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[1]

Presented at the 18th Meeting of the Ottawa Group; Ottawa, Canada; 13-15 May 2024

May 15, 2024

Abstract:

As Machine Learning (ML) is increasingly used to classify large amounts of alternative data into categories of the Consumer Price Index (CPI), National Statistical Organizations (NSOs) need to develop robust processes to mitigate measurement error resulting from ML model misclassification. More specifically, NSOs need to understand the relationship between traditional classification metrics such as accuracy, recall and precision and a price index to determine a threshold of a chosen metric that reduces the error of the price index to an acceptable level. This paper explores the connection between traditional classification metrics and measurements of error such as bias, variance and Root Mean Squared Error of a calculated elementary index against a 'true' index using a public scanner dataset and by carrying out Monte Carlo simulations of misclassification. Further, this paper investigates various methods used to alleviate the impact of misclassification to provide valuable guidance to NSOs regarding the amount of manual intervention that is required to produce a price index at an acceptable level of error. While this research provides insight into the overall relationship between classification metrics and measurements of error that would be difficult to repeat in a production scenario, it provides valuable context and brings forth approaches NSOs may implement in a production environment to reduce the error of elementary indices. The research also utilizes public data to simplify replication and expansion of the experiments.

Keywords: Misclassification, Scanner data, Törnqvist index, Multilateral index, Monte Carlo simulation

[1] Correspondence: to contact the authors, please reach out to serge.goussev@statcan.gc.ca and william.spackman@statcan.gc.ca

1. Introduction

Supervised Machine Learning (ML) is increasingly used by National Statistical Offices (NSOs) for their Consumer Price Index (CPI), specifically to classify alternative data sources (most notably scanner and web scraped data) as part of the processing step to prepare data each month into a format ready for elementary price index compilation. Application of ML is however not a straightforward task. ML models make mistakes, even if on average they perform quite well, hence misclassified products could be used in the price index compilation step and thus impact the calculated elementary aggregate. While a validation process post classification is strongly recommended in the CPI Manual and by Eurostat (section 10.20, Consumer Price Index Manual 2020; Eurostat 2017), it remains unclear what impact various levels of misclassification could have on a compiled index, or exactly what proportion of validation is necessary to mitigate misclassification. Practically speaking, the scale of the data could be considerable and NSOs need to understand how many unique products need to be flagged as this impacts how much resources are needed for the validation step. NSOs thus need to understand which misclassification detection methods are most appropriate. While confidence score, as well as several other methods like price outliers or random samples, are commonly used to flag products (Myklatun 2019; Hov 2021; Greenhough 2022; Spackman, et al. 2023), these methods are commonly evaluated within the context of increasing the accuracy of the overall classification step, rather than directly measuring whether bias remains in the compiled index compared against a 'true' index based on completely correct data. The aim of this working paper is thus to contribute to the discussion of both topics, firstly by simulating misclassification to evaluate its impact against a 'true' compiled elementary index, and secondly by evaluating appropriate flagging methods to mitigate misclassification, helping quantifying how much validation should be necessary.

Within the broader measurement error literature, the impact of misclassification on bias (i.e. misclassification bias) of aggregated statistics has been actively researched. Various studies have shown how misclassification can bias different types of aggregations, such as aggregate turnover of classified records (Meertens, Diks, et al. 2020), counts and proportions (Scholtus and van Delden 2020), or ratios (Van Delden, et al. 2023). These studies also demonstrate approaches to correct the aggregation and reduce misclassification bias, such as by adopting an aggregate corrections step that can be utilized after the initial biased aggregation to correct this bias (Q. Meertens 2021). These studies however have yet to evaluate the more complex scenarios of price indices, and thus a fuller investigation is warranted.

Within the price statistics literature, initial studies have found that classification model performance is tied to bias in the price index, and that misclassification can be present in multiple steps of the classification process (Greenhough 2022; Spackman, et al. 2023; Nietzer 2023; van den Heuvel 2019). It is still unclear however exactly the scale of the problem at various levels of possible model performance. For instance, two key studies on scanner data differ in their conclusions as van den Heuvel (2019) found that accuracy over 0.8-0.85 did not introduce bias on the index, while Nietzer (2023) found that similar levels of performance still led to a considerable proportion of elementary indices violating the Eurostat directive on accuracy, specifically due to misclassification. The differences in the findings are hard to compare however, as the studies utilized different data, different ML models, and calculated elementary indices using different price index methods. To what extent do various levels of misclassification introduce bias into an elementary price index, as well as how much effort is required to mitigate it, are thus still very relevant questions.

To test these objectives, this paper focuses on scanner data with multilateral price indices due to the applicability of this data type and methods to the CPI and the likelihood that the scale of this data may require ML for the classification step. This study also utilizes the publicly available Dominick's Finer Foods dataset (Chicago Booth School for Marketing 2013) and makes the code publicly available in order to support replicability and extendibility

of the research.[2] As the dataset comes pre-categorized and thus can yield a 'true' or correct index, we leverage Monte Carlo misclassification simulations to introduce different levels of random misclassifications and measure the bias, variance, and root mean squared error (RMSE). Five research questions are specifically investigated.

RQ1: Does misclassification affect a Törnqvist price index for one period?

We begin by evaluating how sensitive a weighted Törnqvist is to different levels of misclassifications for two reasons. Firstly, as the annually chained direct Törnqvist is a highly performant price index method that can be applied with scanner data (Lamboray 2021), hence this test will support NSOs looking to leverage this method, and a single-period test will be a useful extension outside of the multilateral focus of the rest of the paper. Secondly, as the GEKS-Törnqvist (also referred to as Caves-Christensen-Diewert-Inklaar or the CCDI) is based on this bilateral index, the single period index thus acts as a valuable benchmark case due to its popularity, versatility (responsive to imputed prices), and satisfaction of performance tests (Eurostat 2022).

RQ2: Does misclassification affect a CCDI over a long period of time?

We next evaluate the CCDI multilateral price index over a long-term period to assess how various levels of misclassification affect this index. We utilize a 25-month window with a mean splice on published extension method to simulate what would occur after 4 years of producing multilateral indices with different levels of misclassification (thus a total index series of 6 years and one month). In our experiment each individual product maintains its assigned category for the entire time period; any misclassified product that "enters" as misclassified remains this way, simulating a production scenario when only new products are classified and validated. Misclassified products will be matched across pairs of months in the window and in theory would have some effect of unknown magnitude.

RQ3 What ML model metrics can be prioritized to minimize bias in the index?

A key question for NSOs when they are selecting a ML classifier is what metrics are most appropriate. This has been studied actively (UK Statistics Authority 2019; Goussev and Spackman 2021) and is currently discussed in the literature within the context of the trade-off between prioritizing classification precision or recall (Spackman, et al. 2023; UK Statistics Authority 2023). Precision for a single category is defined as the number of correct positive predictions (true positives) divided by the total number of predictions to that category. Recall is the number of true positives, divided by the total number of products for that class, in the dataset. Specifically, our focus in this research is to investigate whether the presence of false-positives (products wrongly placed in a category) or false negatives (products wrongly removed from their correct category) yields a better final result in the calculated index. We simulate the various scenarios to add information to NSOs discussing this trade-off.

RQ4: Is the extension method chosen sensitive to misclassification?

We compare the use of the half splice on published (HASP) against the mean splice on published multilateral extension methods, at a single level of misclassification, to see whether there is a difference between these two methods due to misclassification. As these are the two of the most popular extension methods NSOs leverage when applying ML methods, this test is appropriate to validate whether misclassification affects one method more than the other.

---

[2] Code for this study will be made public on the Statistics Canada GitHub page (https://github.com/StatCan) following the 2024 Ottawa Group conference. Please contact the authors if you would like access to the code ahead of time or need help finding it on GitHub.

RQ5: What thresholds of validation, as well as what methods, are most appropriate to mitigate misclassification?

Given that a ML classifier is misclassifying a certain proportion of products, we look to evaluate certain mitigation methods available to NSO's in a production scenario. We consider three possible methods for selecting products, post classification and prior to index compilation, to have their assigned categories manually reviewed by experts. We evaluate the sales proportion cut-off method in detail by, modifying the Monte Carlo misclassification simulations to introducing a sampling process that corrects a certain proportion of records just before compiling the CCDI, but after the misclassification. We thus evaluate and discuss the proportion of validation that is appropriate to mitigate misclassification.

The rest of the paper is structured as follows. Section 2 outlines the Dominick's Finer Foods data utilized, as well as relevant descriptive statistics. Section 3 describes the methodological aspects, such as the Monte Carlo simulations and the measures of bias, variance and RMSE used. Section 4 demonstrates the results for each simulation. Section 5 discusses the findings and what they mean for NSOs, followed by a conclusion that outlines limitations and topics for future research in section 6.

