
  - Inject various levels of random misclassification over a 6-year period and compute the CCDI on 4 different product categories

Research Questions

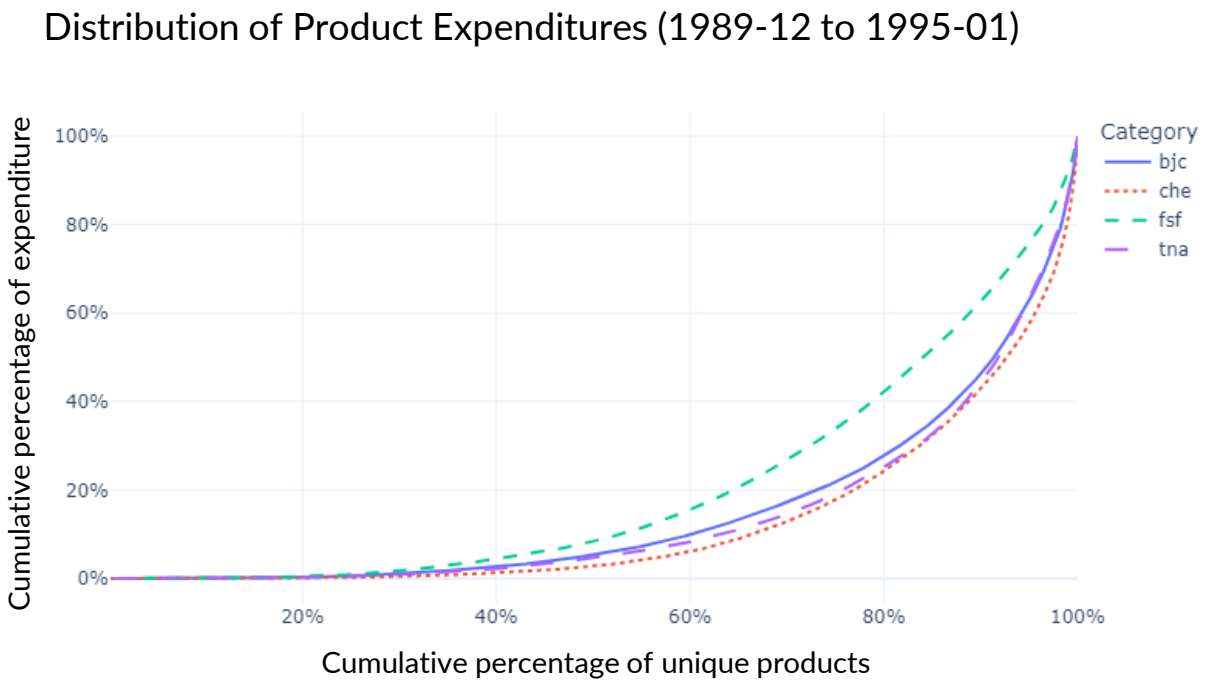
- RQ3: What ML model metrics can be prioritized to minimize bias in the index?
  - Compare the results from RQ1 and RQ2 at various levels of misclassification to see which metrics (precision, recall, F1 score) are most relevant when attempting to minimize index bias
- RQ4: Is the extension method chosen sensitive to misclassification?
  - Compare the experiment from RQ2 with different extension methods. Mean Splice on Published and Half Splice on Published.
- RQ5: What validation thresholds and which methods are most appropriate for mitigating misclassification?
  - If we expect that there are misclassified products in the dataset, what are the best ways to select products for manual review?

