Online price data, high frequency inflation measures, and formulation of monetary policy in New Zealand

Alan Bentley, Karsten Chipeniuk, and Frances Krsinich*

June 14, 2024

Abstract

Web-scraped prices provide a potential source of high-frequency, low-lag, granular data for price inflation. However, these data have limitations preventing their widespread use in official statistics. Among these limitations are that the quantities (and hence basket weights) are not directly observable, that the final price may differ from the advertised price, and the potential for data discontinuities. Despite these limitations, web-scraped price data likely has high value for end-users of official inflation statistics such as central banks. For example, inflation is measured at a quarterly frequency in New Zealand, in contrast with higher frequency data available in peer countries. Notably, this implies that inflation data is released four times a year, while the Reserve Bank of New Zealand adjusts its official cash rate 7 times a year.

In this paper, we investigate how producers and users of price data can collaborate to optimally make use of new and innovative, but potentially challenging, sources of information. Specifically, in joint work between the Reserve Bank of New Zealand and Statistics New Zealand, we explore the use of granular, daily web-scraped data. This covers products offered online by 14 major New Zealand retailers across multiple sectors from 2014 to 2022. Combining detailed features of the data such as product categories with state-of-the-art machine learning techniques, we construct high frequency indicators for CPI in New Zealand. We investigate the scope for timely and accurate nowcasting and forecasting, given the low frequency and short overlapping time series from the official inflation measure. Finally, we conclude by considering implications for monetary policy formulation.

1 Introduction

Following a decade of relatively low and stable consumer price inflation, the COVID-19 pandemic and the associated health response forced economic policymakers to grapple with the question of how best to monitor a rapidly evolving situation. On one hand, the sudden implementation and removal of lockdowns raised the question of how to track economic activity in real time. On the other hand, once restrictions were largely lifted and supply side conditions tightened, central banks around the world were forced to rapidly raise interest rates in response to the highest price increases since the 1970s. Through this experience, it has become increasingly clear that having access to high frequency measures of the overall level of prices would be invaluable to central banks in pursuit of a price stability mandate.

One potential source of information on prices which has become increasingly evident in recent years are those retrieved from the websites of online retailers. In an increasingly digitized world, such prices can be observed at high frequencies across a wide range of products, and as a result have captured the attention of statistical agencies as a potential measurement device. However, there are a number of factors which limit the use of online prices in the calculation of official statistics.

For one thing, official consumer price indices are averages which weight the prices themselves by the relative quantities in which individuals actually consume each good, and these quantities are not listed on retailer websites. Additionally, these prices may not include additional factors such as fees, taxes, or sales,

*The views expressed in this article are those of the authors and do not necessarily reflect the views of the Reserve Bank of New Zealand or Statistics New Zealand. The authors would like to thank Christopher Ball, Lucas Chen, Lydia Dudson, and Marea Sing for their suggestions, discussions, and feedback.

and may differ from the actual price paid. There are also products included in the CPI basket which are not necessarily available online, such as a variety of services. Finally, inconsistencies in the data collection process or the listings themselves could result in data which is inaccurate or discontinuous.

Despite these limitations, it remains highly likely that web-scraped price data has uses within an applied context. One such case is that of monitoring inflation in the case of New Zealand, where Statistics New Zealand measures the consumer price index only four times a year, as opposed to twelve times as in many other countries. This creates a dilemma for the Reserve Bank of New Zealand, which is consequently required to make interest rate decisions at three of its seven annual reviews without additional information on how the inflation situation has evolved. Moreover, as seen during and after COVID, inflation may be difficult to observe in extreme economic scenarios (such as lockdowns) and may take off suddenly when demand and supply do not align.

In this work, we explore the use of this new but challenging source of price information. Specifically, we make use of granular, daily web-scrped price data, which covers products offered online by 14 major New Zealand retailers across multiple sectors from 2014 to 2022. We examine the potential for this data to provide reasonable coverage of the CPI basket, finding that, at a high level, the data provide excellent coverage of broad consumer prices, and include 98.2% of the top level of goods categories tracked by Statistics New Zealand when constructing the CPI basket. In terms individual categories, the only ommission is that of education. However, coverage drops significantly, to just 57.3%, when looking at the mid-level classifications, with the drop primarily being due to the exclusion of housing as well as many services from the data set. Looking at the highest resolution categorization drops coverage further still, to just 40.2%. We then observe the potential for this data to provide useful information on micro price setting behavior by online retailers, by comparing the price movements by two of the retailers in the series.