2. Public scanner data

2.1. Overview

The publicly available Dominick's scanner dataset is used throughout the study (Chicago Booth School for Marketing 2013). The dataset is recommended for NSOs as a benchmark for evaluation of methods (Mehrhoff 2019), and has been leveraged by several studies either directly (Lamboray 2021) or as a source to create synthetic data (Office for National Statistics 2020). The data is available pre-categorized on the Chicago Booth School website[3], and thus a simulation of misclassification can be done by picking a few categories, joining them together to simulate a realistic dataset, and introducing misclassification between a target category and the other categories.

Table 1: Number of new products entering the sample per year

Year	Number of New Products
1990	453
1991	287
1992	1393
1993	312
1994	626
1995	25
1996	279
1997	18

Transformation of the data akin to (Lamboray 2021) and (Mehrhoff 2019) was carried out. Weekly data was first transformed into monthly data by assigning a weekly file to a specific month based on when the week starts. Item code and not the UPC was used to define the individual products and calculate unit prices. While 88 months of data

[3] We noticed some minor mistakes in the classification whereby a few products with the same UPC and product name were present in more than one category file. While this would need to be cleaned up if it were a true production use case, we did not correct these mistakes as it should not seriously impact the simulations carried out. Thus, UPCs in different categories were treated as separate products in our study.

is available in the original dataset, consistent churn in products was detected during the first 6 years (December 1989 to December 1995), whereas the number of new products introduced decreased dramatically after 1995, as evident in table 1. As our purpose was to simulate the situation on realistic data, we thus focused specifically on the first 6 years. December 1989 is the base period for all experiments.

2.2. Descriptive statistics and categories chosen

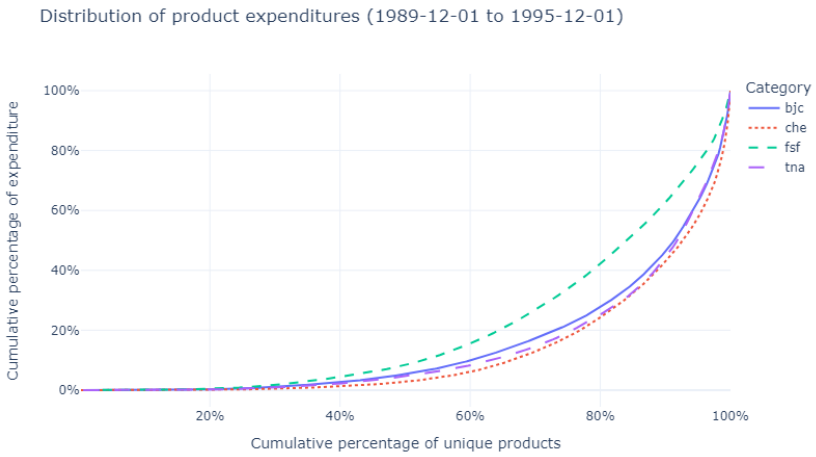
To choose the categories included in the experiments, the number of transactions per month as well as the number of unique matched products present between months throughout the selected period (December 1989 to December 1995) were calculated by category. Any categories with no matched products in any single month were not selected to make certain an index could be calculated each month as more complex processing approaches, such as due to seasonality, would be necessary. Similarly, any categories with a limited number of matched products were not selected to ensure a considerable number of relatives were included in each month's index calculation. Table 2 displays the total number of unique products, as well as the average number of relatives included in each index calculation by category between December 1989 and December 1995. Four categories were selected from those that met the criteria for number of matched products present between months:

Table 2: Unique products and relatives per month

Category	Total Number of Unique Products	Average Number of Relatives per Month
Bottled juices (bjc)	383	141
Cheese (che)	523	239
Fabric softeners (fsf)	159	68
Canned tuna (tna)	191	106

Similar to other scanner data (de Haan, Opperdoes and Schut 1999; Antoniades 2017), Dominick's data for these four categories showed considerable skewness. Figure 1 shows the inequality in the product expenditures, demonstrated by a Lorenz curve. Figure 1 plots the cumulative percentage of the total number of products against the cumulative percentage of total expenditure over the selected period (December 1989 to December 1995) by category. The products are sorted on the horizontal axis in ascending order in terms of expenditure. A Lorenz curve lying on the diagonal would indicate equal expenditure between products. Each of the plotted Lorenz curves are right-skewed, indicating that product expenditures are distributed unequally within the selected categories. A key implication of this distribution is that for weighted price indices, a cut-off sample that validates the highly weighted products should mitigate misclassification in the final index.

Figure 1: Lorenz curves of percentage of cumulative expenditures vs cumulative number of unique products for selected categories (as defined in Table 2) in the Dominick's dataset



3. Methods

3.1. Monte Carlo simulation

As it is challenging to measure the sampling distribution on a single random sample (de Haan, Opperdoes and Schut 1999), it is challenging to model the exact misclassification distribution and probabilities from a single case. A classifier will typically be trained to assign products in a multi-classification scenario (i.e. more than binary case), and typically performs differently across the classes (in this case product categories). As previous findings show, some categories are easier to classify and others are harder, even to the point that annotators may struggle to differentiate products (Greenhough 2022; Spackman, et al. 2023). As classification is usually done on different variables than those that are used in price indices (product name and description for classification versus share of sales and unit price for price indices) and there is no known relationship in the literature between description quality and quantity of sales for this product, it is thus quite possible for a classifier to misclassify a highly weighted product as equally as it would a low weighted one. Furthermore, different classifiers and different natural language preprocessing steps could lead to slightly different decision boundaries for trained classifiers. Finally, data drifts over time and the classifier may perform differently month to month and in general is known to decline in performance over time (Spackman, et al. 2023). It is thus virtually impossible to know all possible cases even for one dataset.

To simplify, we thus estimate the distribution of the calculated price index, by carrying out Monte Carlo simulations. For the bilateral experiments (RQ1), we first select a category of interest and a specific precision and recall set $\{pr, re\}$ as well as a two time periods for calculating the index. All products present in both time periods are selected as the pool of products from which to sample. For each iteration k, we simulate misclassifications to the selected category. To do this we first select a fraction of products labeled to that category from the pool, by simple random sample without replacement, equal to the chosen recall level. Next, we select a fraction of products from the pool that are not in the category of interest, again by simple random sample without replacement, such that the precision of the selected products is equal to the chosen value. The index is then calculated as if the category of interest consisted of only the selected products. We repeat for K iterations, in this instance 5000, to obtain a distribution of indices for the selected category, $\{pr, re\}$ set and time periods. This experiment was repeated for multiple categories, $\{pr, re\}$ sets and time periods.

For experiments that leveraged the multilateral index over time (RQ2 – RQ5), the above method was extended slightly to focus on the aforementioned four categories of interest for the whole 6-year time period at once,

simulating a "steady state" level of misclassification. Note, this also allows a production-like simulation, as a misclassified product stays misclassified for the entire time it is in sample, akin to NSOs classifying and choosing whether or not to validate new products but not re-evaluating existing products.

Each experiment is again conducted as a specific $\{pr, re\}$, this time over the 6-year time period. The pool of products in this experiment is all products with sales transactions, in that time period. For each iteration k, a fraction of products in the four categories of interest, equal to the recall rate are selected from the pool of products by random sample without replacement and kept in their correct category. Next, products are randomly selected without replacement from the remaining pool, and assigned to each of the four categories, such that the precision of each category is equal to the chosen precision. The final results are that each of the four categories of interest have misclassifications at the desired levels, with the rest of the products in that time period left in another, non-interest category. The multilateral index is then calculated for the four categories of interest over the time period. This is then repeated K times, in this case 500, to obtain K different multilateral indices for each of the four categories.

This simulation is meant to roughly match a true production scenario, whereby each product can only be classified to a single category, and a misclassification of a single product impacts two categories, the true category and the one it is misclassified to. For the multilateral simulations, the precision and recall are targeted for the categories of interest only; the remaining products may have different levels.

