ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting

Ben Powell and Guy Nason,

University of Bristol, UK

Duncan Elliott, Matthew Mayhew and Jennifer Davies

Office for National Statistics, Newport, UK

and Joe Winton

Statistics New Zealand, Auckland, New Zealand

[Received August 2016. Final revision July 2017]

Summary. With the increasing relevance and availability of on-line prices that we see today, it is natural to ask whether the prediction of the consumer price index (CPI), or related statistics, may usefully be computed more frequently than existing monthly schedules allow for. The simple answer is 'yes', but there are challenges to be overcome first. A key challenge, addressed by our work, is that web-scraped price data are extremely messy and it is not obvious, *a priori*, how to reconcile them with standard CPI statistics. Our research focuses on average prices and disaggregated CPI at the level of product categories (lager, potatoes, etc.) and develops a new model that describes the joint time evolution of latent daily log-inflation rates driving prices seen on the Internet and prices recorded in official surveys, with the model adapting to various product categories. Our model reveals the differing levels of dynamic behaviour across product category and, correspondingly, differing levels of predictability. Our methodology enables good prediction of product-category-specific CPI immediately before their release. In due course, with increasingly complete web-scraped data, combined with the best survey data, the prospect of more frequent intermonth aggregated CPI prediction is an achievable goal.

Keywords: Dynamic inflation model; High frequency inflation prediction; Inflation estimation; State space model

1. Background to index modelling

The consumer price index (CPI) measures the rate at which the prices of goods and services bought by a typical household rise and fall. The index has numerous important political and financial ramifications. On a political or social level, a population's perceived wealth and prosperity are significantly influenced by the amount that they pay for goods every day, changes in these prices often being more visible than the sums that enter and leave their bank accounts in an automated fashion. Thus the mood of an electorate, for example, may be highly sensitive to the CPI. The index certainly has more tangible macroeconomic consequences also. Since pensions obligations and other large-scale spending plans are often pegged to the index, even very minor

Address for correspondence: Ben Powell, Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK.
E-mail: ben.powell@bristol.ac.uk

changes in the CPI can alter the prices of financial portfolios by millions of pounds. For a more comprehensive introduction to index numbers we defer to Ralph *et al.* (2015).

The strategy of the UK's Office for National Statistics (ONS) for computing the CPI has traditionally involved the careful design of price surveys, implicitly guided by the knowledge of the agents that they employ to collect the price data. A thorough description of their methodology can be found in Office for National Statistics (2014) (where, for reasons that will not affect us here, product categories are referred to as items and category members as products). Recently, however, as the resources for accumulating and processing large amounts of data have increased, statistical institutes such as the ONS are beginning to experiment with the idea that small, well-designed, but costly sampling exercises might be complemented by inferences drawn from the analysis of many more prices that are published on the Internet. Breton *et al.* (2016) and Bunn and Ellis (2009) discuss such experiments.

Web-scraped data sets of prices tend to be much larger and much messier than their hand-picked equivalents, to the extent that human processing and interpretation of the data are impractical. The automation of these tasks is thus a major component of the current project. We shall demonstrate that estimates for disaggregated inflation statistics based on web-scraped data are strongly related to conventional estimates. We do this to justify the relevance of the web-scraped data to inflation estimation in general. Once that relevance has been established, we shall be able to produce and defend CPI statistics at a much finer level of granularity than seen previously. Indeed, we envisage this work potentially leading to the production of inflation measures pertaining to small demographic or geographic populations, on timescales of days rather than months. These CPI figures would provide policy makers with a more detailed view of who inflation is affecting and when than has previously been available. Such information may enable them to take actions that are correspondingly well targeted and timely. A secondary, more immediate, aim of this project is to produce forecasts for the published CPI so that interested parties may compute estimates of official inflation statistics before they are released. Such an early warning system may help in the preparation of actions responding to the official figures. Lastly, we intend the project to demonstrate both our capacity for handling large, incomplete data sets in the production of official statistics, and the appetite of users of official statistics for the information that the data sets contain. Establishing these facts will, in the longer term, inform arguments for the acquisition of data from unconventional sources other than the Internet.

The review of index forecasting methods in Faust and Wright (2013) compares a large number of inflation models and is accompanied by insightful comments regarding their significance in macroeconomic terms. Their set of models all treat inflation as a monthly time series and include members that regress on explanatory variables. In general, however, they are not equipped to assimilate large quantities of high frequency price data. Interestingly, one of their main conclusions is that forecasts that are based on the output of the very simplest mathematical models, and those based on subjective expert advice, tend to outperform those arising from more technically complex models. More specifically, they found that a model that describes an index as the sum of random-walk and a stationary auto-regressive process of order 1 outperforms more complex dynamic stochastic general equilibrium models, such as that described in Smets and Wouters (2007), as do forecasts based on expert judgements, such as those published in the US Federal Reserve's 'Greenbook'. Further evidence for the value of elicited expert opinion can be found in Ang *et al.* (2007). Even models that are informed by, nominally exogenous, explanatory variables, such as the vector auto-regressive and factor-augmented vector auto-regressive models of Ang and Piazzesi (2003) and Boivin *et al.* (2005) respectively fail to beat the simpler models consistently. We hypothesize that the relationships between the macroeconomic variables that are involved are too complex or transitory for the models to describe adequately.

The conclusions of Faust and Wright (2013) reinforce our commitment to simple, transparent models that are tightly constrained by expert beliefs.

The literature that looks at inflation from a faster, disaggregated, microeconomic point of view is less well established but is growing quickly, largely thanks to the work of researchers at the Billion Prices Project and its commercial off-shoot, PriceStats. The Billion Prices Project's academic output, which is nicely summarized in Cavallo and Rigobon (2016), informs many of our modelling decisions and inferences. Aparicio and Bertolotto (2016) have presented CPI forecasts informed by web-scraped prices, as computed with a low order auto-regressive moving average model with exogenous inputs. The exogenous variable here is an aggregated CPI computed from web-scraped prices. Reassuringly, the paper's results show that their model outperforms a range of simple models which do not take this extra information into account.

Building on the work of the Billion Prices Project, this paper describes a more sophisticated model for representing the relationships between inflation, indices, on-line prices, off-line prices and official governmental statistics. Most significantly, we show how important issues, such as the discrepancies between surveyed and web-scraped prices, can be accounted for in a parsimonious way. Additionally, we demonstrate how the statistical time series analysis literature can help us to produce prediction intervals for inflation, and to identify appropriate model hyperparameters.

Our paper proceeds as follows: Section 2 discusses our web-scraped data, the analysis of which motivates our methodology; Section 3 describes our preferred model for prices and their evolution over time; Section 3.2 introduces the sequential learning algorithm for tracking log-inflation and log-index quantities; Section 4 discusses the inferences that can be drawn from the model when fitted to prices of goods at the category level of disaggregation; finally, Section 5 highlights some observations and findings that seem especially relevant to future work in this area. A more thorough description of the algorithms for inference and hyperparameter adjustment can be found in Appendix A.

2. Introducing the data

Our analysis uses two data sets. The first consists of daily web-scraped prices covering 33 product categories harvested from the Web sites of three large supermarkets operating in the UK over a period of approximately 14 months. The supermarkets have a stated policy of charging the same on-line prices throughout the country and, from limited experimentation with running the web scrapers from different geographic locations, this appears to be so. According to the 2014 ONS weighting scheme, the product categories that we have web-scraped data for contribute approximately 13% of the total aggregated CPI. The average number of distinct items encountered by the scraping algorithms within each category is approximately 340, and the average length of time that they follow any particular item is approximately 146 days.

The second data set contains disaggregated CPI values for the same product categories, as well as many more that also contribute to the aggregated CPI, recorded and published by the ONS monthly. Below we focus only on the 33 product categories that are common to both data sets.