We then outline a simple methodology for making use of the data, given the challenges it presents; namely the large number of observations, the noisiness of the data, and the potential for missing values. This amounts to a simple aggregating of prices based on categorization, resulting in a relatively manageable amount of data while retaining all of the frequency and some of the granularity. The resulting series typically have good coverage of the sample period, even when products and retailers enter and exit the sample at various times. They also display varied dynamics, potentially reflecting the different price setting processes of the underlying products.

Turning to predictive power, we attempt a simple nowcasting exercise, in which we use the aggregates previously computed with a machine learning approach to examine whether the daily data is able to provide an indication of consumer price inflation in a given quarter. We find that the granular data is able to provide reasonable estimates of consumer price inflation prior to the post-COVID inflationary period, and does rise significantly during the post-COVID period. However, it consistently undershoots observed inflation after COVID. Moreover, its nowcasts are less accurate than one period ahead forecasts by a simple AR(1) model, raising a number of possibilities for future work.

The rest of the paper is organized as follows. Section 2 reviews related literature about online prices and their similarity to offline prices, their use in measuring inflation, their use in nowcasting and forecasting, and their use in New Zealand. Section 3 describes the data set used here. Section 4 describes our methodology. Section 5 presents our results and Section 6 concludes.

2 Literature

This paper is related to the broader literature on the use of high frequency, large datasets of product prices, and in particular those obtained online. Early studies in this area which focussed on manually collected prices and a small selection of categories (e.g. Brynjolfsson and Smith (2000), Clay, Ramayya, Wolff, and Fernandes (2002), Cavallo, Neiman, and Rigobon (2014), and Cavallo, Neiman, and Rigobon (2015)). These studies largely compared online and offline prices and their dynamics, finding broad similarities between the two.

More recently, Cavallo and Rigobon (2016) argued for the application of new data gathering techniques which would allow for large sets of prices to be obtained online automatically. Potential advantages of this approach the relatively low cost of data collection, comprehensive coverage of the range of products offered by online retailers, and the wealth of information that can be collected for each good. In contrast

to other economic studies on big data, Cavallo and Rigobon (2016) focused on the use of these techniques to measure conventional economic indicators rather than to forecast, following earlier work of **?** which used online inflation indexes to show that the official inflation rate reported in Argentina from 2007 to 2015 was 12 percentage points lower than the measured rate of 20%. The authors subsequently provide a pair of applications of online price indexes, demonstrating that they display a different distribution of price changes compared to those used in official statistics, and are more consistent with the law of one price than the consensus in the broader literature.

Following the initial case for mass collection of online prices for the purposes of measuring inflation in Cavallo and Rigobon (2016), subsequent work of Cavallo (2017) sought to establish definitively that online and offline prices are directly comparable and that the former can be used in place of the latter for applicationns. Specifically, this study crowdsourced the collection of offline prices alongside the web scraping of online prices across 10 countries. The author found that online and offline prices are identical 72 percent of the time on average across countries, with heterogeneity across sectors and countries. Moreover, price differences, where they occur, tend to be small. These findings provide a large scale basis for the findings of the early studies on online prices, and suggest that online prices can be used as a proxy for those offline. Moreover, they suggest that online prices may be a valid source of information for national statistical offices (NSOs) seeking to better measure price indices in their respective regions.

Subsequent work of Harchaoui and Janssen (2018) examined the daily consumer price index produced by the Billion Prices Project, the commercial venture born out of the work of Cavallo and Rigobon (2016). Noting the decline in timeliness of official statistics given advances in information technology, the authors argue that an appropriate methodological approach combined with the daily statistics can enhance timeliness of the CPI figures. They compare mixed data sampling (MIDAS) methods incorporating the daily CPI to a simple autoregressive forecast using only monthly CPI, finding reductions in the average relative root mean squared forecast errors at a 12 month horizon, as well as statistically significant directional accuracy.

Web scraped prices have likewise been examined for measurement of price indexes outside of the United States. For example, in the European Union, the Prices-setting Microdata Analysis Network (PRISMA) was recently established to explore the use of a variety of microdata on prices, including those online (Osbat (2022)). They note that there are disadvantages as well as advantages to this source of data, and that collection and classification can be difficult, particularly given the multilingual environment which makes up the Union. Nonetheless, they highlight the potential for a wide range of applications, including the potential to better understand the price setting process, nowcast, measure, and monitor inflation, and complement online prices with scanner data to understand how individuals experience inflation.