Figure 2: GEKS-Törnqvist (CCDI) index for selected categories in the Dominick's dataset

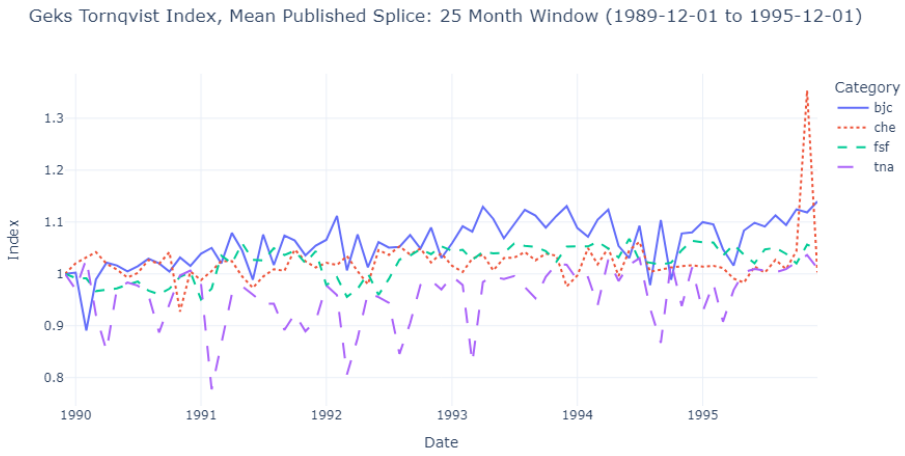


Table 3: Average of monthly price movements for selected categories in the Dominick's dataset

Category	1990	1991	1992	1993	1994	1995
Bottled juices	0.24%	0.41%	-0.08%	0.81%	-0.14%	0.49%
Cheese	0.13%	0.09%	0.26%	-0.50%	0.40%	0.55%
Fabric softeners	0.04%	0.38%	0.12%	0.01%	0.11%	-0.12%
Canned tuna	0.28%	-0.51%	0.85%	0.66%	0.26%	0.09%
Other	-0.10%	0.31%	0.25%	0.17%	0.46%	0.04%

Throughout the selected period, the true indices of the chosen categories, plotted below in figure 2, exhibit differing movements, both in direction and magnitude. To highlight this further, the average monthly movements of the

indices by year and category are displayed in table 3. In addition, the average monthly movement for an index calculated using all products from the remaining 24 categories in the Dominick's scanner data set is also presented in table 3 with the label "other". This provides insight into the price movements of the other products that are potentially being selected for the purpose of misclassification in the experiments.

3.2. Measures tracked for each experiment

Adopting the measures used by (de Haan, Opperdoes and Schut 1999), we construct measures for bias, variance, and RMSE. Specifically, for each set of precision and recall $\{pr, re\}$, $\hat{P}_{k,i}$ denotes the price index, either for the single period bilateral or the at the end of the 6 years for the multilateral case. Finally, k denotes a Monte Carlo replicate or a single iteration of the experiment for category i. Thus:

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

is the estimate of the expected value of the price index, and thus:

$$(2) \quad B\left(\hat{P}_i\right) = \overline{\hat{P}_i} - P_i$$

is the Monte Carlo estimate of bias for a given set of precision and recall $\{pr, re\}$. Similarly:

$$(3) \quad V\left(\hat{P}_i\right) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{P}_{i,k} - \overline{\hat{P}_i})^2$$

is the Monte Carlo estimate of variance, and

$$(4) \quad \widehat{RMSE} = \sqrt{B\left(\hat{P}_i\right)^2 + V\left(\hat{P}_i\right)}$$

is the Monte Carlo estimate of root mean squared error (RMSE) of $\hat{P}_i$.

F_1 score, a commonly used metric for classification tasks which is evaluated in RQ3, is defined as the harmonic mean of precision and recall. More generally the F_β score can be expressed as:

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

Adjusting the value of β, can put more emphasis on precision or recall; for instance, $\beta < 0$ gives more weight to precision, whereas $\beta > 0$ gives more weight to recall.

3.3. Evaluation of metrics and mitigating misclassification

Monte Carlo simulations are run on a set precision and recall $\{pr, re\}$ that are likely to be seen by NSOs that are training ML classifiers. Table 2 summarizes the categories explored. While evaluating the results will provide considerable insight on the discussion of the trade-off between precision and recall directly, to address RQ3, we also compile the F_1 score, and evaluate the bias, variance, and RMSE for a set of scenarios where the metrics are set to identical results. These results can be found in the *Results* section for a single category for simplicity, as well as in full within the appendix.

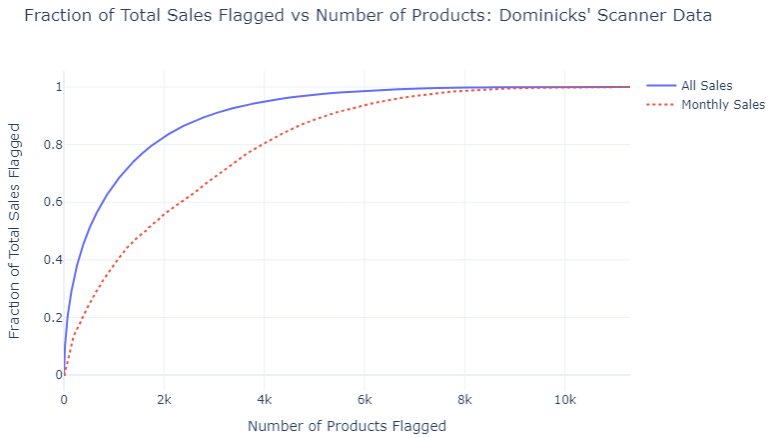
To address approaches to validation (RQ5), several scenarios are analyzed, and a scanner data-specific method is simulated – a cut-off sample of products based on sale proportion in the overall dataset. Firstly, we evaluate the

impact of validating a random sample, as NSOs should flag a certain proportion randomly each month in order to calculate unbiased performance metrics for the production process (Spackman, et al. 2023). Simple random sample does not need to be simulated and its impact can simply be estimated by looking up the results from a slightly higher precision and recall $\{pr, re\}$ scenario. The added effort to validate unique products is estimated according to the scenario. For instance, with random sampling and a $\{pr, re\}$ = {0.8,0.8} scenario, 10% random sample will result in $\{pr, re\}$ = {0.82,0.82}.

Secondly, we consider model confidence, a popular approach for NSOs (Myklatun 2019; Spackman, et al. 2023). NSO's can use the model confidence scores to select which products to flag for review; flagging and correcting those with predicted confidence below a specific threshold. Because the misclassifications in these experiments are randomly generated, there is no associated model confidence available on which to select products; thus, this method was not able to be directly evaluated in this research. We would suggest that if confidence scores are being used to automatically accept predicted classifications, they should be well calibrated using methods such as described by Zadrozny and Charles (2005).

Finally, given the high performance of cut-off sampling based on weight has been found an effective approach for NSOs leveraging scanner data using either a static or dynamic method (de Haan, Opperdoes and Schut 1999; Eurostat 2017), focusing on high selling products within the category should be a very applicable method to sample products for validation. This method of flagging the top selling products was directly simulated in this research. To simulate this scenario, we first misclassify the categories of interest as before, following which, products with the top *n* percent of sales are corrected to the ground truth values, simulating manual review. The multilateral index is then calculated for the full period of interest. To simplify the simulation from a computation point of view, we leverage sales from the whole 6 years rather than for each individual month, when calculating the fraction of total sales. Using monthly sales to calculate the fractional total sales is appropriate within the research to model a "steady state", however its application include two considerations worth noting. Firstly, flagging the top n percent of total sales on a monthly basis would flag more unique products for review, compared to flagging the top n percent of total sales for the entire time period, as shown in figure 3 below. Additionally, products can have low sales in the first month they are observed and thus wouldn't be flagged, but in subsequent months could meet the flagging threshold. In this case, the product category may be corrected to the correct category in a month other than the first month observed. These impacts of correcting products was not considered in this study but may be relevant in a production scenario and should be further investigated.

Figure 3: Comparison of fraction of total sales flagged vs number of products flagged, monthly and for the entire period of interest.



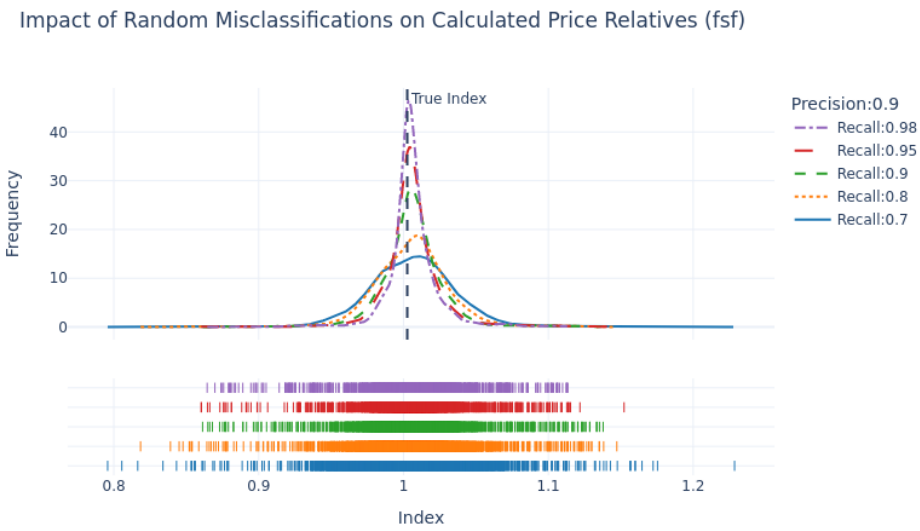
4. Results

4.1. Bilateral simulation results

Results for the bilateral case were generated by running 5000 simulated misclassifications at each desired precision and recall level, for each category of interest. These simulations were conducted on each category for a number of different classification periods; we present the results of the fabric softener category for the periods of January and February 1991 as an example of the results.