Plots of the web-scraped data for two product categories are presented in Fig. 1 and Fig. 2. In Fig. 1 the extent of the data missingness can be seen from the white gaps in the arrays. The varying degree of missingness over time also impresses on us the need for a model to quantify the changing precision of index estimates based on the data. Unfortunately, we currently lack the ability to distinguish missing prices reliably because of stock issues and failures in the web scraping procedure. The web scraping algorithms that are used to collect the data did also attempt to identify promotions and to flag otherwise exceptional prices on the basis of contextual information on the supermarket Web sites. Their success at doing so, however, was

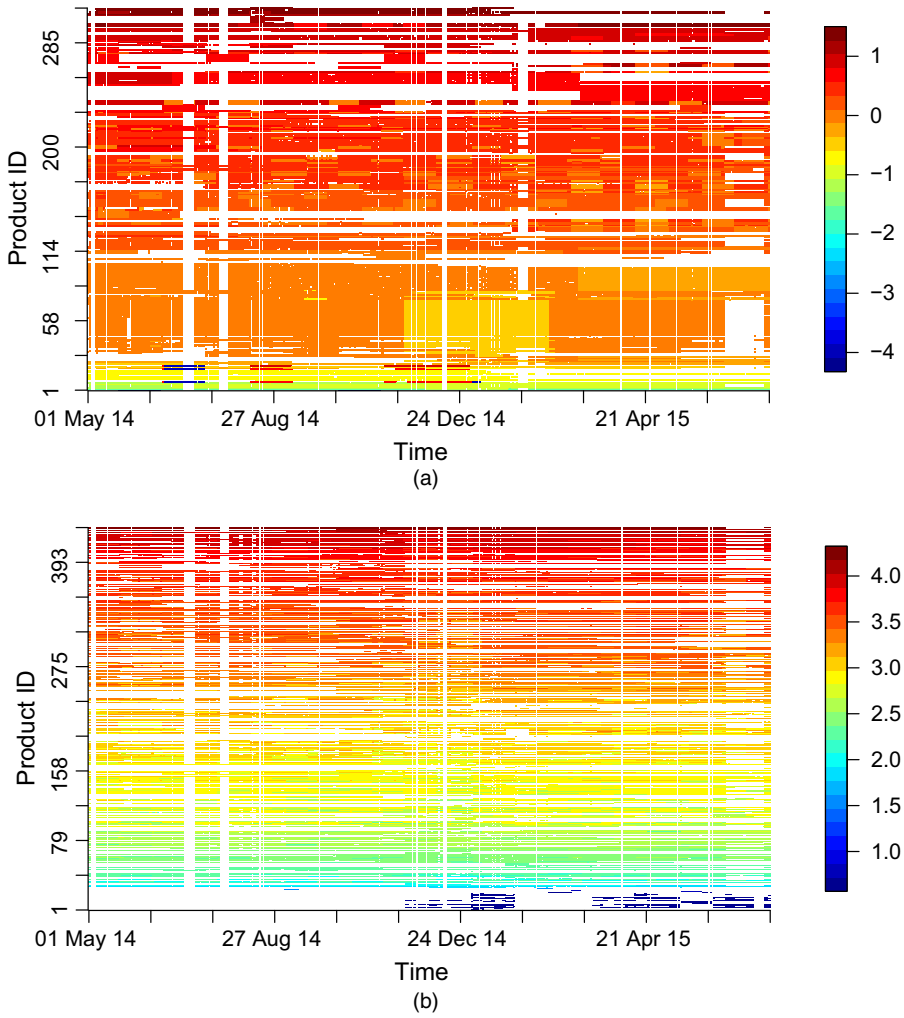

Fig. 1. Web-scraped daily log-price data for products over time (the individual products within a category, which correspond to rows of the arrays, have been sorted according to their mean value; □, missing values): (a) pasta category; (b) whisky category

poor and this aspect of the data is not used in the analysis below. Instead, we have chosen to downweight large day-by-day log-prices changes by passing them through a soft thresholding function as a preprocessing step. This thresholding strategy represents an initial attempt to mitigate the destabilizing effect of large price jumps that are commonly attributable to product discounting, which may be seen as a specific form of price friction. Discussion of observational evidence for different types of on-line price friction (or jumpiness) can be found in Lünnemann and Wintr (2011). We expect that more sophisticated treatments of price friction or jumpiness will eventually lead to improved inflation estimates. For now, however, in an effort to minimize model complexity, we avoid making further modifications to the pricing model.

The price variation between cheap and expensive products within a product category tends to dominate price variation over time. This is manifested in Fig. 1 in the vertical colour gradient being more noticeable than the horizontal colour gradient. To obtain a better view of

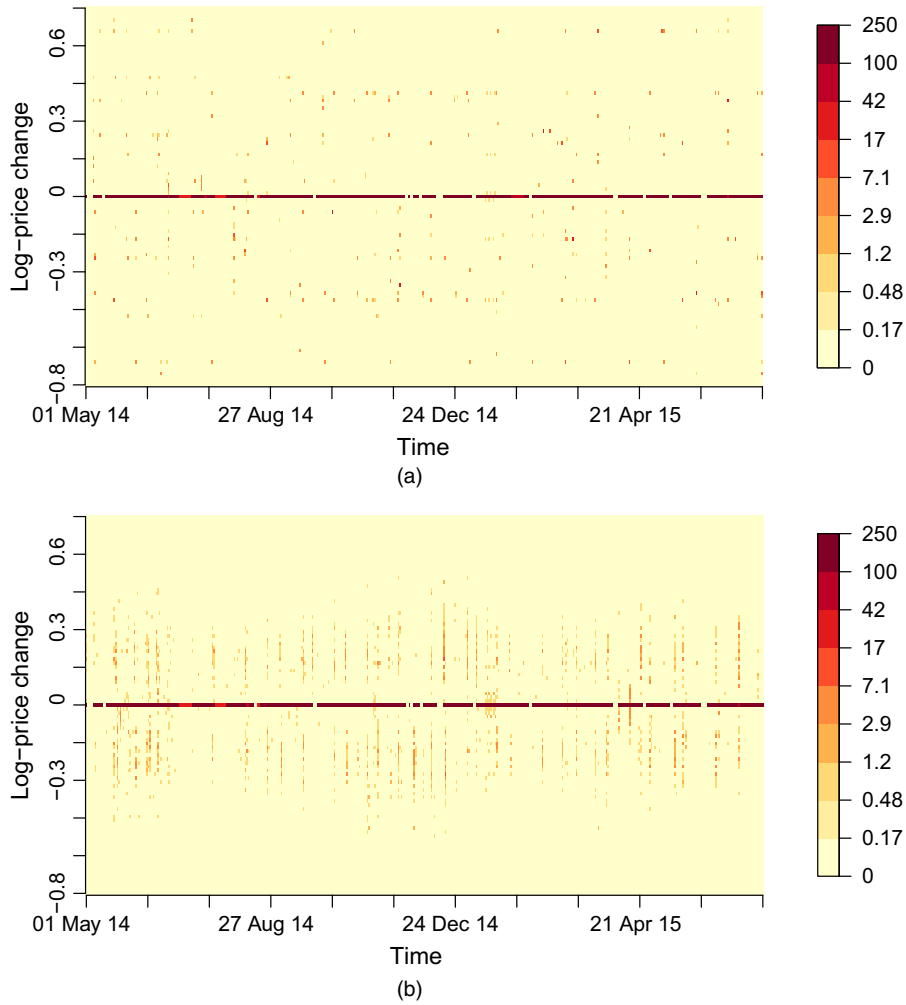

Fig. 2. Each column of these arrays encodes a histogram of observed changes in log-price for products in a category (the colours relate to counts of log-price changes on a particular day; note that the colour axis is on an inverse hyperbolic sine scale to provide greater distinction between low counts): (a) pasta category; (b) whisky category

the changes in log-price over time we compute the 1-day log-price changes for each available item. We then produce histograms of these changes for each day. Fig. 2 shows the histograms stacked side by side and viewed from above. It is obvious from these plots that the majority of log-price changes (approximately 98%) are exactly 0. It is also evident that the two products that are considered here are behaving differently: non-zero changes in the log-price for whisky appear to be more common than for pasta and are more concentrated in terms of their absolute value.

3. Modelling

This section describes a model for describing the relationship between inflation, price data and the published CPI for individual product categories. For example, we might consider the product

category 'lager', of which products 'Brand X's Pilsner' and 'Brand Y's Kölsch' are members. For the time being we shall analyse the product categories independently. For forecasting the aggregated CPI, this is obviously not an optimal strategy because of the dependences between categories. Our current aim, however, is to gain insight and predictive power for the disaggregated indices for which we have data. By doing so we may investigate a range of index behaviours. These are interesting in their own right and stretch our inflation model in different directions, but they will also contribute to our understanding of the aggregated CPI, whose analysis is reserved for future work. We also note that the model that is described below is one of many that may be used to structure inferences from the price data. Although the model may not be demonstrably optimal, it has proved robustly useful in our work here. For brevity, the nature of our model is only sketched out in the following section. A comprehensive description is provided in Appendix A.

3.1. Price setting

Considering the task at hand, the review by Faust and Wright (2013) is especially significant for its identification of simple models for inflation forecasting outperforming complex models in terms of predictive skill. Specifically, they draw attention to what they called the 'fixed ρ inflation gap' model, which describes log-inflation, denoted π_t, as the sum of a random walk τ_t and an auto-regressive AR(1) process g_t. The random-walk term provides a slowly varying local mean, which Faust and Wright (2013) suggested might be attributed to long-term monetary policy. The AR(1) process then allows for medium-term deviations around the local mean. Our work below adopts a variant on this model, electing to replace the random-walk component with another AR(1) component whose auto-correlation length is forced to be longer than that of g_t. The resulting model is written as

$$\pi_t = \tau_t + g_t, \tag{1}$$

$$\tau_t = \phi_\tau \tau_{t-1} + \epsilon_t, \tag{2}$$

$$g_t = \phi_g g_{t-1} + \varepsilon_t, \qquad 0 < \phi_g < \phi_\tau < 1, \tag{3}$$

where the innovation terms ϵ_t and ε_t have mean 0 and are mutually uncorrelated. The replacement of the random walk with a slowly varying AR(1) process is made primarily for computational convenience since it renders the long-term variance of π_t finite.