Looking at a single member of the Union, Jaworski (2021) considers the particular case of food price inflation in Poland during the COVID-19 pandemic, a period during which lockdowns hindered or even prevented traditional price collection methods. Noting increasing attempts by NSOs to use online prices in official statistics (e.g. Breton, Clews, Metcalfe, Milliken, Payne, Winton, and Woods (2015), Horrigan (2013), Krsinich (2015), and Nygard (2015)), the author web-scapes food prices and computes real-time daily index. He finds that the online indices are reliable estimates of month over month and year over year inflation up to 30 days prior to the release of official statistics.

This paper is also related to a strand of literature which aims to nowcast or forecast using online data with a variety of more advanced statistical tools, rather than to measure an index. One such study is Aparicio and Bertolotto (2020) which uses online prices for ten major economies in a regression specification including the CPI and offline fuel prices. They find that this specification is able to provide lower root mean square forecast errors one, two, and three months ahead than survey forecasts or simple empirical specifications for many of the countries considered. Similarly, returning to the context of food prices specifically, Macias, Stelmasiak, and Szafranek (2023) estimate ten different empirical nowcasting specifications, both with and without online prices. They find that frameworks with online prices can provide better nowcasts than baseline specifications without them, and that careful curation of the data and selection of the model can provide additional benefits.

Alongside the above, econometrically conventional specifications, literature continues to expand on the use of recent advances in machine learning technology. One such example, Almosova and Andresen (2023), compares US inflation forecast errors between conventional econometric time series models, a simple neural network (NN), and a long short-term memory recurrent (LSTM) neural network. They find that the LSTM outperforms an autoregressive specification, a random walk, and the NN model on nonseasonally adjusted

data and is on par with a seasonal ARIMA model. They find that all of the specifications they consider display similar performance on seasonally adjusted data, apart from a random walk, which performs relatively poorly. In the context of economic activity, Qureshi, Chu, and Demers (2020) examines the use of extreme gradient boosting (XGBoost) with Google Trends and Statistics Canada data to forecast GDP in Canada. They find that gradient boosting improves monthly real GDP forecast errors relative to an AR(1) specificationwhen only official data is used, and that further gains are provided by using XGBoost with Trends data. They further find that Trends data on Finance, Property Development, and Roleplaying Games are the feautres most useful for forecasting real GDP.

Finally, this paper is related to the nine year body of literature which has aimed to make use of online prices in the New Zealand context. Krsinich (2015) followed up on theoretical literature by the same author to apply the fixed-effects window-splice (FEWS) index to online prices. The FEWS method exploits fixed effects to control for unobserved product characteristics while the window splice accounts for the introduction or removal of products. The author finds that the FEWS index for several product categories closely matches the result given by a chained Jevons index for several products at daily and monthly frequencies, and that including weight quantities from scanner data provides little improvement in the majority of categories. Building further on this work, Bentley and Krsinich (2017) provides the high level case for the move towards the adoption of big data, and online prices in particular, in the measurement of the CPI, and also gave an application of the use of administrative data to measure rent price changes.

Lynch, Olivecrona, and Stansfield (2019) applied online food prices with the FEWS approach to compute multiple food price indexes with varying methods of aggregation, comparing a number of online indexes with Statistics New Zealand's conventional published index. They found that aggregating prices based on the retailer menu provided the closest comparison to the published index. Finally, Stansfield and Krsinich (2021) provided a further update on Statistics New Zealand's progress towards incorporating online, scanner, and administrative data into their price measurement activities. While the earlier literature in the New Zealand context emphasized usig online prices for measuring price inflation, in this work we explore the direct application of the underlying granular data by the end user.

3 Data

We use granular daily price data, which includes prices of products scraped from the webpages of 14 major New Zealand retailers between January 1, 2014 and September 25, 2022. In addition to the prices themselves, the data set includes an identifier for each product, the name of each product, the date on which the price was observed, and approximate Classification of Individual Consumption According to Purpose (COICOP) categories. These categories in turn loosely correspond to categories used by Statistics New Zealand when applying basket weights to compute the CPI.