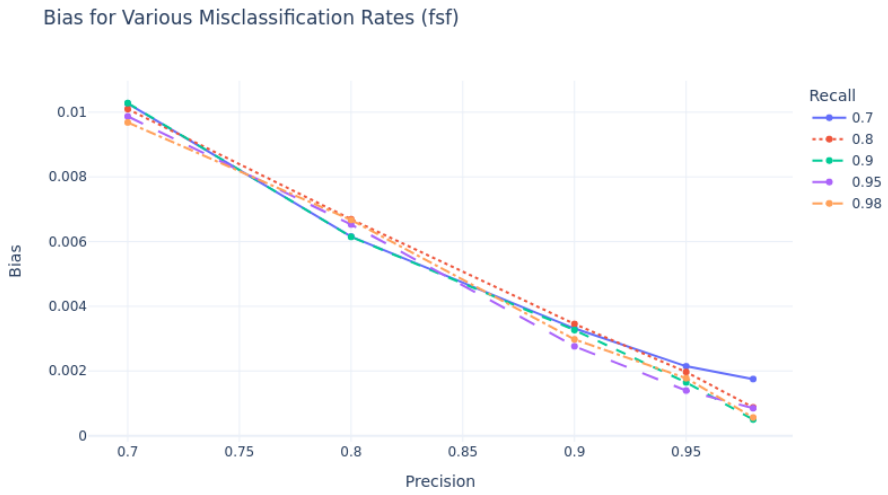
Presented in figure 4 is a distribution plot for the resulting probability distribution of the calculated index, at selected precision and recall rates. Each curve represents the probability distribution for a specific level of category precision and recall with the vertical-coloured lines below representing the calculated index from a single iteration of the experiment. At a fixed single category precision rate, we can see that a higher recall score, qualitatively decreases the variance of the distribution of indices, as indicated by the narrowing of the probability distribution curve. We can also observe that the probability distribution curves, in this example, are all centered around the true price index value, indicating that the bias of the calculated index is relatively constant for the different recall rates at a fixed precision rate.

Figure 4: Probability distribution curve of bilateral Törnqvist index after random misclassifications. 1991-01 to 1991-02



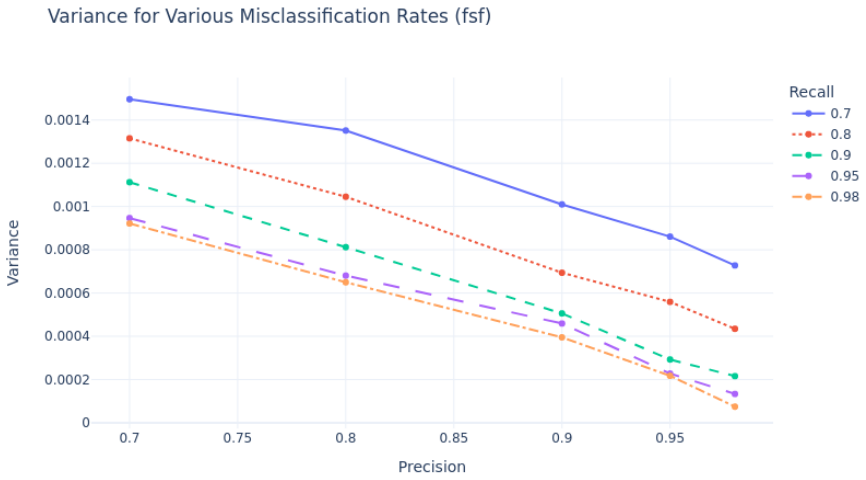
It can be observed, in general, that the absolute value of bias tends to decrease as precision is increased; but is less sensitive to recall, as shown in figure 5. Each line represents a fixed recall level, with precision level varying on the x-axis. The y-axis shows the calculated bias of the estimated index, for the specific $\{pr, re\}$ level. This trend makes intuitive sense as recall is lowered by removing products from the category randomly. This trend of decreasing bias with increasing precision appears to hold for all product categories tested.

Figure 5: Bias of calculated bilateral Törnqvist index after random misclassifications. 1991-01 to 1991-02



Conversely, the variance of calculated indices tended to decrease both with increasing precision and recall. This is highlighted in figure 6. This figure is the same as figure 5, with the variance replacing the bias on the y-axis.

Figure 6: Variance of calculated bilateral Törnqvist index after random misclassifications. 1991-01 to 1991-02

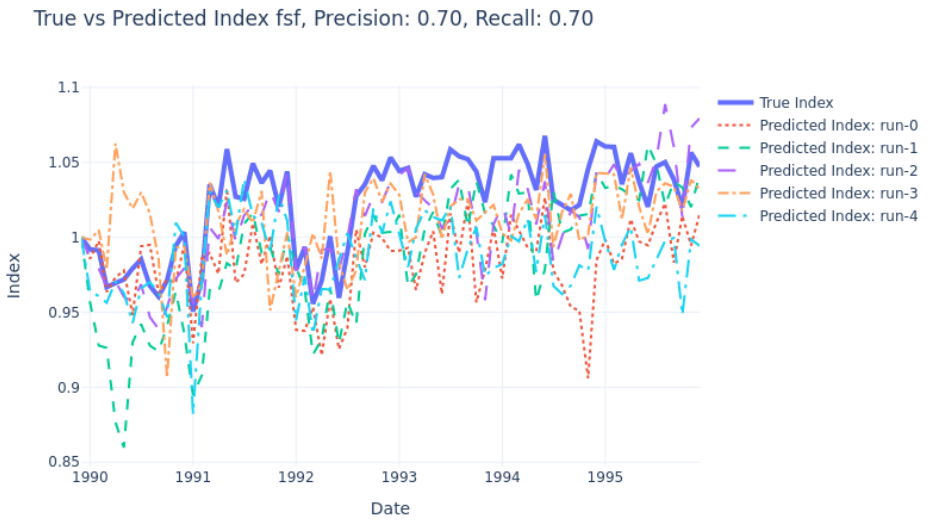


4.2. Multilateral simulation

Results for the multilateral case were generated by running 500 simulated misclassifications at each desired precision and recall level simultaneously for the four categories of interest. The calculation of a multilateral price index over a six-year period is a relatively computationally intense process; as such the number of simulations was restricted to a level that could be run on a single machine in a reasonable amount of time. This restriction was imposed to enable sharing the simulation code with as wide of an audience as possible. As a sensitivity test, we ran 1000 and 2000 simulations at a single $\{pr, re\}$ level and found that, although there were some variations in the results, the mean index calculated at December 1995 varied by less than 0.3 percent from the value obtained in the experiment with 500 simulations. Not only this, but when computing the Geweke diagnostic, values are within the recommended range for convergence, as long as the number of simulations exceeds 100 (Geweke 1991).

To demonstrate the results from the multilateral simulation, we can visualize the potential impact of misclassification on the multilateral index. In figure 7, the true index is compared to five individual simulations at $\{pr, re\}$ = {0.7,0.7} for the fabric softener category. Each line in the figure represents one realization of the random misclassification and the calculated index.

Figure 7: Calculated CCDI for individual misclassification simulations.



Isolating the bias and variance of all simulations shows similar trends to the bilateral experiments. The following two figures show the bias and variance of the index calculated at December 1995 for the fabric softener category. These show similar trends to the bilateral experiment, suggesting that the bilateral experiment could serve as a good proxy for how the index will behave in a multilateral scenario. Figure 8 shows the bias of the calculated index at December 1995 for the fabric softener category; figure 9 shows the variance for the same period.

Figure 8: Bias of calculated CCDI at 1995-12 after random misclassifications.

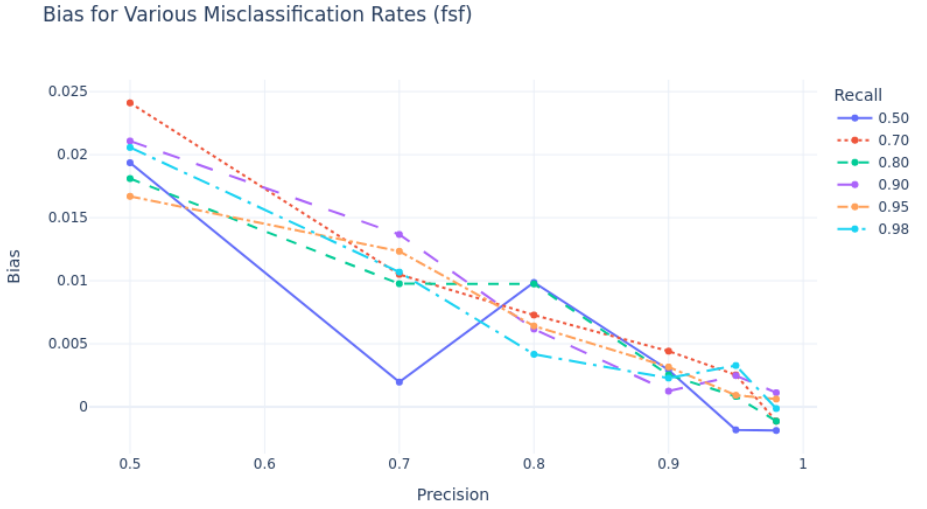
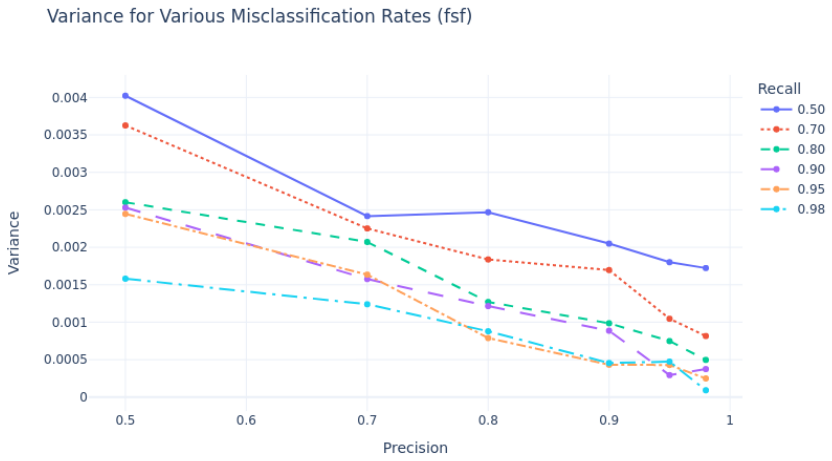