Dataset

- Dominick's Finer Foods scanner dataset
  - Recommended for NSOs as benchmark for evaluation of methods (Mehrhoff 2019)
  - Pre-categorized so misclassification can be simulated and compared to the true index
  - Leveraged by other NSOs (Lamboray (2021), ONS (2020)) for public research needs

- Transformations:
  - Weekly to monthly data
  - Item code to define unique products
  - Subset (December 1989 – December 1995)

- Categories chosen based on number of unique products + number of matched products monthly:
  - Bottled juices (bjc)
  - Cheese (che)
  - Fabric softeners (fsf)
  - Canned tuna (tna)



Experiment Design

Calculate “True” price index value(s) for category/categories of interest

Misclassify Products. i.e. Change the product categories for a specified fraction of products to meet the target misclassification rate

Calculate index value(s) for category/categories of interest based on products classified to those categories

Repeat multiple times K to obtain a distribution of calculated indices

Methodology

- Perform K Monte Carlo simulations of misclassification, estimating the index $\hat{P}_{k,i}$ for each iteration.
- Calculate the mean index: $\overline{\hat{P}_i}$, bias: $B(\hat{P}_i)$, variance $V(\hat{P}_i)$ and root mean squared error $\widehat{RMSE}$
- Repeat the K simulations for different levels of misclassification (Precision and Recall)
- Record F_1 score

$$\overline{\hat{P}_i} = \frac{1}{K}\sum_{k=1}^{K} \hat{P}_{k,i}$$

$$B(\hat{P}_i) = \overline{\hat{P}_i} - P_i$$

$$V(\hat{P}_i) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{P}_{k,i} - \overline{\hat{P}_i}\right)^2$$

$$\widehat{RMSE} = \sqrt{B(\hat{P}_i)^2 + \mathrm{V}(\hat{P}_i)}$$

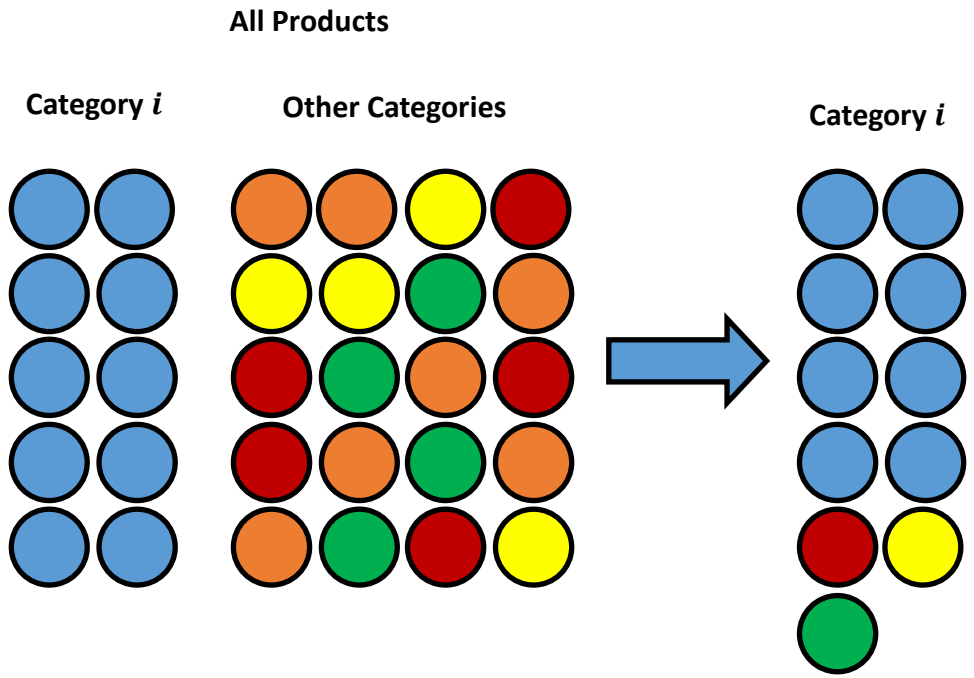
$$F_\beta = \frac{(1+\beta^2) * Precision * Recall}{\beta^2 * Precision + Recall}$$

Example of Misclassifications

1. Use all matched unique products from the time period of interest

2. Randomly select unique products from category i such that the recall is the desired level. i.e. for recall=0.8, select 80% of unique products that are in category i

3. Randomly select products from the other categories such that the precision is equal to the desired level. In this example 0.73