Table 1 reports basic summary statistics about the data set as a whole, across all retailers, products, and dates. The total number of observations in the raw data set is close to 509.5 million, which as previously mentioned are spread across 14 retailers. There is significant heterogeneity in how observations are spread across retailers, not all of whom have data across the full sample period. Specifically, the maximum number of observations for a single retailer is about 92.6 million; were every retailer have this number of observations, the total number of observations would be well over 1 billion. Likewise, were every retailer to only have the minimum number of 3.8 million, then the full data set would only contain 53 million observations.

Turning to the distribution of prices in the full dataset, we find a mean price of $171.53. We find that prices vary greatly around this amount, with a large standard deviation of 845.91. Unsurprisingly given the low mean and high variance, prices are highly skewed to the more expensive side, with a standardized skewness of 2284.85. Despite this, the distribution has very thin tales, with the kurtosis being just 2.7, a value less than that of a normal distribution (3.0).

Finally, we examine how well the data covers the full range of consumer products. Of the 11 possible top layer COICOP categories, 10 appear in the data set, representing 98.2 percent of the full 2020 CPI basket. However, at a more granular level, the data set only covers 26 of the 43 possible categories in the middle layer of the classification scheme, and in turn this only represents 57.3 percent of the full 2020 consumption basket. At the most granular level, this drops to 56 out of 109 categories, a further drop to 40.2 percent of the basket. Moreover, we do not report the proportion of products in each of the categories which appears.

Statistic	Value
Number of Observations	509455833
Retailers	14
Max. Obs.	92640599
Min. Obs.	3799640
Mean Price	171.53
St. Dev. of Prices	845.91
Skewness of Prices	2284.85
Kurtosis of Prices	2.70
Level 1 Categories	10
Level 2 Categories	26
Level 3 Categories	56

Table 1: Summary statistics for the full data set.

It is possible that some categories are overrepresented and others underrepresented in the data set.

Table 2 reports the level 2 categories which fail to appear in the data set, as well as their corresponding level 1 category. Two categories at level 1 are not missing any subcategories at level 2 at all, namely Clothing and Footwear and Communication. On the other hand, the level 1 category of Education does not appear in the data set at all, and consequently all four of its level 2 subcategories fail to appear.

Turning to the remaining categories, we make a number of observations. Apart from Education, most level 1 categories are only missing one or two level 2 categories, and these are largely services. This latter observation is to be expected, as the data collection largely focused on scraping the websites of major retailers, and not prices listed online by service providers. However, some services do appear in the webscraped data, such as postal and telecommunication ones. Notably, alcoholic beverages do appear in the data set, as beer and wine is available in New Zealand via delivery from grocery chains, but cigarettes and tobacco are more tightly controlled and hence do not appear online. Finally, housing costs such as rents, new housing purchase, and property rates are not captured in the data.

Having described some general features of the data, we next turn to the empirical methods that have been applied to investigate uses of this data to better understand pricing dynamics and inflation in New Zealand. However, we note that there are likely to be many other broad features of interest. These include, but are not limited to, summary statistics broken down by retailer, sample coverage by retailer, and summary statistics for the distribution of price changes.

4 Methodology

In order to make use of the data set, we must methodologically address a number of challenges that come with working with it. While the breadth and frequency of the data raises its potential value, it also increases the memory requirements for storing it to the point that the full data set for a single retailer fills the RAM of a 15 GB laptop. Additionally, retailers enter and exit the sample at different times, as do products. Finally, even when products are included, some fields such as classification may be missing.

In order to address the size of the data set and make it manageable to work with across all retailers while retaining the high frequency element, we first aggregate products on each day according to their level 2 COICOP classification. Specifically, we compute the average price for goods within each individual across all retailers on each date t. By doing so, we limit the number of observations to the number of level 2 categories multiplied by the number of days in the sample, leaving us with just 63,800 values. Where a category does not appear on a given day, it's value is left as unobserved.

We then perform a simple nowcasting exercise using extreme gradient boosting (XGBoost). This is a

Level 1	Level 2
Food	Restaurant meals and ready-to-eat food
Alcoholic beverages and tobacco	Cigarettes and tobacco
Housing and household utilities	Actual rentals for housing Home ownership Property rates and related services
Household contents and services	Other household goods and services
Health	Out-patient services Hospital services
Transport	Passenger transport services
Recreation and culture	Recreational and cultural services Accommodation services
Education	Early childhood education Primary and secondary education Tertiary and other post-school education Other education
Miscellaneous goods and services	Insurance Credit services

Table 2: Level 1 and 2 categories which do not appear in the data set.

machine learning algorithm which has gained popularity in the decade since its introduction in 2014. At the base of this algorithm are simple decision trees, with nodes that split based on the value of one of the features in the data set. XGBoost combines the predictions of several such trees, adding them one by one to minimize the errors made by previous ones via gradient descent. The net result of this procedure is an algorithm which is able to learn nonlinear relationships between input data and a target variable by combining the outputs of several much more primitive nonlinear objects.