Figure 9: Variance of calculated CCDI at 1995-12 after random misclassifications.



4.3. Multilateral extension method sensitivity

As a comparison, a single additional experiment was run using the half splice on published (HASP) extension method in place of the mean splice on published method used for the rest of the multilateral experiments. The test was run at the misclassification rates of $\{pr, re\}$ = {0.8,0.8}. The true and mean calculated index with misclassifications at December 1995 is presented in table 4 below.

Table 4: Comparison of results for 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

To validate if the estimate of bias varies depending on the extension method chosen, we ran an ANOVA to compare the mean bias from the 500 experiments for the two extension methods: obtaining a P value of 0.082. This result is interpreted to indicate that the bias calculated for the two different splicing methods was not conclusively different thus both methods are similarly sensitive to misclassification.

4.4. Most applicable metrics

Tables 5 and 6 show a summary of results from the multilateral and bilateral experiments respectively. For each of the different precision and recall levels, we show the corresponding F1 score as well as the estimated variance, bias and RMSE values. The first six rows show the results for different precision, with fixed recall at 0.90 whereas the next six rows show the results for different recall levels, with a fixed precision of 0.90.

We can see from the results that F1 score may not be an ideal choice as a metric for measuring classification performance, in the context of classifying products for price index calculations. For instance, in table 5 rows 1 and 7 both have an F1 score of 0.64; however, the bias of the estimated index in row 1 is approximately seven times higher than in row 7. If the goal for NSOs it to minimize bias in the estimated index, it may be appropriate to favor precision over recall, when evaluating classifiers. Providing more weight to precision could be done either by selecting models based on precision score only, or by choosing and F_β score with $\beta < 0$. The results in tables 5 and 6 are for the fabric softener category only, but we see similar trends in other categories, as presented in the appendix.

Table 5: Effect of performance measures with varying recall and precision for multilateral experiment (CCDI with mean pub splice extension method December 1995). All results are for Fabric Softener (fsf)

Row	Summary of Experiment	Category: {Precision, Recall}	Variance	Bias	RMSE	F1
1	Fixed recall, varying precision	fsf: {0.50, 0.90}	0.00253	0.02108	0.05453	0.64286
2		fsf: {0.70, 0.90}	0.00158	0.01367	0.04199	0.78750
3		fsf: {0.80, 0.90}	0.00122	0.00617	0.03540	0.84706
4		fsf: {0.90, 0.90}	0.00089	0.00125	0.02980	0.90000
5		fsf: {0.95, 0.90}	0.00029	0.00248	0.01731	0.92432
6		fsf: {0.98, 0.90}	0.00037	0.00112	0.01939	0.93830
7	Varying recall, fixed precision	fsf: {0.90, 0.50}	0.00205	0.00292	0.04536	0.64286
8		fsf: {0.90, 0.70}	0.00170	0.00442	0.04142	0.78750
9		fsf: {0.90, 0.80}	0.00099	0.00253	0.0315	0.84706
10		fsf: {0.90, 0.90}	0.00089	0.00125	0.02980	0.90000
11		fsf: {0.90, 0.95}	0.00043	0.00316	0.02103	0.92432
12		fsf: {0.90, 0.98}	0.00046	0.00228	0.02146	0.93830

Table 6: Effect of performance measures with varying recall and precision for single reference period experiment, 1991-01 to 1991-02. All results are for Fabric Softener (fsf)

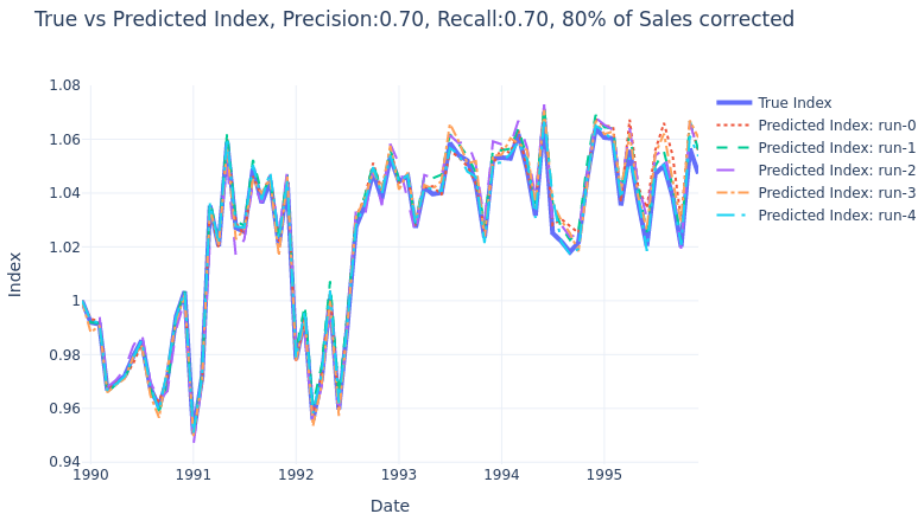
Row	Summary of experiment	Category: {Precision, Recall}	Variance	Bias	RMSE	F1
1	Fixed recall, varying precision	fsf: {0.70, 0.90}	0.00106	0.01004	0.03406	0.78750
2		fsf: {0.80, 0.90}	0.00081	0.00783	0.02951	0.84706
3		fsf: {0.90, 0.90}	0.00052	0.00351	0.02316	0.90000
4		fsf: {0.95, 0.90}	0.00031	0.00108	0.01757	0.92432
5		fsf: {0.98, 0.90}	0.00021	0.00033	0.01462	0.93830
6	Varying recall, fixed precision	fsf: {0.90, 0.70}	0.00103	0.00323	0.0322	0.7875
7		fsf: {0.90, 0.80}	0.00073	0.00304	0.02726	0.84706
8		fsf: {0.90, 0.90}	0.00052	0.00351	0.02316	0.90000
9		fsf: {0.90, 0.95}	0.00042	0.00324	0.02077	0.92432
10		fsf: {0.90, 0.98}	0.00038	0.00389	0.01995	0.93830

4.5. Applying corrections based on product sales

As for the multilateral simulations, the results for this section were generated by running 500 simulated misclassifications at a specific precision and recall level and then applying the correct category to all products in the top *n* percent of total sales. For this experiment we focused on a single level of misclassification $\{pr, re\}$ = {0.7, 0.7}, flagging different fractions of total sales. This level was chosen as it represents a modest level of misclassifications and we wished to have enough misclassifications that the impact of corrections would be easily visualized.

Figure 10 shows the results of five individual realizations of the simulation, misclassified at $\{pr, re\}$ = {0.7, 0.7}, with all products in the top 80 percent of sales corrected. Similar to figure 7, several runs, which are individual realizations of this misclassification and correction process, are show as a demonstration. Compared to the equivalent figure 7, with no corrections, it can be seen that the index tracks much closer to the ground truth in each of the iterations shown. Flagging 80 percent of total sales requires that only approximately 16 percent of the products in the dataset be flagged, or 35 percent if flagging every month (figure 3).

Figure 10: Calculated CCDI for individual misclassification simulations at {pre, re} = {0.7,0.7} *followed by corrections of 80% of sales for fsf category (Fabric softener)*



For the four categories of focus, at lower thresholds of correction, the bias of the calculated index at December 1995 was not always reduced, and for some product categories simulated it even increased or fluctuated (figure 11). While simulation noise could have been a contributing factor, this also suggests that reviewing 80 percent of sales for all categories may be insufficient to guarantee an unbiased index for each category for all time periods. Investigating additional categories in the Dominick's dataset that were outside the main focus of this study did show a trend in decreasing bias by increasing the correction threshold (see Figure A-1 in the Annex). These additional categories were selected with higher average sales compared to the four categories of focus for this study, meaning that a cut-off threshold for correction for the whole dataset was very likely to correct more products in each of these additional categories. As a decline in bias with an increase of the correction threshold is more prevalent in higher sale additional categories, and less prevalent in lower sale categories of focus, this suggests that utilizing a cut-off sales threshold needs to be designed in a way that is most appropriate for the dataset. For instance, combining a correction threshold across the whole dataset with an additional threshold by each category, as well as other methods, may be necessary. Variance consistently decreased at all levels of corrections in the categories of focus. Table 7 provides additional details related to figure 12, showing the mean reduction in variance, compared to the correction threshold applied.