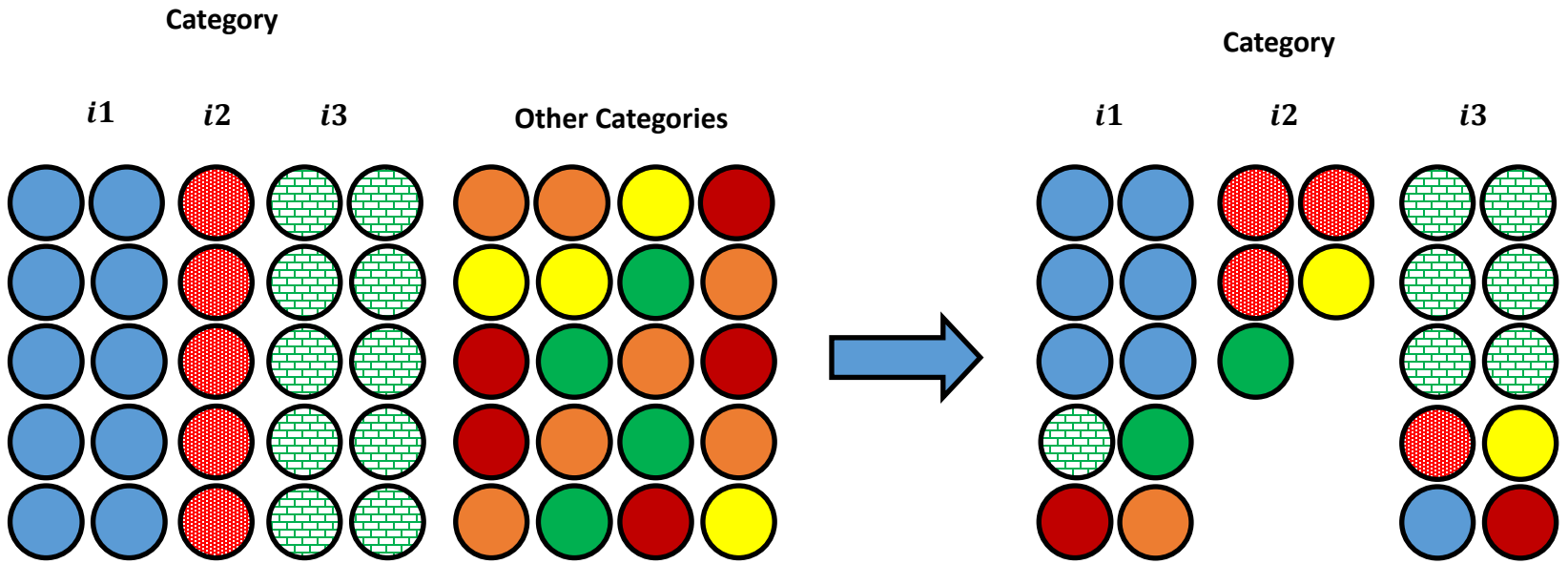
$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{8}{10}$$