The cumulative sum of the log-inflation terms, starting from a specified base period t_0, is then identified with a daily log-index quantity

$$d_t = \sum_{s=t_0}^{t} \pi_s. \tag{4}$$

We assume that, from one day to the next, a retailer will adjust the log-price of product i, denoted $y_{i,t}$, approximately in line with log-inflation. Explicitly, we write

$$y_{i,t+1} - y_{i,t} = \pi_t + e_{i,t}, \tag{5}$$

where the $e_{i,t}$-terms encode mutually uncorrelated product-specific price innovations with zero means and common variances. This lack of correlation between products is consistent with the assumption that, having accounted for inflation, all products have their prices changed independently from each other and independently from previous price changes. Deviating from these assumptions, by explicitly modelling the effects of cross-price elasticities for example, is not considered here.

Our strategy for relating the ONS's published log-inflation figures and those estimated from the web-scraped data begins by imagining two identically distributed log-inflation processes, π_t^{surv} (for *survey*) and π_t^{web} (for *web scraped*), of the type described in equation (1). We induce correlation between them by supposing that the innovation terms (ϵ_t and ε_t) that drive them are correlated. As described more precisely in Appendix A.1.2, we allow for differing degrees of correlation in the long-term and short-term components of log-inflation.

Hereafter, we shall refer to the quantities

$$\{\tau_t^{\text{surv}}, \tau_t^{\text{web}}, g_t^{\text{surv}}, g_t^{\text{web}}\}, \tag{6}$$

as well as the log-inflation and log-index quantities that are derived from their sums, as latent parameters. In Sections 3.2 and Appendix A, we describe how their values can be encoded as elements of a state vector that we adjust sequentially by using linear combinations of log-price changes and published log-CPI values. The log-index d_t^{surv} is of particular importance here since it is what we currently want to predict. This is so because we need to demonstrate how our model relates to conventional, published index estimators. We note that in the future, particularly when our web scraping capacity increases, the implied superiority of the survey-based estimates may not be appropriate.

The quantities that interact with the state vector in more than a simple additive way,

$$\Omega = \{\sigma_\tau^2 = \text{var}(\epsilon_t), \sigma_g^2 = \text{var}(\varepsilon_t), \phi_\tau, \phi_g, \rho_\tau = \text{corr}(\epsilon_t^{\text{web}}, \epsilon_t^{\text{surv}}), \rho_g = \text{corr}(\varepsilon_t^{\text{web}}, \varepsilon_t^{\text{surv}})\}, \tag{7}$$

will be treated as hyperparameters. Our approach to specifying these, which is described fully in Appendix A.3, is based on a combination of *a priori* specification and empirical Bayes methodology.

3.2. Inference

We now briefly discuss the inferential machinery for deriving estimates for the model parameters and hyperparameters. Further mathematical details appear in Appendix A. An expression describing the joint evolution of the log-inflation and log-index quantities can be conveniently written using vector notation as

$$\mathbf{u}_{t+1} = \mathbf{G}\mathbf{u}_t + \mathbf{e}_t. \tag{8}$$

The vector $\mathbf{u}_t$ in equation (8) is referred to as the model's state vector and contains the latent process values, whereas the innovation vector $\mathbf{e}_t$ contains mean zero noise terms. The evolution matrix $\mathbf{G}$ contains hyperparameters to be specified *a priori* or inferred by using the approximate likelihood methods described in Appendix A.3. Observed log-price changes and published log-CPI figures can both be equated with noisy linear combinations of elements of the state vector. Hence, we can proceed by employing standard Kalman filter (Chui and Chen, 2009) or dynamic linear model methodology (West and Harrison, 1997) to adjust our estimates for $\mathbf{u}_t$ as time progresses and more data are received.

4. Fitting the model to real price data

4.1. Remarks on the fitting procedure

Numerical optimization of the log-posterior density to compute empirical Bayes estimates for the hyperparameters was carried out using the Nelder–Mead simplex method (Nelder and Mead, 1965) implemented in R (R Development Core Team, 2009). Analysis of the Hessian of the log-posterior for the hyperparameters at convergence points confirmed that the optimizer was

indeed locating local maxima. Reinitializations of the optimizer, however, occasionally resulted in different points of convergence, indicating that the log-likelihood may be multimodal. As such, we must entertain the possibility that the optimizer is locating suboptimal hyperparameter point estimates. Optimizer convergence typically required around 1000 passes of the filter over the data for each product category. The need to perform so many iterations underlines the importance of our inflation model's simplicity and the small size of the state vector, which determines the computational demand of each pass.

4.2. Remarks on inferred consumer price index values

We proceed by presenting the inferences produced by our model for each product category, conditionally on the optimized hyperparameters for them. Fig. 3 plots daily expectations and variances for the log-CPI, d_t^{surv}, for a subset of product categories, given all the data preceding that day. The variances are captured by the shaded blue regions around the expectations, whose width is four times the standard deviation of our implied distribution for d_t^{surv}. By appealing to Gauss's inequality (Pukelsheim, 1994), we identify these regions with approximate conservative 95% credible regions. We can see immediately that the forecasts for this subset of product categories tend to be highly accurate, and the magnitudes of forecast errors are well quantified by the credible regions. Equivalent plots for all 33 product categories are presented in Fig. 4.

Table 1 contains empirical scores for the published log-CPI values in terms of their differences from $E(d_t^{\text{surv}})$ at the CPI publication dates. We reserve particular attention for the root-mean-squared error RMSE and median absolute deviation MAD of the log-CPI from its forecasted values. Presented are the error statistics computed by using versions of our model both with and without the web-scraped data. To provide reference figures we also present statistics computed by using a persistence forecast strategy whereby the published log-CPI is forecasted to be the same as in the previous month. We shall refer to these as the *informed*, *uninformed* and *persistence* models respectively. Informally, we observe that the RMSE-statistics across product categories are highly variable and, given the small number of official CPI releases in the time range, sensitive to a small number of large forecast errors. On average, however, the RMSE-statistics can be seen to fall from 0.029 to 0.028 to 0.025, as we introduce the model structure and then the web-scraped data.

These figures ought to be compared with the findings of Aparicio and Bertolotto (2016) who considered forecasts of the aggregated UK log-CPI covering all consumer products monthly. Specifically, we concentrate on Table 1 of Aparicio and Bertolotto (2016). Their calculations suggest that regressing on aggregates of web-scraped log-prices can lead to reductions in RMSE from 0.00191 (for their basic uninformed AR(1) model) to 0.00118 (for their best performing model informed by on-line prices including fuel prices). There are several reasons why these figures are smaller than those computed for the disaggregated log-CPI. Firstly, on the assumption that the inflation rates for different product categories are correlated, forecasting the aggregated log-CPI is inherently at least as easy as forecasting the disaggregated log-CPI. This is simply a consequence of the variance of the mean of a set of independent random variables decreasing as the set becomes larger. Secondly, we hypothesize that the reduction in RMSE attributable to the web-scraped data is smaller for the disaggregated log-CPI (approximately $15\% \approx 0.025/0.029 \times 100\%$ rather than $40\% \approx 0.0018/0.00191 \times 100\%$) may be due to our web scrapers' misclassification of items into product categories. In practice misclassification, or failure to replicate the classifications made by ONS price surveyors, means that we are often feeding our model inappropriate information. Obviously, for more finely disaggregated product categories our web scrapers need to be increasingly more discerning. Conversely, as we consider