Figure 11: Bias of calculated CCDI at 1995-12 after random misclassifications at $\{\mathrm{pr}, \mathrm{re}\} = \{0.7, 0.7\}$ *followed by corrections of a fraction of total products, based on weight.*

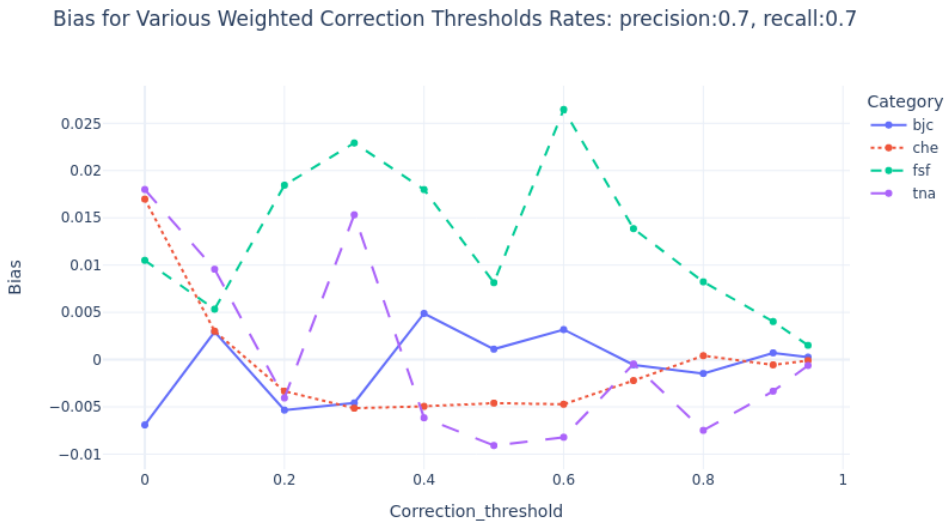


Figure 12: Variance of calculated CCDI at 1995-12 after random misclassifications at $\{\mathrm{pr}, \mathrm{re}\}$ = {0.7,0.7} *followed by corrections of a fraction of total products, based on weight.*

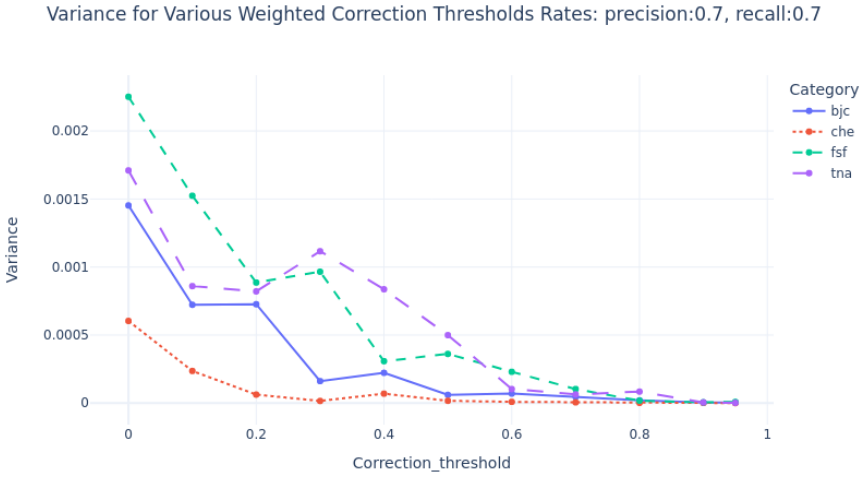


Table 7: Mean reduction of variance of calculated CCDI at 1995-12 after random misclassifications at $\{\mathrm{pr}, \mathrm{re}\}$ = {0.7,0.7} *followed by corrections of a fraction of total products, based on weight.*

Correction Threshold	Mean Reduction in Variance
0.00	0.00
0.10	0.48
0.20	0.63
0.30	0.70
0.40	0.78
0.50	0.87

0.60	0.94
0.70	0.97
0.80	0.98
0.90	1.00

5. Discussion

We find as a result of the bilateral experiments on Dominick's scanner data that even modest levels of misclassification can introduce errors to the weighted Törnqvist index. These errors tend to be proportional to the level of misclassification, with the bias of the estimated index being particularly sensitive to the individual category precisions. When extended to a multilateral index (CCDI), modest levels of misclassification were shown to introduce errors in the calculated index in a similar manner. Though the trends were consistent between product categories, the exact values for bias and variance varies between category, due to the price index calculations dependence on the specific distribution of prices and sales volumes.

When viewing the results of the experiments in aggregate, it is shown that it is important to try to minimize both false positives (increased precision) and false negatives (increased recall) in a production setting. Though both metrics were found to be important, if a trade-off between the two must be made, the precision of the specific category was found to be the most important for reducing the bias of the estimated index, with recall and precision being equally important for reducing variance.

In practice, the errors made by a ML classifier are not guaranteed to be random; the relationship between the specific misclassification levels and evaluated metrics will be more complex. This research provides a useful demonstration of how the index responds to misclassification under specific conditions; future research could evaluate the impact of misclassification under different assumptions.

Comparing two popular extension methods for CCDI, we find the distribution of estimated bias to be equal between the two methods; we suggest that the conclusions from these experiments will be applicable to that extension method. With a relatively low P value of 0.0823 however, further investigation of other index methods and extension methods are warranted.

Applying corrections to the product category, based on the top n percent of sales volumes, was found to be an efficient method, particularly for reducing variance in the calculated price index, and in some cases reducing bias as well. Validating products with 80 percent of total sales volume only required review of 16 percent of total products (35 percent if reviewed monthly) and reviewing products representing 90 percent of sales would require review of only 26 percent of products (46 percent if done monthly). Though selecting products for correction in this way was found to be an efficient option, NSOs should be cautious when selecting a sales threshold to reduce the impact of misclassifications to an acceptable level. Additional approaches such as cut-off thresholds for the whole dataset and each category, as well as other methods, maybe appropriate. Furthermore, it is advisable to combine flagging methods chosen such as cut-off thresholds with a simple random sample by category in order to calculate unbiased performance metrics for the production process. For instance, an 80% cut off combined with a random sample will flag in 24.4 percent of total products (or 41.5 percent if reviewed monthly), and a 90% cut off will flag 33.4 percent of products (51.4 percent if done manually).

In practice, our findings suggest it is valuable for NSOs to run a similar bilateral or multilateral experiment on their own scanner data and chosen price index method, select the desired bias, variance or RMSE, and determine the required category precision and recall required to meet this target. A product review process can then be designed

to ensure that these performance metrics are met or exceeded in each month of production. In practice, these experiments can be done on a few categories rather than for a whole dataset as ML classifiers typically make mistakes between specific related categories; hence such experiments could be guided by an evaluation of the confusion matrix of the ML model being prepared for production. As NSOs may not have access to a multi-year historical dataset, even for a set of representative categories, the bilateral index simulation appears to be a useful substitute for a full multilateral, due to the consistency of results.

It should be noted that the purpose of this study was to model the distribution of possible scenarios and measure bias, variance, and RMSE against a known 'true' index. NSOs may not be able to do this type of analysis in practice, given the requirement of an extended dataset with known labels, and will need to conduct detailed tests and use alternative methods, such as bootstrap methods on random samples every month to estimate bias and variance and to calibrate or correct for misclassification.

In a production scenario, the classification performance metrics must be estimated each month by reviewing a random sample of products, as mentioned above. For a single category, it may be simpler to use validation to increase the precision of that class; for instance, the maximum precision could always be achieved by simply reviewing all products predicted to that category; however, the maximum recall requires the review of all products not predicted to that category.

6. Conclusions and future studies

This paper evaluates the impact of the misclassification of scanner data in the calculation of both bilateral and multilateral price indices. Experiments carried out using the Dominick's Finer Foods scanner dataset using both a weighted Törnqvist bilateral index as well as a GEKS-Törnqvist multilateral index (CCDI), found that, even at limited levels of misclassification, the calculated price indices were affected. In general, the variance of the calculated indices decreased as both precision and recall increased; whereas bias decreased alongside increasing precision and was less impacted by increasing recall. While we note that the impact also extends to the same multilateral index using the HASP extension method, further studies could examine the impact of misclassification using different indices and/or extension methods as well as investigating the impact of non uniform misclassification probabilities.