$$Precision = \frac{8}{11}$$

Example of Misclassifications



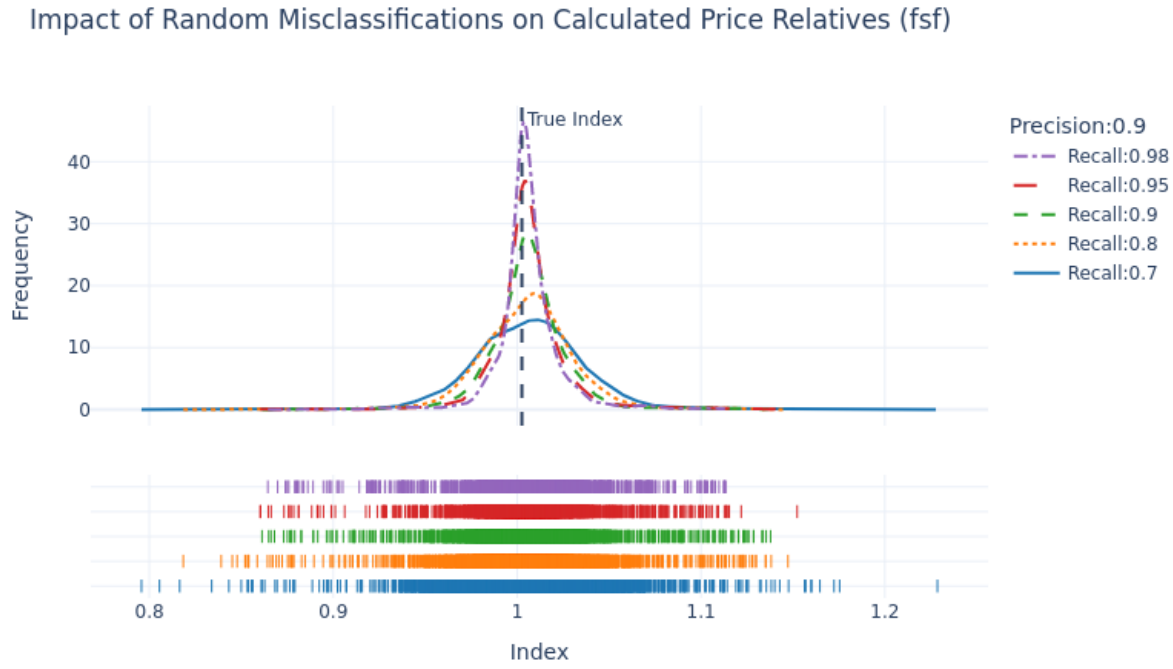
$$Recall = \frac{6}{10} = \frac{3}{5}$$

$$Precision = \frac{6}{10} = \frac{3}{5}$$

For multilateral experiments, take all products in the 6-year time period and misclassify in a similar way. We are only concerned for the precision and recall of the categories of interest, in this case $i1$, $i2$ and $i3$

Misclassification Impact on Bilateral Index

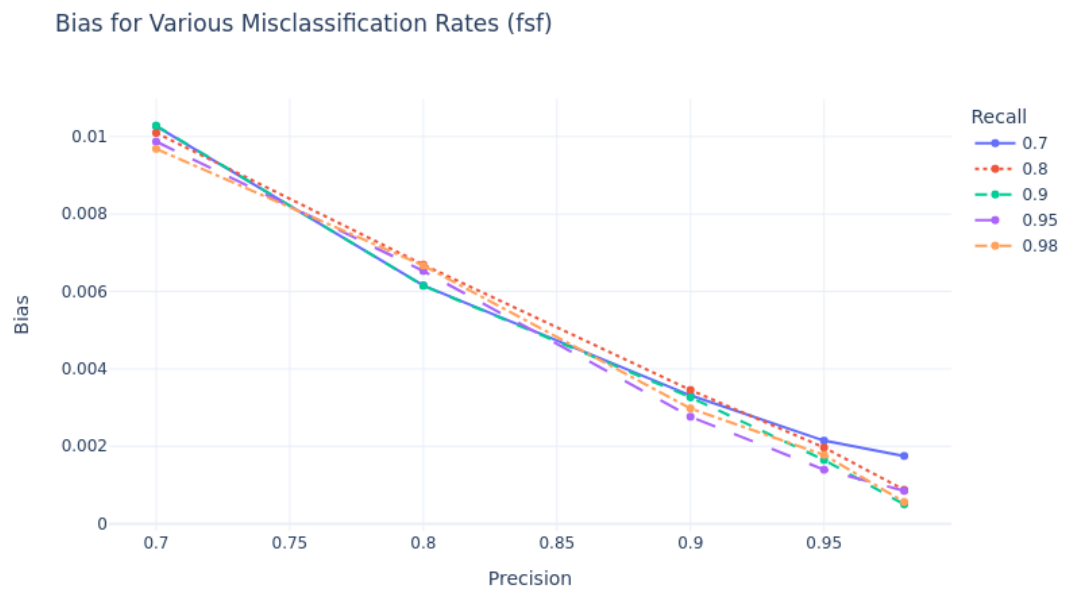
- Does misclassification affect a Törnqvist price index for one period?
- Plot of Distribution of calculated Törnqvist price index for example fabric softner “fsf” category 1991-01 to 1991-02
- Random misclassifications can cause inaccuracies in the calculated index, even at relatively low rates (high precision and high recall