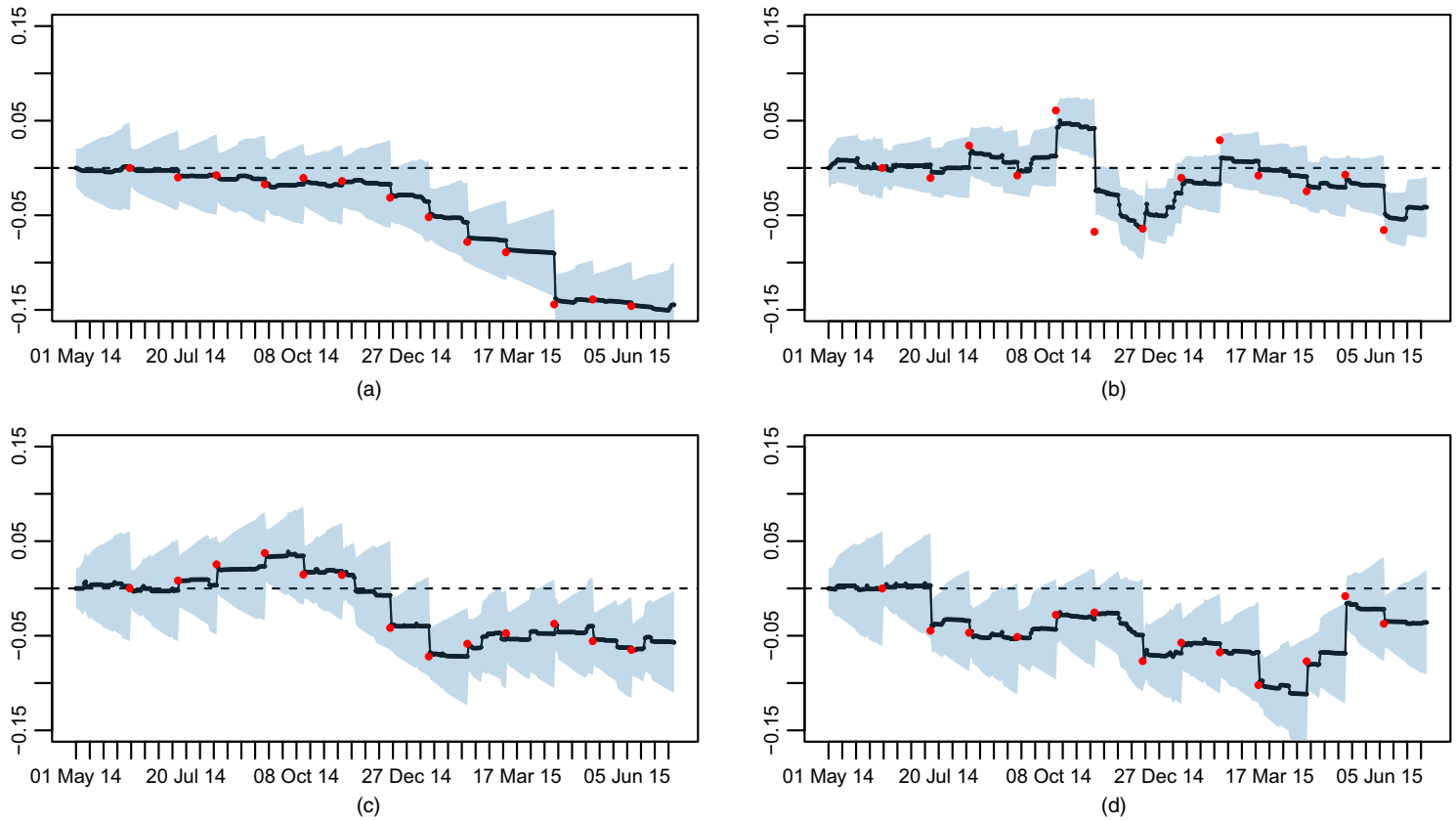

Fig. 3. Daily expectations and variances for product categories (a) white sliced loaf, (b) whisky, (c) pasta and (d) cider: ———, interpolated forecasted values of the published log-CPI informed by preceding observations of web-scraped data and monthly published CPI; , pointwise approximate 95% credible regions; ●, published log-CPI values

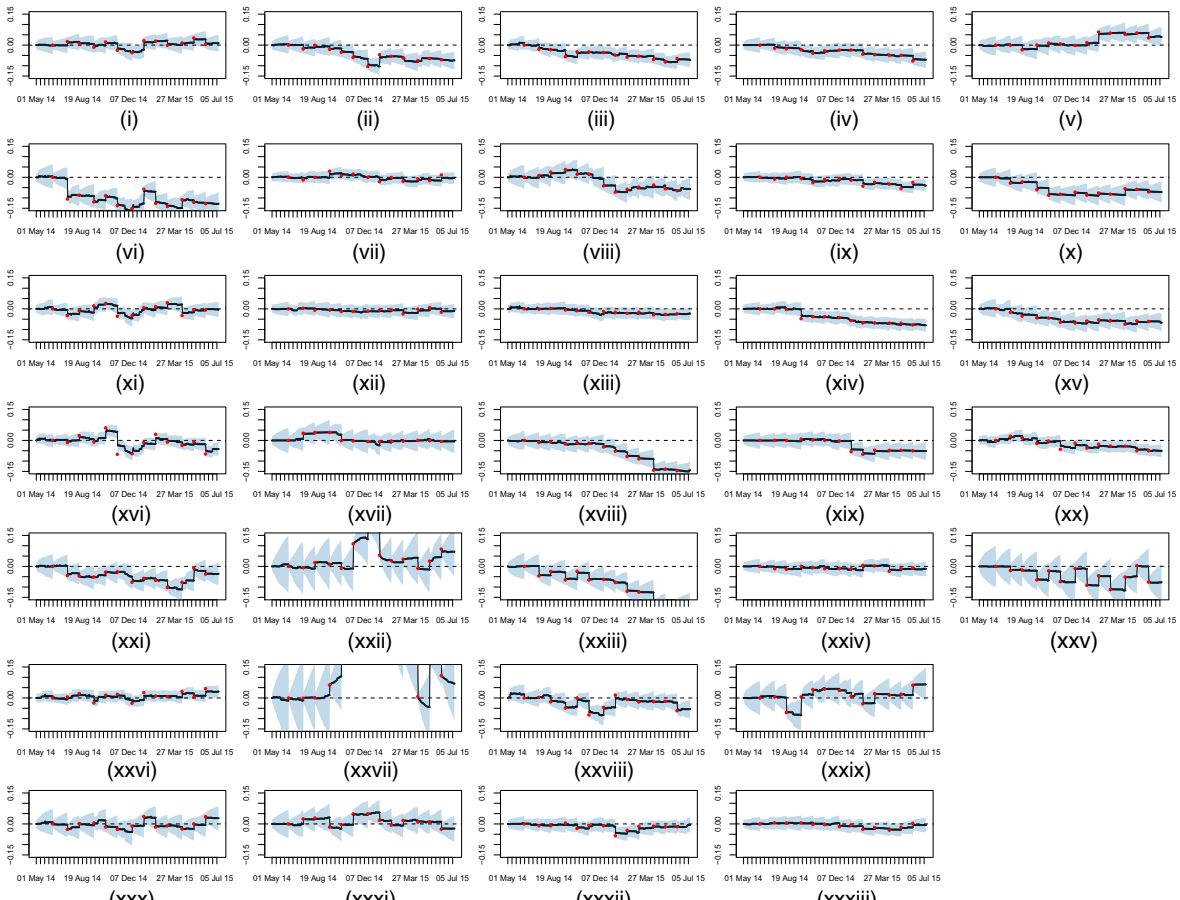

Fig. 4. Plots of estimated CPI for all product categories (———, interpolated forecasted values of log-CPI (d_t^{surv}) informed by preceding observations of web-scraped data and monthly published CPI; ■, pointwise approximate 95% credible regions; ●, published log-CPI values): (i) brandy; (ii) bitter; (iii) plain cereal; (iv) other fruit juice; (v) tomatoes; (vi) rum; (vii) lager; (viii) pasta; (ix) cheddar; (x) potatoes (new); (xi) vodka; (xii) apples; (xiii) biscuits; (xiv) whole milk; (xv) potatoes (old); (xvi) whisky; (xvii) bananas; (xviii) white sliced loaf; (xix) semi-skimmed milk; (xx) tea bags; (xxi) cider; (xxii) grapes; (xxiii) brown sliced loaf; (xxiv) yoghurt; (xxv) butter; (xxvi) red wine; (xxvii) strawberries; (xxviii) cola; (xxix) *fromage frais*; (xxx) white wine; (xxxi) sugary cereal; (xxxii) orange juice; (xxxiii) onions

Table 1. Fitted category-specific hyperparameters and empirical forecast metrics for the web-scraped prices†