Experiments were carried out, also using a multilateral GEKS-Törnqvist index, to determine the appropriate number of products to review to minimize the consequences of misclassification. Various percentages of the top sales were used to determine which products' classifications should be corrected across different levels of misclassification. While bias did not always decrease in our case, in general, variance decreased as the percentage of top sales used to flag products needing correction increased. This method of alleviating the impact of misclassification, specifically in terms of diminishing variance, proves to be an efficient option to implement in production dependent on the skewness of the distribution of product expenditures. The evidence of bias in the resulting index, even with relatively high sales thresholds suggest that further investigation is warranted into the recommended sales thresholds for manual review, especially if this flagging method is used in isolation. Under the assumptions of this experiment, ensuring that a minimum of the top 80 percent of products, based on fraction of total sales, are correctly classified, provided a reasonable trade-off between impact on the calculated indices and number of products that would need to be reviewed. Though this may not generalize to all situations, this could be a good starting point from which to evaluate other methods.

In addition, we would recommend randomly sampling and reviewing the category of products predicted to each category, in order to get an estimate each month, for the precision and recall for each class. Any class with low precision should have additional products reviewed and corrected. This study did not assess price relatives to flag

products whose classifications should be reviewed for a number of reasons. Firstly, outlier filtering is a key topic of research due to the sensitivity of multilaterals to dramatic outliers, thus it is likely to be studied well by NSOs as part of their research process. Secondly, price outlier flagging can vary by category in the CPI and the data type, thus focusing on it in this study was not warranted as it would not provide broader guidance. Finally, decomposition methods or analytics by NSO production staff are likely to focus on products with noticeable price relatives even if they are not flagged for review, hence any mistakes in misclassification would likely be caught. However, it may still be a method that is applicable given the specific data and specific use case the NSO is dealing with.

We note that products were sampled by fractional sales, independent of which category they belong to. Ideally one would sample the top sales from each category, however in a classification scenario, we do not know the true class. A future experiment could be to test sampling the top n precent of total sales for each individual category, based on the prediction from the classifier.

This study also maintained consistent levels of misclassification across the whole dataset, including the initial 25 months, to simulate a "steady state" assessment of misclassification. In practice, it may take a while for this "steady state" to occur. NSOs typically have much more time to classify and validate the initial launch window and tend to use ML for production processes. Thus, the proportion of misclassified products will start off very small and gradually increase over time until a "steady state" is reached (Spackman, et al. 2023). Various factors, such as the amount of validation carried out monthly, or the amount of data drift and how often models are retrained, will impact how quickly this "steady state" is reached. Furthermore, the NSOs could use bulk efforts to fix a proportion of the unique products periodically, further offsetting how quickly this "steady state" is reached.

Bibliography

Antoniades, Alexis. 2017. "Distribution as Expenditure." Working paper.

Chicago Booth School for Marketing. 2013. "Dominick's Data Manual."

de Haan, Jan, Eddy Opperdoes, and Cecile M. Schut. 1999. "Item sampling in the consumer price index: a case study using scanner data." *Survey Methodology* 25 (1): 31-41.

Eurostat. 2022. *Guide on Multilateral Methods in the Harmonised Index of Consumer Prices.* European Commission.

Eurostat. 2017. *Practical Guide for Processing Supermarket Scanner Data.* European Commission.

Geweke, John. 1991. "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments." *Feder Reserve Bank of Minneapolis.*

Goussev, Serge, and William Spackman. 2021. "Machine Learning Classifier Evaluation Criteria: application to price statistics." *Statistics Canada International Symposium Series: Proceedings.* Statistics Canada.

Greenhough, Liam. 2022. "Modernising the measurement of clothing price indices using web-scraped data: classification and product grouping." *17th Meeting of the Ottawa Group.* Rome, Italy.

Hov, Kjersti Nyborg. 2021. "Machine learning in the Norwegian CPI: A classification tool." *Group of Experts on Consumer Price Indices.* online.

Lamboray, Claude. 2021. "Index Compilation Techniques for Scanner Data: An Overview." *Group of Experts on Consumer Price Indices.* online.

Manual, Consumer Price Index. 2020. *Concepts and Methods.* Geneva: ILO/IMF/OECD/Eurostat/UNECE/The World Bank, International Labour Office (ILO).

Meertens, Q A. 2021. "Misclassification bias in statistical learning." *PhD thesis.* Amsterdam School of Economics Research Institute, April 28. https://hdl.handle.net/11245.1/4b031bbd-5a46-4181-b0f1-52b38a3b63a6.

Meertens, Quinten. 2021. *Misclassification bias in statistical learning.* Universiteit van Amsterdam.

Meertens, Quinten, Cees Diks, Jaap van den Herik, and Frank W. Takes. 2020. "A data-driven supply-side approach for estimating cross-border Internet purchases within the European Union." *Journal of the Royal Statistical Society Series A: Statistics in Society* 183 (1): 61-90.

Mehrhoff, Jens. 2019. "Promoting the use of a publically available scanner dataset in price index research and for capacity building." *Ottawa Group.* Rio de Janeiro, Brazil.

Myklatun, Kristian Harald. 2019. "Utilizing Machine Learning in the Consumer Price Index." *28th Nordic Statistical Meeting, Helsinki.*

Nietzer, Daniel Helmut Bernd. 2023. *Hierarchical Learning for Classifying Scanner Data and the Consequences of Misclassification for the German Consumer Prices Index.* Internal report, not for publication. Göttingen: University of Göttingen.

Office for National Statistics. 2020. *New index number methods in consumer price statistics: Research into the use of new index number methods to calculate price indices using web-scraped and scanner data.* September 1.

Oyarzun, Javier, and Laura Wile. 2022. "Quality Control of Machine Learning Coding: A Statistics Canada Experience." *UNECE.*

Scholtus, Sander, and Arnout van Delden. 2020. *On the accuracy of estimators based on a binary classifier.* Discussion Paper, CBS.

Spackman, William, Greg DeVilliers, Christian Ritter, and Serge Goussev. 2023. "Identifying and mitigating misclassification: A case study of the Machine Learning lifecycle in price indices with web-scraped clothing data." *Meeting of the Group of Experts on Consumer Price Indices.* Geneva: UN.

UK Statistics Authority. 2023. *Clothing Classification and Product Grouping.* Paper, Technical Advisory Panel on Consumer Prices - Technical.

UK Statistics Authority. 2019. *Guidelines for selecting metrics to evaluate classification in price statistics production pipelines.* Paper, Advisory Panel on Consumer Prices – Technical, UK Statistics Authority, Technical Advisory Panel on Consumer Prices - Technical.

UN Task Team on Scanner Data. 2023. "Classifying alternative data for consumer price statistics: Methods and best practices." *Meeting of the Group of Experts on Consumer Price Indices.* Geneva.

van Delden, Arnout, Joep Burger, and Marco Puts. 2023. "Ten propositions on machine learning in official statistics." *AStA Wirtschafts- und Sozialstatistisches Archiv* 195–221.

van Delden, Arnout, Sander Scholtus, and Joep Burger. 2016. "Accuracy of mixed-source statistics as affected by classification errors." *Journal of official statistics* 32 (3): 619-642.

Van Delden, Arnout, Sander Scholtus, Joep Burger, and Quinten Meertens. 2023. "Accuracy of Estimated Ratios as Affected by Dynamic Classification Errors." *Journal of Survey Statistics and Methodology* 11 (4): 942-966.

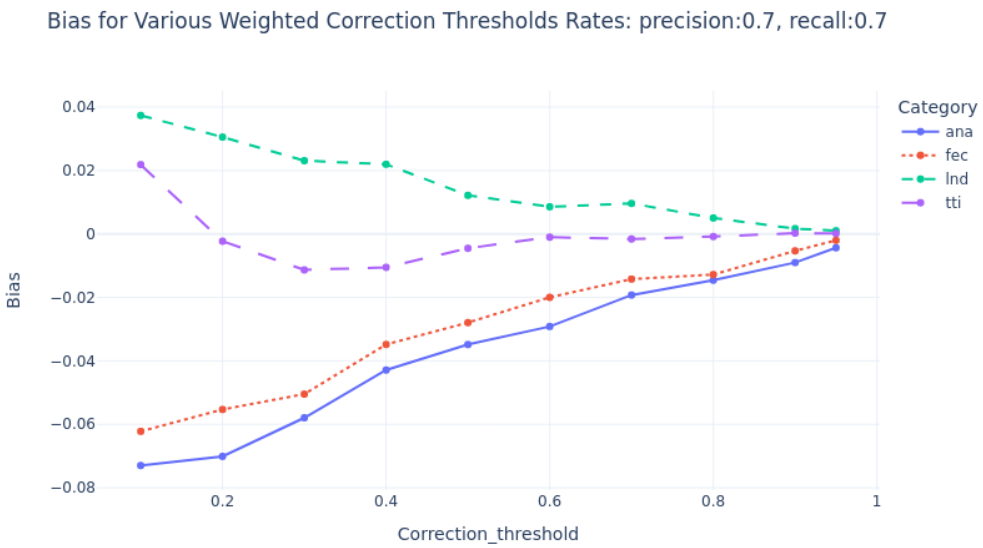
van den Heuvel, Edward P.J. 2019. *Quality of classification of CPI goods and its impact on price indices.* Internal report, not for publication, CBS.