To do our exercise, we first collect the observations of our categorical price averages according to the quarter in which the corresponding date falls. This results in 1,800-1,840 features, depending on the quarter. In order to make the number of features consistent for all observations of the CPI, the extra 20 or 40 features are treated as unobserved in quarters with 91 or 90 days, respectively. The explained variable is then the year-over-year percentage inflation rate, available quarterly from Statistics New Zealand.

We initially train the model on the first 16 observations for CPI inflation. We then construct expanding window nowcasts, using the adding the daily category price aggregates derived from the data to predict that quarter's CPI inflation outturn. We then add the category price aggregates and the inflation value to the training set, and repeat the exercise. At each step, we compute the absolute error between the out-of-sample prediction and the observed value to evaluate the model.

5 Results

We now present the results of the empirical exercise described at the end of the previous section. First, we examine the behaviour of the price indexes computed for each level 2 COICOP category.

Table 4 gives the numerical values of the average price of goods aggregated into categories across all dates for which there is an observation for that category, as well as the corresponding standard deviations. The results are broadly in line with expectation, in that categories containing large durables, such as furniture and appliances, tend to be more expensive than smaller items such as clothing or consumables. The most volatile products are those related to fuel, although this appears to primarily come from extreme movements in late 2021 and early 2022, at the end of the series. The least volatile are food products, with the exception of Postal services for which there is just a single item in the data set for much of 2018, whose price never changed.

Figure 1 plots a number of these aggregated daily price series throughout time, all of which are available for most of the sample period. The series show distinct dynamic properties, with some such as alcoholic beverages and household appliances having large outliers, and others such a glassware, tableware and household utensils appearing to be more stable. Food prices in particular papear to stabilize following the initial part of the series, possibly due to the data collection procedures maturing. Several of the series display clear upward trends following the COVID-19 pandemic, which was a period of historically high consumer price inflation.

Next, Figure 2 shows graphically the results of using these daily averages to nowcast year-over-year consumer price inflation using XGBoost. The plot shows both the target variable, the observed inflation rate, as well as the value predicted out of the machine learning model from the quarter's daily level 2 price aggregates. The model appears to perform reasonably well early in the sample, during a period in which inflation was low and stable. Notably, this period is relatively similar to that which the model was trained on, namely the 16 quarters from 2014Q1 through 2017Q4.

While the baseline model does pick up an increase in CPI inflation in the historically high inflation period which occurred following the COVID-19 pandemic, it nonetheless consistently undershoots the observed value beginning in the first quarter of 2021. Moreover, the model prediction begins to decline as CPI inflation reached its peak. This suggests that, despite several of the series showing an upwards trend during this period in Figure 1, these trends are not sufficient for the model to anticipate the degree of passthrough to broader consumer prices.

There are a few potential reasons for this mismatch. For one, the data set does not extend back to a previous period of similarly high inflation, and may require further observations of the target variable for training purposes. For another, this inflationary episode was initially driven by supply side factors such as shipping and petrol prices, goods which do not appear in our data set. Likewise, it cannot account for price changes for a broad range of services, as well as new housing and rents, over this period.

Next, we consider the quantitative performance of the model, with results shown in Table 4. As seen

Category	Average Price	Std. Dev. of Price
Alcoholic beverages	18.63	4.54
Audio-visual, photographic and information processing equipment	360.61	102.26
Clothing	37.52	22.77
Electricity, gas and other fuels	167.93	503.64
Food	5.77	0.50
Footwear including repair	74.92	46.97
Furniture and furnishings, carpets and other floor coverings	1234.31	291.51
Glassware, tableware and household utensils	24.53	5.28
Goods and services for routine household maintenance	19.34	5.11
Household appliances	867.06	176.33
Household textiles	51.98	15.63
Maintenance and repair of the dwelling	111.75	80.81
Medical products, appliances and equipment	19.80	7.95
Newspapers, books and stationery	27.99	14.68
Non-alcoholic beverages	5.91	1.37
Operation of personal transport equipment	36.81	32.42
Other major durables for recreation and culture	285.77	238.73
Other recreational items and equipment, gardens and pets	51.79	23.10
Other services n.e.c.	13.90	1.90
Personal care	34.64	14.77
Personal effects n.e.c.	137.46	75.36
Postal services	115.99	0.00
Purchase of vehicles	474.61	416.94
Telephone and telefax equipment	291.84	69.34
Telephone and telefax services	21.48	2.69
Tools and equipment for house and garden	47.63	17.40