Product	*Hyperparameters*			*Forecast metrics*					
	$\sqrt{var(g_t)}$ $(\times 10^{-2})$	ϕ_g	ρ_g	*Informed RMSE* $(\times 10^{-2})$	*Uniformed RMSE* $(\times 10^{-2})$	*Persistence RMSE* $(\times 10^{-2})$	*Informed MAD* $(\times 10^{-2})$	*Uninformed MAD* $(\times 10^{-2})$	*Persistence MAD* $(\times 10^{-2})$
Butter	0.20	0.33	0.55	4.26	4.30	5.61	3.86	3.73	5.37
Lager	0.39	0.29	0.39	1.42	1.72	1.87	0.95	1.02	1.71
Tea bags	0.45	0.32	0.50	1.72	1.76	2.12	1.04	1.08	1.92
White wine	0.61	0.23	0.18	2.61	2.47	2.76	1.90	1.63	2.22
Whisky	0.64	0.55	0.91	3.54	4.66	5.27	1.08	3.31	3.58
Semi-skimmed milk	0.65	0.44	0.42	1.91	1.96	1.68	0.43	0.46	0.28
Apples	0.76	0.50	0.57	0.99	1.03	1.19	0.76	0.75	0.63
Cheddar	0.84	0.44	0.77	1.29	1.50	1.73	0.81	0.85	1.28
Vodka	1.09	0.47	0.79	2.43	3.01	3.24	2.05	2.28	2.35
Yoghurt	1.15	0.50	0.11	1.14	1.09	1.15	0.71	0.64	0.62
Pasta	1.18	0.70	0.88	1.67	2.40	2.22	0.58	1.47	1.27
Brown sliced loaf	1.18	0.63	0.72	3.32	3.38	3.56	2.67	3.07	3.41
Onions	1.19	0.46	0.57	1.03	1.10	1.13	0.76	0.64	0.87
Biscuits	1.20	0.32	0.14	0.60	0.55	0.59	0.49	0.45	0.42
Brandy	1.27	0.51	0.89	1.88	2.49	2.69	1.61	2.02	1.95
Orange juice	1.37	0.53	0.49	1.79	1.85	1.97	0.70	0.94	1.15
Plain cereal	1.37	0.43	0.34	1.60	1.57	1.64	1.20	1.24	1.26
Bananas	1.44	0.86	0.13	1.42	1.55	1.54	0.39	0.27	0.27
Whole milk	1.51	0.48	0.54	1.41	1.43	1.44	0.34	0.26	0.43
Red wine	1.57	0.49	0.72	2.09	2.76	3.15	1.43	2.42	2.73
Potatoes (new)	1.58	0.54	0.37	1.98	1.96	1.92	1.03	1.03	1.12
White sliced loaf	1.67	0.90	0.87	1.56	1.96	2.02	0.69	1.00	0.99
Bitter	1.77	0.51	0.27	2.30	2.43	2.51	1.33	1.55	1.60
Tomatoes	1.78	0.66	0.20	2.00	1.98	1.97	1.01	1.01	0.87
Strawberries	2.01	0.97	0.77	11.72	13.98	13.10	3.54	7.88	6.46
Other fruit juice	2.13	0.47	0.39	1.22	1.26	1.28	0.60	0.53	0.39
Sugary cereal	2.26	0.74	0.11	2.77	2.74	2.73	2.33	2.32	2.30
Rum	2.83	0.48	0.32	4.37	4.90	5.07	2.60	2.94	3.12
Cider	2.85	0.69	0.77	2.73	3.30	3.31	1.01	2.43	2.43
Potatoes (old)	3.34	0.43	0.49	1.18	1.25	1.24	1.14	1.01	0.93
Grapes	3.84	0.88	0.40	5.98	6.56	6.59	3.24	3.28	3.08
Fromage frais	3.93	0.57	0.29	4.16	4.00	4.02	2.18	3.28	3.19
Cola	3.95	0.44	0.64	3.40	3.82	3.90	2.60	2.64	2.60

†Omitted hyperparameters were considered fixed and common to all product categories. See Appendix A.3 for further details.

more highly aggregated categories the failings of the web scrapers to discriminate between products become less significant.

Further quantitative descriptions for the reductions in forecast error can be produced by looking more closely at the squared error statistics for predictions made by the three models:

$$\delta^2_{\mathrm{a},t} := \{d_t^{\mathrm{surv}} - E(d_t^{\mathrm{surv}} \mid \mathrm{Model_a})\}^2, \qquad t \in \Omega_{\mathrm{pub}}, \tag{9}$$

where Ω_{pub} is the set of official CPI publication dates and a is an index for identifying either the persistence, uninformed or informed model. Initial work in which we computed ratios of RMSE-statistics

$$R_{\mathrm{a,b}} := \frac{\sqrt{\left(n^{-1} \sum\limits_{t \in \Omega_{\mathrm{pub}}} \delta^2_{\mathrm{a},t}\right)}}{\sqrt{\left(n^{-1} \sum\limits_{t \in \Omega_{\mathrm{pub}}} \delta^2_{\mathrm{b},t}\right)}}, \tag{10}$$

for a pair of models labelled a and b, proved to result in figures that were highly sensitive to unusually large errors. For product categories exhibiting one particularly large jump in log-CPI at a particular point during the observation period, for example, expression (10) essentially just measures the models' performance at this time. Product categories that normally exhibit only very low inflation rates are especially vulnerable to this. On the understanding that sudden large price jumps are commonly attributable to extraneous economic factors that none of the models can anticipate, the RMSE ratio is a particularly inappropriate metric for assessing them.

A more illuminating analysis derives from consideration of the transformed errors

$$\zeta^2_{\mathrm{a},t} := \zeta^2(\delta^2_{\mathrm{a},t}) = \delta_0^2 + (\delta_0^2 + \nu) \log\left(\frac{\delta^2_{\mathrm{a},t} + \nu}{\delta_0^2 + \nu}\right), \qquad \nu = \delta_0^2 = 0.01^2. \tag{11}$$

These new error quantities are approximately equal to $\delta^2_{\mathrm{a},t}$ when $|\delta^2_{\mathrm{a},t} - \delta_0^2|$ is small, with the parameter ν governing exactly how local the approximation is. The transformation serves to reduce the leverage of unusually large forecast errors, providing our analysis with robustness against extreme events. We note that a simpler log-transform, reached in the limit that takes ν to 0, leaves us dangerously sensitive to very small forecast errors, and that the $\zeta^2_{\mathrm{a},t}$ can be identified with negative log-likelihoods as described by t-distributions. We can now look at scaled differences between the error statistics at each publication date

$$Z_{\mathrm{a,b},t} := (\delta_0^2 + \nu)^{-1}(\zeta^2_{\mathrm{a},t} - \zeta^2_{\mathrm{b},t}) = \log\left(\frac{\delta^2_{\mathrm{a},t} + \nu}{\delta^2_{\mathrm{b},t} + \nu}\right) \tag{12}$$

and consider how different they are from 0. In accordance with methodology that was established by Diebold and Mariano (1995) for comparing the predictive accuracy of competing forecasts, we appeal to the approximate normality of the long-term means of the $Z_{\mathrm{a,b},t}$-statistics to compute approximate confidence intervals for them. In doing so we may quantify the significance of the error reductions that are attributable to the web-scraped data in a way that discounts anomalously large shocks. Fig. 5 plots the approximate confidence intervals for the means of the $Z_{\mathrm{a,b},t}$ measuring the error differences for various models. These intervals are computed as

$$\bar{Z}_{\mathrm{a,b}} \pm z_{0.975} n^{-1/2} \hat{s}_{\mathrm{a,b}}, \tag{13}$$

where $\bar{Z}_{\mathrm{a,b}}$ is the sample mean of the $Z_{\mathrm{a,b},t}$ over time, $\hat{s}^2_{\mathrm{a,b}}$ is an estimate of the variance of the $Z_{\mathrm{a,b},t}$ that approximately accounts for auto-correlation, computed as a windowed sum of the