Van Loon, Ken. 2020. *Scanner data and web scraping in the Belgian CPI.* National Academies.

Zadrozny, Bianca, and Elkan Charles. 2005. "Transforming classifier scores into accurate multiclass probability estimates." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* 694-699.

Appendix

Figure A-1: Bias of calculated CCDI at 1995-12 after random misclassifications at $\{\mathrm{pr}, \mathrm{re}\}$ = {0.7,0.7} *followed by corrections of a fraction of total products, based on weight (alternate categories).*



Categories*: Analgesics (ana), Front End Candies (fec), Laundry Detergent (lnd), Bathroom Tissue (tti)*

Table A-1: Varying precision with constant recall for single reference period experiment

Category: {Pr, Re}	variance	bias	RMSE	F1
bjc: {0.70, 0.90}	0.00078	0.00884	0.02927	0.7875
bjc: {0.80, 0.90}	0.00061	0.00652	0.02561	0.847059
bjc: {0.90, 0.90}	0.00032	0.00353	0.01814	0.9
bjc: {0.95, 0.90}	0.00018	0.00191	0.01348	0.924324
bjc: {0.98, 0.90}	0.00007	0.00074	0.0083	0.938298
che: {0.70, 0.90}	0.00026	0.00366	0.01658	0.7875
che: {0.80, 0.90}	0.00024	0.0023	0.01579	0.847059
che: {0.90, 0.90}	0.0002	0.00111	0.01417	0.9
che: {0.95, 0.90}	0.00018	0.00047	0.01328	0.924324
che: {0.98, 0.90}	0.00016	-0.00002	0.01274	0.938298
fsf: {0.70, 0.90}	0.00106	0.01004	0.03406	0.7875
fsf: {0.80, 0.90}	0.00081	0.00783	0.02951	0.847059
fsf: {0.90, 0.90}	0.00052	0.00351	0.02316	0.9
fsf: {0.95, 0.90}	0.00031	0.00108	0.01757	0.924324
fsf: {0.98, 0.90}	0.00021	0.00033	0.01462	0.938298
tna: {0.70, 0.90}	0.00129	0.07239	0.0808	0.7875
tna: {0.80, 0.90}	0.00118	0.0483	0.0593	0.847059
tna: {0.90, 0.90}	0.00095	0.02462	0.03946	0.9
tna: {0.95, 0.90}	0.00078	0.01253	0.03068	0.924324
tna: {0.98, 0.90}	0.00067	0.00554	0.02639	0.938298

Table A-2: Varying recall with constant precision for single reference period experiment

Category: {Pr, Re}	variance	bias	RMSE	F1
bjc: {0.90, 0.70}	0.00041	0.0039	0.02072	0.7875
bjc: {0.90, 0.80}	0.00036	0.00354	0.01926	0.847059
bjc: {0.90, 0.90}	0.00032	0.00353	0.01814	0.9
bjc: {0.90, 0.95}	0.0003	0.00343	0.01753	0.924324
bjc: {0.90, 0.98}	0.00028	0.00337	0.01709	0.938298
che: {0.90, 0.70}	0.0006	0.0007	0.02443	0.7875
che: {0.90, 0.80}	0.00038	0.00127	0.01944	0.847059
che: {0.90, 0.90}	0.0002	0.00111	0.01417	0.9
che: {0.90, 0.95}	0.00014	0.00109	0.01173	0.924324
che: {0.90, 0.98}	0.0001	0.00122	0.00999	0.938298
fsf: {0.90, 0.70}	0.00103	0.00323	0.0322	0.7875
fsf: {0.90, 0.80}	0.00073	0.00304	0.02726	0.847059
fsf: {0.90, 0.90}	0.00052	0.00351	0.02316	0.9
fsf: {0.90, 0.95}	0.00042	0.00324	0.02077	0.924324
fsf: {0.90, 0.98}	0.00038	0.00389	0.01995	0.938298
tna: {0.90, 0.70}	0.00265	0.02766	0.05847	0.7875
tna: {0.90, 0.80}	0.00166	0.02721	0.04898	0.847059
tna: {0.90, 0.90}	0.00095	0.02462	0.03946	0.9
tna: {0.90, 0.95}	0.0006	0.02291	0.03358	0.924324
tna: {0.90, 0.98}	0.00052	0.0244	0.03345	0.938298

Table A-3: Varying recall with constant precision for mean pub splice (500) experiment

Category: {Pr, Re}	variance	bias	RMSE	F1
bjc: {0.90, 0.50}	0.00058	-0.00302	0.02431	0.64286
bjc: {0.90, 0.70}	0.00048	-0.00178	0.02195	0.7875
bjc: {0.90, 0.80}	0.0005	-0.00129	0.02242	0.84706
bjc: {0.90, 0.90}	0.00043	-0.00225	0.02076	0.9
bjc: {0.90, 0.95}	0.00032	-0.00251	0.018	0.92432
bjc: {0.90, 0.98}	0.00036	-0.00256	0.01902	0.9383
che: {0.90, 0.50}	0.00067	0.00563	0.02655	0.64286
che: {0.90, 0.70}	0.00034	0.00458	0.01905	0.7875
che: {0.90, 0.80}	0.00022	0.00554	0.01586	0.84706
che: {0.90, 0.90}	0.00012	0.00508	0.01221	0.9
che: {0.90, 0.95}	0.00011	0.00483	0.01165	0.92432
che: {0.90, 0.98}	0.00011	0.0044	0.01132	0.9383
fsf: {0.90, 0.50}	0.00205	0.00292	0.04536	0.64286
fsf: {0.90, 0.70}	0.0017	0.00442	0.04142	0.7875
fsf: {0.90, 0.80}	0.00099	0.00253	0.0315	0.84706
fsf: {0.90, 0.90}	0.00089	0.00125	0.0298	0.9
fsf: {0.90, 0.95}	0.00043	0.00316	0.02103	0.92432
fsf: {0.90, 0.98}	0.00046	0.00228	0.02146	0.9383
tna: {0.90, 0.50}	0.00243	0.01389	0.05124	0.64286
tna: {0.90, 0.70}	0.0012	0.00639	0.03529	0.7875
tna: {0.90, 0.80}	0.00083	0.00872	0.03002	0.84706
tna: {0.90, 0.90}	0.00073	0.0078	0.02813	0.9
tna: {0.90, 0.95}	0.00056	0.00657	0.02457	0.92432
tna: {0.90, 0.98}	0.00048	0.00531	0.02259	0.9383

Table A-4: Varying precision with constant recall for mean pub splice (500) experiment

Category: {Pr, Re}	variance	bias	RMSE	F1
bjc: {0.50, 0.90}	0.00169	-0.014	0.04346	0.64286
bjc: {0.70, 0.90}	0.00122	-0.00857	0.0359	0.7875
bjc: {0.80, 0.90}	0.00098	-0.00406	0.03162	0.84706
bjc: {0.90, 0.90}	0.00043	-0.00225	0.02076	0.9
bjc: {0.95, 0.90}	0.00019	-0.00135	0.01395	0.92432
bjc: {0.98, 0.90}	0.0001	-0.00057	0.01014	0.9383
che: {0.50, 0.90}	0.00056	0.03029	0.03845	0.64286
che: {0.70, 0.90}	0.00034	0.01655	0.02476	0.7875
che: {0.80, 0.90}	0.00025	0.00895	0.01818	0.84706
che: {0.90, 0.90}	0.00012	0.00508	0.01221	0.9
che: {0.95, 0.90}	0.00011	0.00283	0.01108	0.92432
che: {0.98, 0.90}	0.00007	0.00094	0.00835	0.9383
fsf: {0.50, 0.90}	0.00253	0.02108	0.05453	0.64286
fsf: {0.70, 0.90}	0.00158	0.01367	0.04199	0.7875
fsf: {0.80, 0.90}	0.00122	0.00617	0.0354	0.84706
fsf: {0.90, 0.90}	0.00089	0.00125	0.0298	0.9
fsf: {0.95, 0.90}	0.00029	0.00248	0.01731	0.92432
fsf: {0.98, 0.90}	0.00037	0.00112	0.01939	0.9383
tna: {0.50, 0.90}	0.00205	0.02953	0.05404	0.64286
tna: {0.70, 0.90}	0.00167	0.01718	0.04436	0.7875
tna: {0.80, 0.90}	0.00085	0.01084	0.03115	0.84706
tna: {0.90, 0.90}	0.00073	0.0078	0.02813	0.9
tna: {0.95, 0.90}	0.00041	0.00242	0.02028	0.92432
tna: {0.98, 0.90}	0.00033	0.00302	0.01852	0.9383