Table 3: Average prices and standard deviation of prices aggregated by COICOP level 2.

graphically above, the model performs relatively well prior to 2021, with nowcasts which are off by 0.49 percentage points, but poorly thereafter, with nowcasts that are off by 2.65 percentage points. As a result, across the full sample the the nowcast error is 1.17 percentage points.

While these results indicate that the high frequency price data may have some indicative power for broader consumer prices, particularly outside of times when goods not available online are behaving very differently online, the nowcast errors are much worse than the one period forecasts from a simple AR(1) estimated from 1990Q1 to 2018Q1, as shown in Table 4. This is largely mitigated by including the lag of the CPI as a feature which XGBoost can use to predict the next period outturn, however the errors are still higher in this case, despite the fact that it is within XGBoost's ability to mimic an AR(1) model. Further work will be required to determine what numerical advantages are offered by high frequency web scraped price data.

6 Conclusion

In this paper, we have examined the use of high frequency, granular, web scraped prices for better understanding inflation dynamics in New Zealand. New, rich sources of data such as this have recently been explored by academics and officials at National Statistical Offices for the purposes of measuring consumer price indexes as well as forecast and nowcast inflation. This includes a decade of work by Statistics New

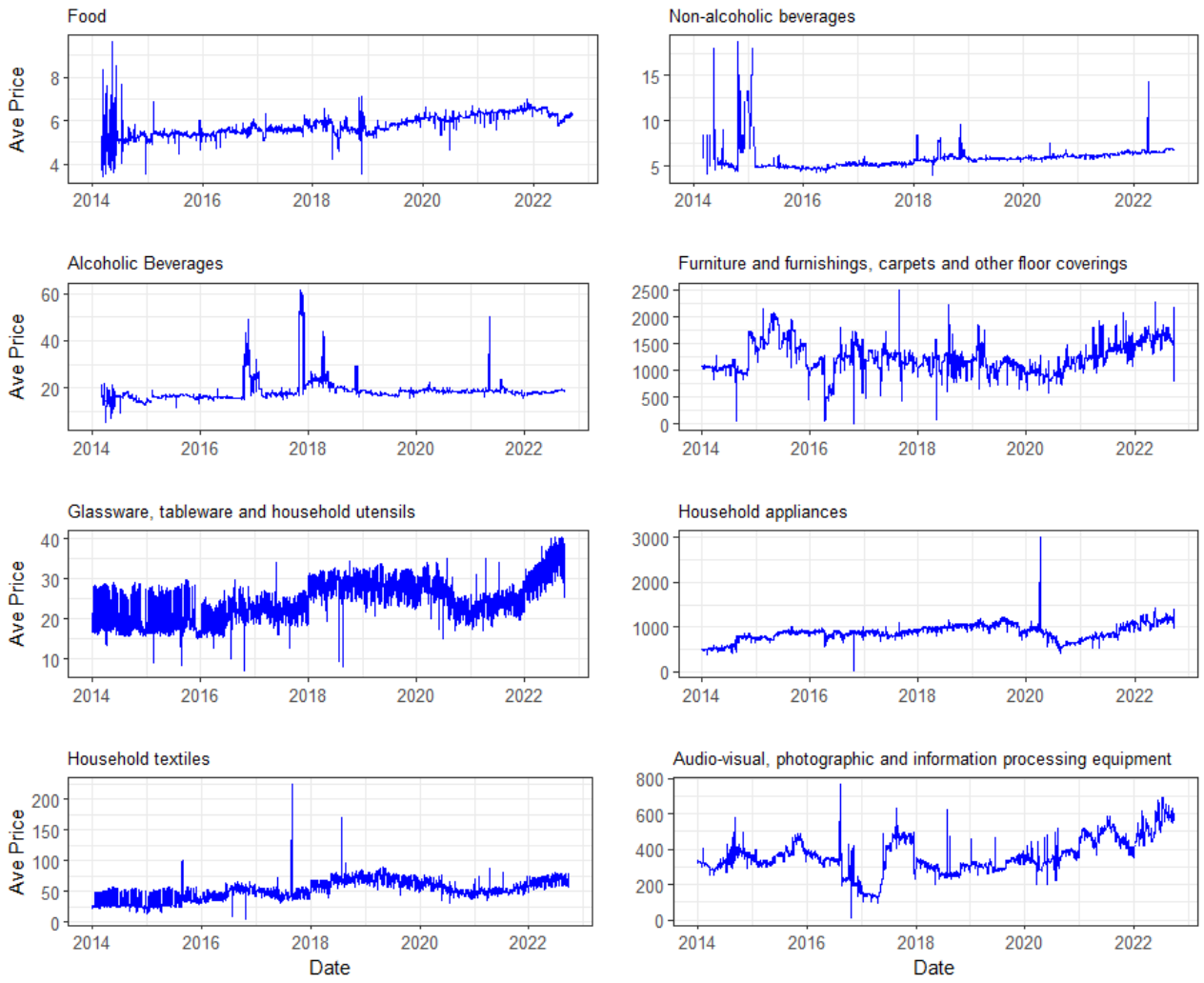

Figure 1: Aggregates of prices into eight of the twenty level 2 COICOP categories.