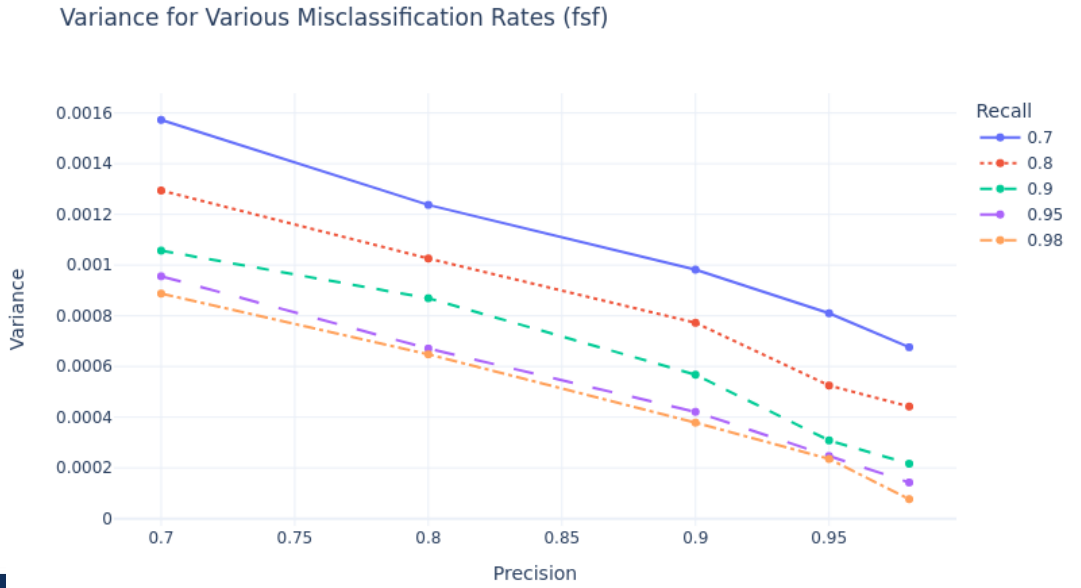


Category: {Pr, Re}	variance	bias	RMSE	F1
fsf: {0.90, 0.50}	0.00205	0.00292	0.04536	0.64286
fsf: {0.90, 0.70}	0.00170	0.00442	0.04142	0.78750
fsf: {0.90, 0.80}	0.00099	0.00253	0.0315	0.84706
fsf: {0.90, 0.90}	0.00089	0.00125	0.02980	0.90000
fsf: {0.90, 0.95}	0.00043	0.00316	0.02103	0.92432
fsf: {0.90, 0.98}	0.00046	0.00228	0.02146	0.93830

Bilateral Index Cont.