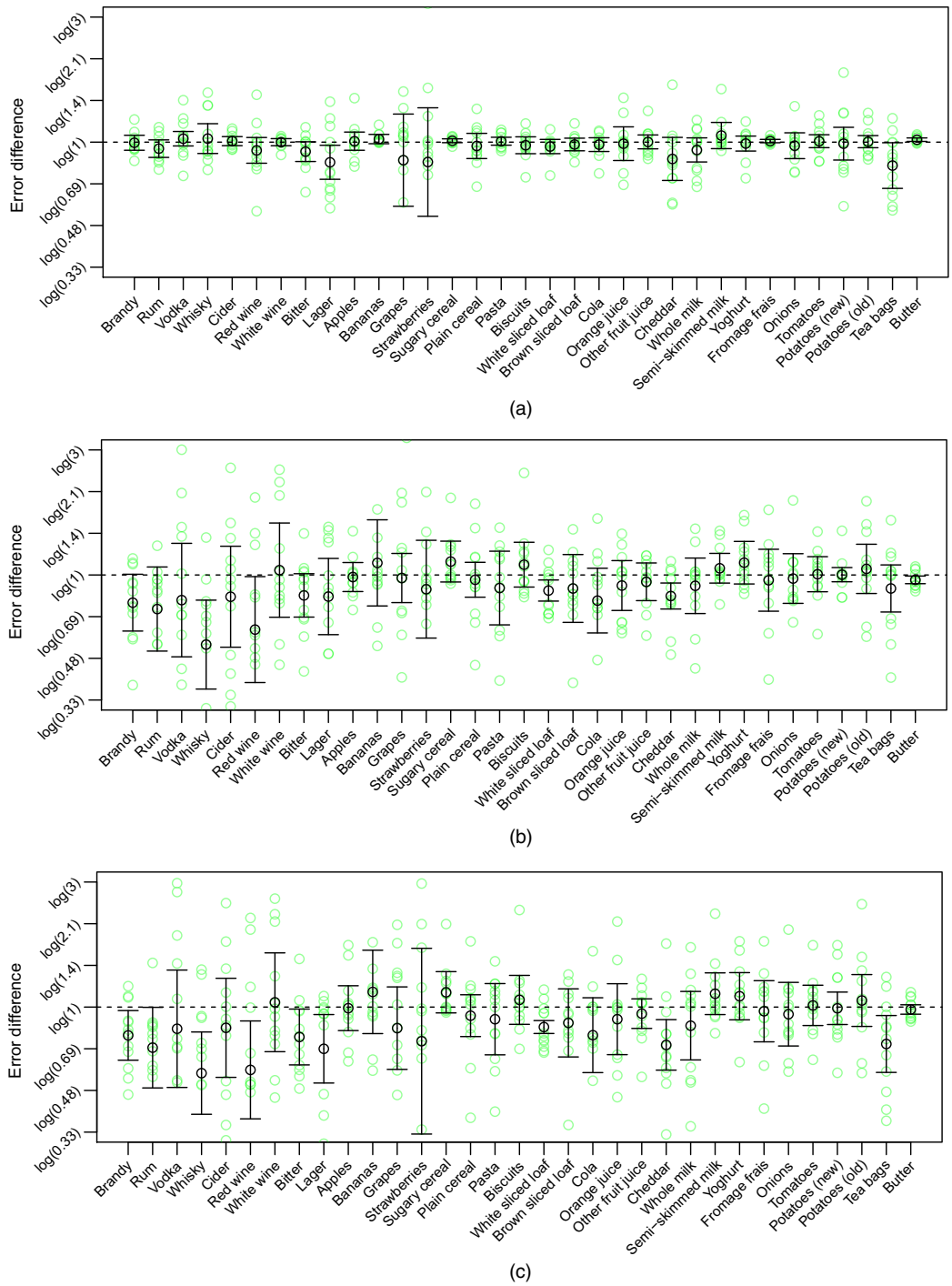

Fig. 5. Loss of uninformed model minus (a) loss of persistence model, (b) loss of uninformed model and (c) loss of persistence model (⌶, approximate 95% confidence intervals for the long-term mean of the error differences between models for each product category; ○, sample means; ○, differences themselves): for each subfigure, values below zero correspond to the first-named model outperforming the second

empirical autocovariances, $z_{0.975} \approx 1.96$ is the 97.5% percentile of the unit normal distribution and $n = 12$ is the number of published CPI figures that are available for us to calibrate to. When reading these plots it is helpful to remember that values of $Z_{\mathrm{a,b},t}$ less than 0 correspond to instances when forecast a outperforms forecast b in terms of the specified loss function. The confidence intervals for the long-term means in Figs 5(b) and 5(c) that exclude zero provide strong evidence of the potential for web-scraped data to improve month-ahead log-CPI forecasts for certain product categories. The presence of intervals that do include zero show how, for other products, producing such evidence will require longer time series of data to be acquired. Nevertheless, the greater part of the majority of the intervals are below zero, suggesting that future data are more likely than not to confirm the relevance of the web-scraped data.

4.3. Remarks on inferred hyperparameters for product categories

It is clear from the plots in Fig. 4 that the log-prices of different products display different dynamic behaviour and are therefore predictable to varying degrees. These behaviours are reflected in the optimized category-specific hyperparameters that are reported in Table 1. The hyperparameters that are not reported in Table 1 were considered common to all categories. Their values, along with prior constraints on the category-specific hyperparameters, are detailed in Appendix A.3.

Although differences in price behaviour are ultimately of great interest to consumers and policy makers, we concentrate now only on their contributions to the predictability of published CPI statistics. We begin by noting how the published log-index for some product categories is inherently more predictable than others, because primarily for some product categories prices vary only very gradually. Differences in this respect are captured in the values for the variance of the g_t-processes for different product categories. We also note how the inherent predictability of product-specific indices is conceptually distinct from the improvement in predictability that can be achieved by utilizing either historical values of the published CPI or the relevant web-scraped data. Given that there is a large amount of short-term variation to be explained, we expect the improvement in predictability to be significant when either the g_t-process is smooth or the correlation between the on-line and survey-based processes is high and many prices are scraped.

To give an example for the type of inference that the model allows for, we draw attention to the following, more specific, observations gleaned from the hyperparameter estimates of Table 1 and the robust error statistics.

(a) Short-term log-inflation, as quantified by the long-term variance of g_t, for onions, milk and apples is very low, whereas for strawberries and whisky it is highest. We note that large values of g_t correspond not to large values of the log-CPI in Figs 3 and 4, but to large values of its gradient. These inferences suggest the greater importance of strawberries and whisky to short-term aggregate inflation measures.

(b) The ϕ_g-hyperparameter tells us about the stability of the short-term log-inflation process and reflects the extent to which log-CPI can be extrapolated from day to day. High inferred values suggest that the log-prices of strawberries and white bread vary most smoothly in the short term, whereas white wine and lager vary less predictably. These inferences suggest that high frequency price data for strawberries and bread contain less extra information than those for wine and lager.

(c) The survey-based and on-line short-term log-inflation processes for alcoholic spirits appear to be the most highly correlated, whereas sugary cereal and yoghurt show the least correlation. These statements, informed by the inferred values of the hyperparameter ρ_g,

imply that our web scrapers are successfully identifying representative price data for spirits but not for cereal and yoghurt.

(d) The consequences of the high correlations in log-inflation are borne out by the mean decreases in forecast error after accounting for the web-scraped data. Fig. 5(b) illustrates the degree and consistency of the decreases, corroborating the value of the web scrapers for alcoholic drinks, pasta and white bread in particular.

5. Discussion

With the work that is presented in this paper, we have demonstrated that the log-CPI for individual product categories are commonly at least partially predictable from web-scraped price data. In doing so, we have produced an uncertainty-qualified product-specific daily CPI forecast that is not seen elsewhere in published official statistics. As a by-product of the model fitting that precedes the forecasting, our methodology also produces estimates for hyperparameters encoding the variability of prices for different products. These estimates can help us to identify products that are particularly influential in changes to aggregated CPI quantities, and also to identify products whose web-scraped prices correlate well, or badly, with survey prices.

In the medium term we hope to scrape prices for a wider range of product categories, enabling us to produce estimates for more highly aggregated CPI quantities. We also plan to link the scraping, estimate adjustment and plotting codes to produce graphical forecasts that update in real time. In the longer term, we also hope to acquire direct or inferred information regarding the quantity of product sales. This information will then allow us to weight observed prices in a more meaningful way. Requesting quantity data directly from retailers is a natural next step, but one that is likely to require some negotiation.

In terms of methodology, we consider the following model extensions to be particularly promising: linking products via a hierarchical structure to recognize the tendency of prices to move in complementary or antagonistic modes; the incorporation of series discontinuities corresponding to significant changes in pricing policy (such as changes in value-added tax rates), the consideration of higher order seasonal auto-regressive integrated moving average models for describing the log-inflation process and the development of more realistic distributions for price changes, which ought to render our likelihood calculations more relevant and trustworthy. In particular, we suspect that significant improvements in the correspondence between web-scraped and survey-based CPI figures could be attained if the model were better equipped to accommodate or anticipate product promotions.

Our work has also highlighted some computational issues that are worthy of further attention. Firstly, we would like to develop a more satisfactory solution to the maximization problem through which we estimated our model's hyperparameters. It is possible that the maxima that are involved here will become more distinct if we insist that certain groups of products share common hyperparameters. This could be achieved if we were to invest in developing a more sophisticated hierarchical model structure, as mentioned above. A second issue is the efficient storage of price data. Although, in a time of big data problems, it is easy to find statisticians who are tolerant of, or even excited by, data sets that place a large burden on their computing resources, we can identify trivial ways to compress our data that speed up our computations enormously. Within each product category, our model looks at only daily means of log-price changes to adjust its forecasts. Summarizing the data like this commonly results in a compression of two or three orders of magnitude. The third computational issue relates to the web-scraping software itself. Although we have, so far, implicitly described model noise terms as being part of the pricing regime, we must recognize that this is only partly true. A significant amount of noise

can also be attributed to our web scrapers feeding inappropriate values in our data sets. Work on this topic is already under way at the ONS (Mayhew and Clews, 2016), but additional, and perhaps continuous, effort needs to be assigned to the identification of spurious web-scraped data.