Zealand to incorporate these prices into their own indices, and now this work which involves an end user of inflation data.

We noted that the coverage of the dataset at a high level appears good, but that this coverage drops at more granular levels of classification due to the omission of services and housing-related prices. We noted potential issues with using this data, such as noise, missing values, and entry and exit of products and retailers. We also noted the potential advantages, such as the potential for timely indicators or a better understanding of micro price setting behavior.

We then outlined a simple scheme to aggregate the information into a smaller set of features while retaining the high frequency element and some of the granularity of the data, by aggregating prices across the mid-level classification provided in the data. We observed that the resulting series had broadly sensible statistical properties and distinct dynamic properties. Moreover, we observed that some of them began to show a clear upward trend following the COVID-19 pandemic.

Finally, we concluded by attempting a simple nowcasting exercise with XGBoost, to see the extent to which the Price Stats data can pinpoint broad consumer price inflation. In this initial exercise, we found that the granular data is able to provide a rough approximation to CPI, particularly prior to the supply driven inflationary episode which began in 2021. However, we found that the estimates it provides are inferior to one period ahead forecasts using an AR(1) model.

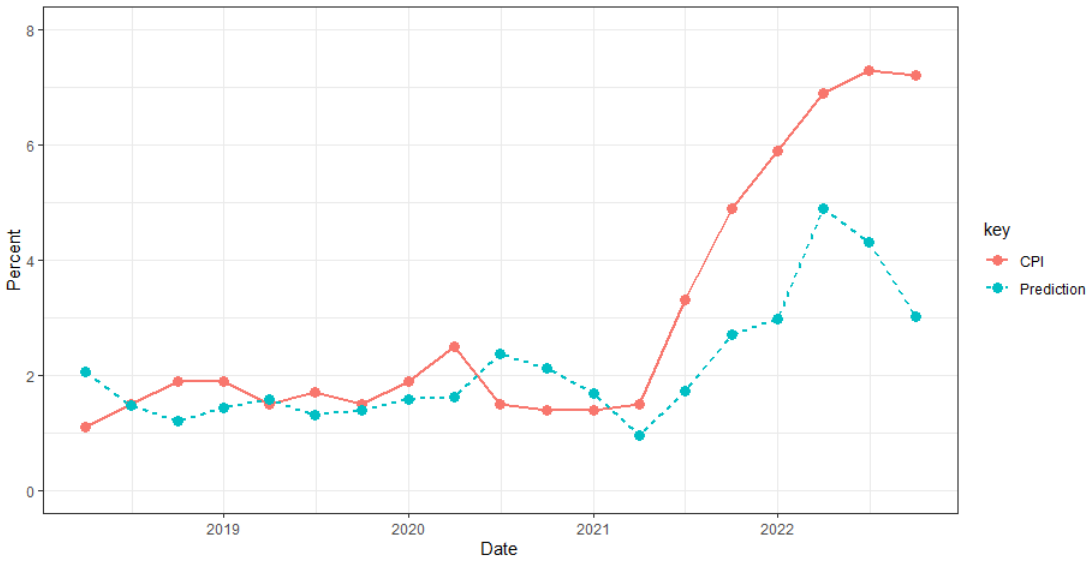

Figure 2: Predicted values of year-over-year CPI inflation versus observed values.