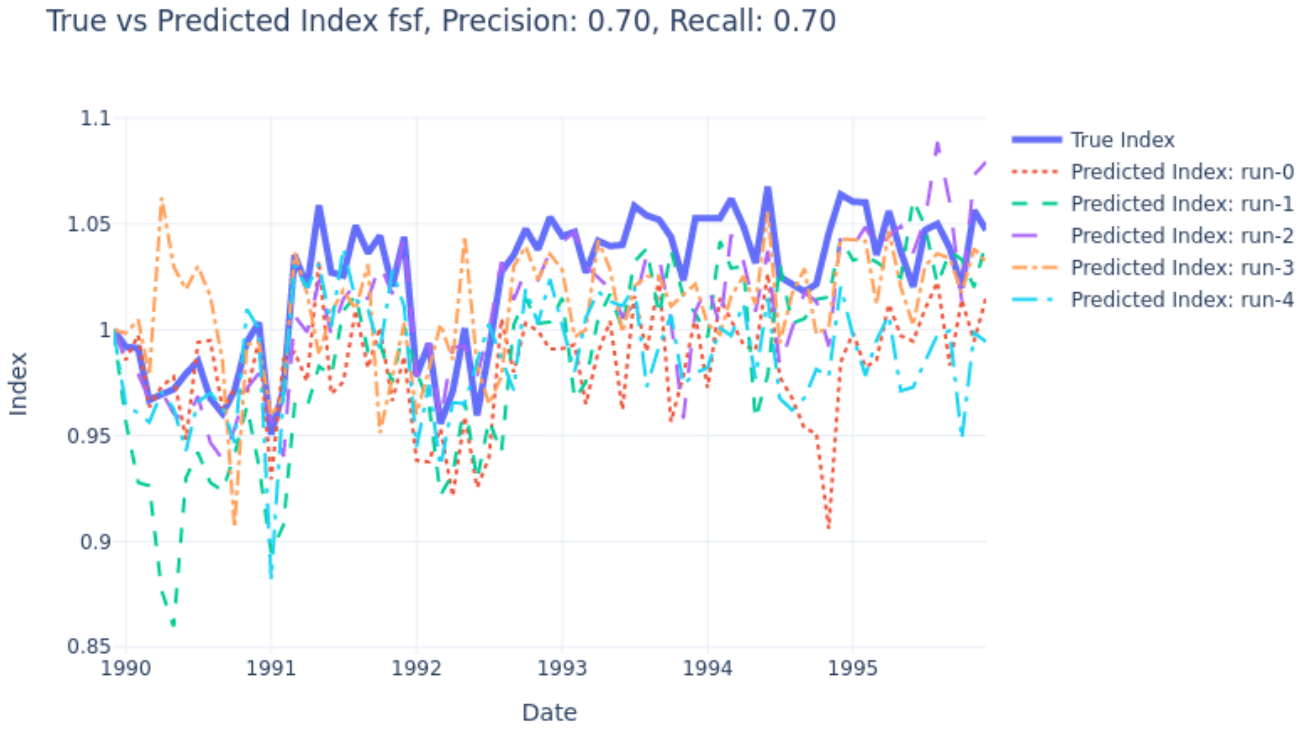
- Impact of various precision and recall levels on Bias and Variance for bilateral index for single category: fabric softener (fsf)
- Bias is reduced, predominantly at higher precision rates
- Variance is reduced at both higher precision and recall rates





Misclassification Impact on Multilateral

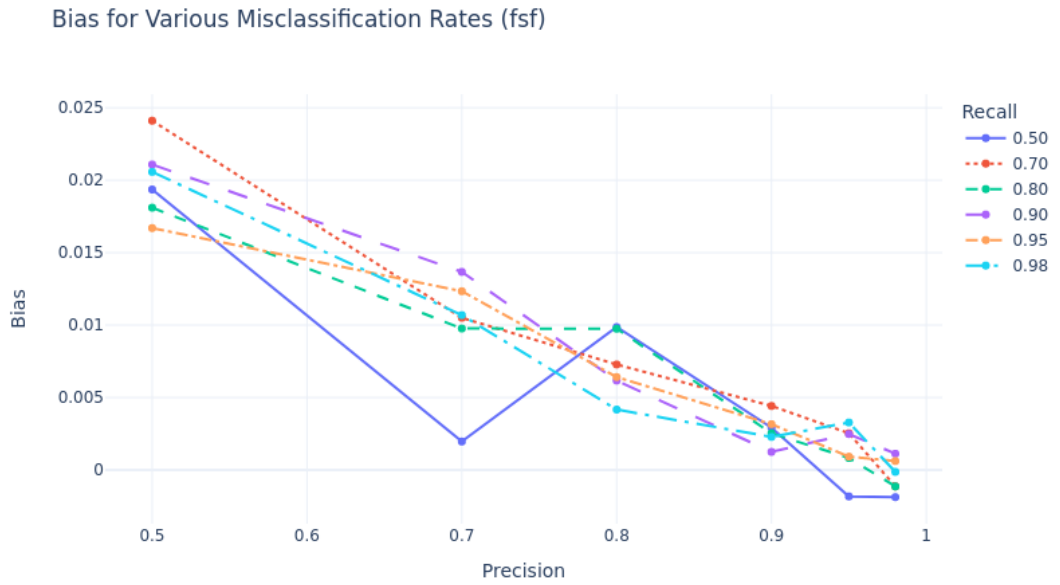
- Does misclassification affect a GEKS-Törnqvist (i.e. CCDI) over a long perio of time?
- Plot of multilateral indices (CCD) for individual runs at $\{pr = 0.7, re = 0.7\}$
- Index can deviate from the true index as a result of misclassifications.
- Deviations persist over entire period of interest.