At a more general level, we intend the findings that are documented in this paper to contribute to the discussion on the utility of web-scraped price data for index production. We have considered the relationship between web-scraped prices and conventional survey-based indices because it can be quantified, documented and scrutinized. There is potential, however, for Web-based index calculations to move beyond what can realistically be achieved with surveys: potential, for example, to produce high frequency index values for very specific combinations of products. At this point we lose the 'gold standard' official statistics against which to check the accuracy of our new index values. Our research suggests that, for some types of product, on-line prices are relatively stable and are representative of the prices that intelligent surveyors do or would collect. For others, high variability and the difficulty in scraping representative prices may undermine indices based on web-scraped data.

Having established that we can, to a quantifiable degree, compute high frequency estimates for the CPI by using web-scraped data, the question remains whether the ONS should do so. Although the answer is likely to depend on more considerations than can be fully addressed here, it is worth reviewing some of them briefly to describe the context to our statistical findings. As mentioned previously, high frequency and timely estimation of the CPI offers many potential benefits but also raises some possible risks. The CPI is a key indicator used by the Bank of England's Monetary Policy Committee and, although it is currently one of the most timely publications of data by the ONS, with estimates published in the second or third week following the reference period, providing estimates in as near to realtime as possible could enable more timely interventions in monetary policy. Equally markets can react to the publication of CPI data—see for example Jain (1988), or Mitchell and Mulherin (1994)—and so having a more regular publication may affect the way in which traders respond to information on inflation. Moreover, current publications of the CPI do not include measures of uncertainty, and so the inclusion of prediction intervals may also be useful for users in taking appropriate action based on available information. Of course, these are areas for further research and as most attention is given to the aggregate level CPI our research is one step in that direction.

We should note that there are also potential risks in using the web-scraped data, as they create an opportunity for on-line retailers to set prices specifically to influence the CPI. This represents a data collection issue that is to some extent separate from the issues that have been considered so far but is nevertheless an important problem to address. There is also a risk of providing inaccessible information to the user of statistics. For example, users are typically used to a published point-in-time estimate of inflation. Advanced users may be comfortable with interpreting prediction intervals, but others may not. In terms of the best presentation of information from our model there are many decisions to be made and engagement with users of the data will be important if it is to form part of a regular output. Ultimately we believe that providing timely and more detailed estimates is fully consistent with the principles and protocols of the code of practice for official statistics (UK Statistics Authority, 2009) and so contributes towards providing a coherent and trustworthy service to users of statistics.

Finally, it is worth noting that the CPI is not only an important indicator used by a range of users who are interested in the economy and society but is also used in the production of other official statistics such as national accounts. Increasing the timeliness of the CPI therefore provides an opportunity to explore increasing the timeliness of other important economic indicators.

6. Data access statement

We are unable to make public the data analysed in this paper. ONS policy informing the sharing of web-scraped data, in particular those of potential commercial sensitivity, is, at the time of writing, still being refined. Unfortunately, until such a policy is finalized no sharing can be authorized.

Acknowledgements

The authors are grateful for comments and advice made by staff at the ONS, particularly from the Index Number Methodology, Time Series Analysis, Prices Economic Analysis and Prices Development teams. Credit is due also to two reviewers whose comments and suggestions improved the paper substantially. BP and GPN gratefully acknowledge support by Engineering and Physical Sciences Research Council grant EP/K020951/1.

Appendix A: Model fitting and parameter adjustment

A.1. Adjusting estimates for the log-inflation processes

A.1.1. Parameter initialization

Running our estimation algorithm requires first specifying initial values for the expectation and variance of the state vector $\mathbf{u}_t$. The theoretical long-term means and variances of the log-inflation subprocesses are finite and computable (see chapter 4 of Canova (2007) for example). Those for the aggregated, random-walk-type quantities d_t^{web} and d_t^{surv} are set to 0, according to their definitions, when the state vector reaches the base period t_0.

A.1.2. Parameter propagation

Given current values for the expectation and variance for the state vector, we propagate them forwards in accordance with expression (8) so that

$$E(\mathbf{u}_t) = \mathbf{G}\, E(\mathbf{u}_{t-1}), \tag{14}$$

$$\operatorname{var}(\mathbf{u}_t) = \mathbf{G}\, \operatorname{var}(\mathbf{u}_{t-1})\mathbf{G}^{\mathrm{T}} + \operatorname{var}(\mathbf{e}_t). \tag{15}$$

where $E(\cdot)$ and $\operatorname{var}(\cdot)$ denote the expectation and variance operators, and

$$\mathbf{u}_t = \begin{pmatrix} \tau_t^{\text{web}} \\ g_t^{\text{web}} \\ d_t^{\text{web}} \\ \tau_t^{\text{surv}} \\ g_t^{\text{surv}} \\ d_t^{\text{surv}} \end{pmatrix}, \qquad \mathbf{e}_t = \begin{pmatrix} \epsilon_t^{\text{web}} \\ \varepsilon_t^{\text{web}} \\ 0 \\ \epsilon_t^{\text{surv}} \\ \varepsilon_t^{\text{surv}} \\ 0 \end{pmatrix}, \tag{16}$$

$$\mathbf{G} = \begin{pmatrix} \phi_\tau & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_g & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi_\tau & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_g & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}, \qquad \operatorname{var}(\mathbf{e}_t) = \begin{pmatrix} \sigma_\tau^2 & 0 & 0 & \sigma_\tau^2\rho_\tau & 0 & 0 \\ 0 & \sigma_g^2 & 0 & 0 & \sigma_g^2\rho_g & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_\tau^2\rho_\tau & 0 & 0 & \sigma_\tau^2 & 0 & 0 \\ 0 & \sigma_g^2\rho_g & 0 & 0 & \sigma_g^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{17}$$

with the implication that

$$\begin{aligned} \sigma_\tau^2 &= \operatorname{var}(\epsilon_t^{\text{web}}) = \operatorname{var}(\epsilon_t^{\text{surv}}), \\ \sigma_g^2 &= \operatorname{var}(\varepsilon_t^{\text{web}}) = \operatorname{var}(\varepsilon_t^{\text{surv}}), \end{aligned} \tag{18}$$

$$\begin{aligned} \rho_\tau &= \operatorname{corr}(\epsilon_t^{\text{web}}, \epsilon_t^{\text{surv}}), \\ \rho_g &= \operatorname{corr}(\varepsilon_t^{\text{web}}, \varepsilon_t^{\text{surv}}). \end{aligned} \tag{19}$$

A.1.3. Learning from log-price movements

Learning from daily log-price movements involves first computing the mean of log-price differences

$$\bar{y}_{\text{diff},t} = |\mathcal{I}_t|^{-1} \sum_{i \in \mathcal{I}} (y_{i,t} - y_{i,t-1}) \tag{20}$$

where $\mathcal{I}_t$ denotes the set of indices identifying the products with observed log-price changes at time t. The second-order properties of $\bar{y}_{\text{diff},t}$ and its relationship to the state vector are encoded in the quantities

$$E(\bar{y}_{\text{diff},t}) = h^{\mathrm{T}} E(\mathbf{u}_t), \tag{21}$$

$$\text{var}(\bar{y}_{\text{diff},t}) = h^{\mathrm{T}} \text{ var}(\mathbf{u}_t) h + |\mathcal{I}_t|^{-1} \text{var}(e_{i,t}), \tag{22}$$

$$\text{cov}(\mathbf{u}_t, \bar{y}_{\text{diff},t}) = h^{\mathrm{T}} \text{ var}(\mathbf{u}_t) \tag{23}$$

where $h = (1, 1, 0, 0, 0, 0)$. Given these, we can update our estimates for $\mathbf{u}_t$ by using the adjustment equations

$$E(\mathbf{u}_t) \leftarrow E(\mathbf{u}_t) + \text{cov}(\mathbf{u}_t, \bar{y}_{\text{diff},t}) \text{ var}(\bar{y}_{\text{diff},t})^{-1} \{\bar{y}_{\text{diff},t} - E(\bar{y}_{\text{diff},t})\}, \tag{24}$$

$$\text{var}(\mathbf{u}_t) \leftarrow \text{var}(\mathbf{u}_t) - \text{cov}(\mathbf{u}_t, \bar{y}_{\text{diff},t}) \text{ var}(\bar{y}_{\text{diff},t})^{-1} \text{cov}(\mathbf{u}_t, \bar{y}_{\text{diff},t})^{\mathrm{T}}. \tag{25}$$

Equations (24) and (25) describe estimate adjustments that minimize expected squared error in a sense that is elaborated on in chapter 3 of Goldstein and Wooff (2007). The same equations feature in section 4.3 of West and Harrison (1997), although they are encoded in more abstract notation.