Model	Mean Abs. Error		
	2018Q1-2021Q1	2021Q2-2022Q3	Full Sample
XGBoost	0.49	2.65	1.17
XGBoost + CPI	0.39	1.62	0.78
AR(1)	0.35	1.31	0.67

Table 4: XGBoost nowcast errors and forecasts errors from a simple AR(1) estimated from 1990Q1 to 2018Q1, as well as one estimated from 2014Q1 to 2018Q1.

These early results suggest the potential for additional work. It remains to be seen whether the granular data can give a better read of core CPI inflation, or components of the index. Another possibility is to add additional features to the data, such as an indicator of petrol prices, to give it better coverage of the CPI basket. While the exercise presented here was a naive initial use of the machine learning algorithm, there are a number of techniques for making the algorithm's job easier which should be explored, as well asother ways of aggregating or selecting which features to use. Finally, to the extent that the model can be improved, there is the potential to create a high frequency indicator for the CPI. This could be of use to the Reserve Bank of New Zealand, both given the low frequency of New Zealand's CPI data, and also for use in crises like the initial days of the COVID-19 pandemic and resulting lockdowns.

References

Almosova, A., and N. Andresen (2023): "Nonlinear inflation forecasting with recurrent neural networks," *Journal of Forecasting*, 42, 240–259.

Aparicio, D., and M. Bertolotto (2020): "Forecasting inflation with online prices," *International Journal of Forecasting*, 36, 232–247.

Bentley, A., and F. Krsinich (2017): "Towards a big data CPI for New Zealand," *Paper presented at the Ottawa Group*.

Breton, R., G. Clews, L. Metcalfe, N. Milliken, C. Payne, J. Winton, and A. Woods (2015): "Research indices using web scraped data," *Office for National Statistics*.

Brynjolfsson, E., and M. Smith (2000): "Frictionless commerce? A comparison of internet and conventional retailers," *Management Science*, 46(4), 563–585.

Cavallo, A. (2017): "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers," *American Economic Review*, 107(1), 283–303.

Cavallo, A., B. Neiman, and R. Rigobon (2014): "Currency unions, product introductions, and the real exchange rate," *Quarterly Journal of Economics*, 129(2), 529–595.

——— (2015): "The price impact of joining a currency union," *IMF Economic Review*, 63(2), 281–297.

Cavallo, A., and R. Rigobon (2016): "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, 30(2), 151–178.

Clay, K., K. Ramayya, E. Wolff, and D. Fernandes (2002): "Retail strategies on the web: price and non-price competition in the online book industry," *Journal of Industrial Economics*, 50(3), 351–367.

Harchaoui, T., and R. Janssen (2018): "How can big data enhance the timeliness of official statistics? The case of the U.S. consumer price index," *International Journal of Forecasting*, 34, 225–234.

Horrigan, M. (2013): "Big data: a perspective from the BLS," *Amstat News*.

Jaworski, K. (2021): "Measuring food inflation during the COVID-19 pandemic in real time using online data: a case study of Poland," *British Food Journal*, 123(13), 260–280.

Krsinich, F. (2015): "Price indexes from online data using the fixed-effects window-splics (FEWS) index," *Paper presented at the Ottawa Group*.

Lynch, D., S. Olivecrona, and M. Stansfield (2019): "Creating a digital food price index from web scraped data," *Paper presented at the Ottawa Group*.

Macias, P., D. Stelmasiak, and K. Szafranek (2023): "Nowcasting food inflation with a massive amount of online prices," *Nowcasting food inflation with a massive amount of online prices*, 39, 809–826.

Nygard, R. (2015): "The use of online prices in the Norwegian consumer price index," *Paper presented at the Ottawa Group*.

Osbat, C. (2022): "What micro price data teach us about the inflation process: web-scraping in PRISMA," *SUERF Policy Brief*, 470.

Qureshi, S., B. Chu, and F. Demers (2020): "Forecasting Canadian GDP growth using XGBoost," *Carleton Economics Working Papers*, 20(14).

Stansfield, M., and F. Krsinich (2021): "Bigger, better, faster: further progress in using non-traditional data to measure price inflation," *Paper presented at the 61st Annual Conference of the New Zealand Association of Economists*.