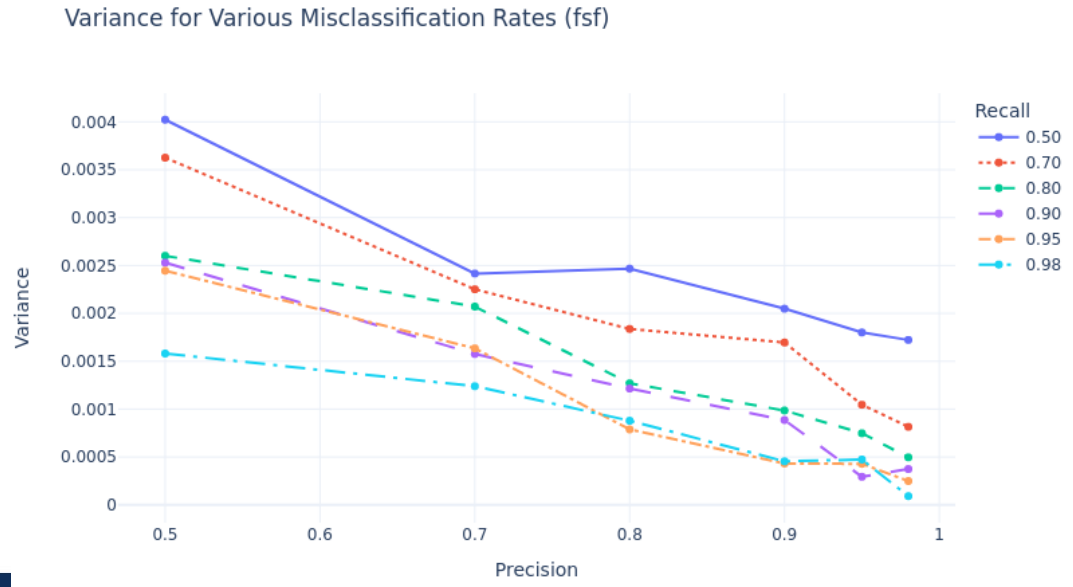


Multilateral Cont.

- Index for fabric softener (fsf) at 1995-12

- Observe directionally similar trends to the bilateral experiment





Comparison of Extension Methods

- Compared two popular extension methods as modest misclassification level:
  - Precision = 0.8
  - Recall = 0.8
- Moderate difference in Bias of estimated index at December 1995.

Extension Method	True Index	Mean Index	Bias
Mean Splice on Published	1.047	1.057	0.010
Half Splice on Published	1.052	1.057	0.006

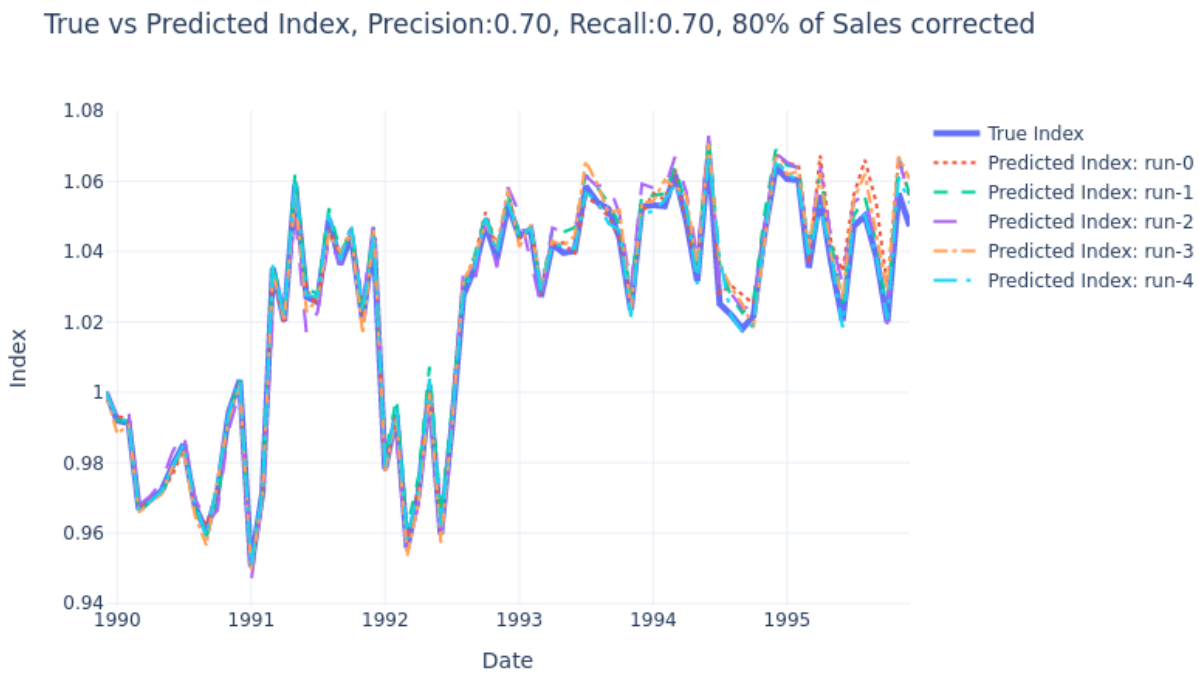
Comparison of Metrics

- When training machine learning classifiers, we can choose which metrics to optimize; F_1 score is a common choice
- When comparing potential classifiers, F_1 score, which weights precision and recall equally, may not correlate well to what we care about: accuracy of the estimated index.
- May wish to put apply weight to precision using F_β with $\beta < 1$

Row	Summary of Experiment	{Precision, Recall}	Variance	Bias	RMSE	F1
1	Fixed recall, varying precision	**{0.50, 0.90}**	**0.00253**	**0.02108**	**0.05453**	**0.64286**
2		{0.70, 0.90}	0.00158	0.01367	0.04199	0.78750
3		{0.80, 0.90}	0.00122	0.00617	0.03540	0.84706
4		{0.90, 0.90}	0.00089	0.00125	0.02980	0.90000
7	Varying recall, fixed precision	**{0.90, 0.50}**	**0.00205**	**0.00292**	**0.04536**	**0.64286**
8		{0.90, 0.70}	0.00170	0.00442	0.04142	0.78750
9		{0.90, 0.80}	0.00099	0.00253	0.0315	0.84706
10		{0.90, 0.90}	0.00089	0.00125	0.02980	0.90000

Selecting Products to Correct

- Individual products can have their predicted category reviewed by human experts.
- With volume of scanner data, reviewing all records is likely impossible or impractical.
- Want to use this finite resource of human reviewers, in the most efficient way possible.
- Sales are not evenly distributed amongst products. Can review 80% of sales by reviewing ~35% of individual products.
- Applying corrections to the top 80% of products based on sales volumes reduced the error in the CCDI .



Conclusions

- Misclassification, even at limited levels, affected calculated price indices
  - High classifier performance does not mean an error-free index
  - Review processes must be uniquely tailored based on the data
- Qualitative results
  - Variance decreased as precision + recall increased
  - Bias decreased as precision increased but less affected by increasing recall
  - Bilateral indices may proxy multilateral misclassification outcomes
- Review Strategies
  - Prioritize reviewing products based on their index formula weights (e.g. sales)
  - Further evaluation of and combination with additional review options is recommended
- Limitations and topics for further research
  - NSOs need to measure likely bias from residual misclassification with only a sample of products
  - Misclassification is not necessarily random; should evaluate the impact of different types of misclassification

Questions?

Please reach out to william.spackman@statcan.gc.ca and serge.goussev@statcan.gc.ca.