A.1.4. Learning from the published consumer price index

If an observation of the published log-CPI for the product category, d_t^{surv}, is also available, we can update the moments for $\mathbf{u}_t$ again. Doing so requires another application of the linear adjustment equations

$$E(\mathbf{u}_t) \leftarrow E(\mathbf{u}_t) + \text{cov}(\mathbf{u}_t, d_t^{\text{surv}}) \text{ var}(d_t^{\text{surv}})^{-1} \{d_t^{\text{surv}} - E(d_t^{\text{surv}})\}, \tag{26}$$

$$\text{var}(\mathbf{u}_t) \leftarrow \text{var}(\mathbf{u}_t) - \text{cov}(\mathbf{u}_t, d_t^{\text{surv}}) \text{ var}(d_t^{\text{surv}})^{-1} \text{cov}(\mathbf{u}_t, d_t^{\text{surv}})^{\mathrm{T}}. \tag{27}$$

where

$$E(d_t^{\text{surv}}) = h^{\mathrm{T}} E(\mathbf{u}_t), \tag{28}$$

$$\text{var}(d_t^{\text{surv}}) = h^{\mathrm{T}} \text{ var}(\mathbf{u}_t) h + \sigma_{\text{round}}^2, \tag{29}$$

$$\text{cov}(\mathbf{u}_t, d_t^{\text{surv}}) = h^{\mathrm{T}} \text{ var}(\mathbf{u}_t), \tag{30}$$

with $h = (0, 0, 0, 0, 0, 1)$ and $\sigma_{\text{round}} = 0.01$ approximately accounting for errors in published CPI figures.

A.2. Computational cost of parameter adjustment

It may be shown that the adjustment calculations for the latent process involve only matrix calculations with computational costs scaling with the cube of the state vector's size. Scaling of the computational cost as the time interval of interest increases is linear. These are two factors that substantiate the resulting algorithm's claims of computational efficiency and ought to be compared with general, unstructured Kriging-type calculations whose cost typically scales with the cube of the total number of data quantities. High computational speed is of great importance when we recompute the calculations with different hyperparameters to identify appropriate values for them.

A.3. Inference for model hyperparameters

For this work we have chosen to partition the set of model hyperparameters into two sets, one of which contains hyperparameters that are considered to be common to all product categories, and another containing

those specific to product categories, which are identified with the index j below:

$$\Omega_j = \{\Omega^{\text{common}}, \Omega_j^{\text{specific}}\}, \tag{31}$$

$$\begin{aligned} \Omega^{\text{common}} &= \{\phi_\tau, \sigma_\tau, \rho_\tau, \text{var}(e_{i,t})t\}, \\ \Omega_j^{\text{specific}} &= \{\phi_{g,j}, \sigma_{g,j}, \rho_{g,j}\}. \end{aligned} \tag{32}$$

The common hyperparameters are specified at values

$$\begin{aligned} \phi_\tau &= 0.9995, \\ \sigma_\tau &= (1 \times 10^{-4})\sqrt{(1-\phi_\tau)}, \end{aligned} \tag{33}$$

$$\begin{aligned} \rho_\tau &= 0.8, \\ \sqrt{\text{var}(e_{i,t})} &= 1 \times 10^{-2} \end{aligned} \tag{34}$$

on the basis of our beliefs about inflation processes, whereas the category-specific hyperparameters are adjusted to optimize a constrained log-likelihood function approximately,

$$f(D_j \mid \Omega_j^{\text{specific}}) = \log\{\pi(\Omega_j^{\text{specific}})\} + \sum_{t \in \Omega_{\text{pub}}} \log\{\pi(d_t^{\text{surv}} \mid D_{j,t-1}, \Omega^{\text{common}}, \Omega_j^{\text{specific}})\}, \tag{35}$$

in a manner equivalent to a Bayesian maximum *a posteriori* estimation procedure. The symbol D_t in equation (35) is used to denote all the observational data that are available up to time t. The likelihood in question is seen as an approximate marginal likelihood for the hyperparameters, from which the latent process values have been integrated out. More specifically, we choose to quantify the marginal likelihoods within the sum in equation (35) by using t-distributions, translated and scaled to match the adjusted expectations that are described in Sections A.1.3 and A.1.4.

The constraint on the optimization, which is manifested in the first term on the right-hand-side of equation (35), is interpretable as a log-prior distribution on the hyperparameters for all product categories. Our chosen prior is a product of independent normal priors on transformed versions of the relevant hyperparameters:

$$\pi(\Omega_j^{\text{specific}}) = \pi_\phi\{\Phi^{-1}(\phi_{g,j}/\phi_\tau)\}\, \pi_{\sigma,\phi}\{\log(\sigma_{g,j}) - \log(1-\phi_{g,j})/2\}\, \pi_\rho\{\Phi^{-1}(\rho_{g,j})\}, \tag{36}$$

$$\pi_\phi(x) = N\left\{E(x) = \Phi^{-1}(0.6), \text{var}(x) = 1\right\}, \tag{37}$$

$$\pi_{\sigma,\phi}(x) = N\left\{E(x) = \log(1 \times 10^{-3}), \text{var}(x) = 1\right\}, \tag{38}$$

$$\pi_\rho(x) = N\left\{E(x) = \Phi^{-1}(0.5), \text{var}(x) = 1\right\}, \tag{39}$$

where $\Phi^{-1}(\cdot)$ is the quantile function for the unit normal distribution and the quantity $\log(\sigma_{g,j}) - \log(1-\phi_{g,j})/2$ is the long-term log-standard deviation of the $g_{t,j}$-process. The prior distributions approximately encode our beliefs for the hyperparameters' values. They also mitigate the danger of overfitting to a likelihood that approximates only the type of correspondence between estimates and observations that we would expect. Another reason for constraining the optimization is that a moderately high dimensional search over the surface of a complex objective function is, in general, still a considerable computational challenge. On grounds of pragmatism, therefore, we force our optimizers to look closely only at hyperparameter values close to those which we consider *a priori* sensible.

In presenting the expectations and variances for the latent process conditionally on the optimized hyperparameters as our forecasts, our approach to predicting log-inflation may be categorized as an empirical Bayes method. Background and further discussion of such methods can be found in Casella (1985) and Efron (2012). As with almost every part of the modelling exercise, our chosen strategy for hyperparameter specification is open to criticism and modification. Informal experimentation has led us to believe that the strategy that is presented here represents a practicable compromise between bias and variance, leading to inferences that are robust and in accordance with prior beliefs about price changes.

References

Ang, A., Bekaert, G. and Wei, M. (2007) Do macro variables, asset markets, or surveys forecast inflation better? *J. Monet. Econ.*, **54**, 1163–1212.

Ang, A. and Piazzesi, M. (2003) A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *J. Monet. Econ.*, **50**, 745–787.

Aparicio, D. and Bertolotto, M. (2016) Forecasting inflation with online prices. *Technical Report*. Department of Economics, Massachusetts Institute of Technology, Cambridge.

Boivin, J., Bernanke, B. and Eliasz, P. S. (2005) Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.*, **120**, 387–422.

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, N., Payne, C., Smith, T., Winton, J. and Woods, A. (2016) Research indices using web-scraped data: May 2016 update. *Technical Report*. Office for National Statistics, Newport.

Bunn, P. and Ellis, C. (2009) Price-setting behaviour in the United Kingdom: a microdata approach. *Bnk Engl. Q. Bull.*, **49**, 28–36.

Canova, F. (2007) *Methods for Applied Macroeconomic Research*, vol. 13. Princeton: Princeton University Press.

Casella, G. (1985) An introduction to empirical Bayes data analysis. *Am. Statistn*, **39**, 83–87.

Cavallo, A. and Rigobon, R. (2016) The Billion Prices Project: using online prices for measurement and research. *J. Econ. Perspect.*, **30**, 151–178.

Chui, C. and Chen, G. (2009) *Kalman Filtering: with Real-time Applications*. Berlin: Springer.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *J. Bus. Econ. Statist.*, **13**, 253–263.

Efron, B. (2012) *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1. Cambridge: Cambridge University Press.

Faust, J. and Wright, J. (2013) *Handbook of Economic Forecasting*, vol. 2, ch. 1, pp. 3–56. Amsterdam: Elsevier.

Goldstein, M. and Wooff, D. (2007) *Bayes Linear Statistics, Theory and Methods*. Chichester: Wiley.

Jain, P. C. (1988) Response of hourly stock prices and trading volume to economic news. *J. Bus.*, **61**, 219–231.

Lünnemann, P. and Wintr, L. (2011) Price stickiness in the US and Europe revisited: evidence from internet prices. *Oxf. Bull. Econ. Statist.*, **73**, 593–621.

Mayhew, M. and Clews, G. (2016) Using machine learning techniques to clean web-scraped price data via cluster analysis. In *Survey Methodology Bulletin*. Newport: Office for National Statistics.

Mitchell, M. L. and Mulherin, J. H. (1994) The impact of public information on the stock market. *J. Finan.*, **49**, 923–950.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–313.

Office for National Statistics (2014) *Consumer Prices Indices: Technical Manual*. Newport: Office for National Statistics.

Pukelsheim, F. (1994) The three sigma rule. *Am. Statistn*, **48**, 88–91.

Ralph, J., O'Neill, R. and Winton, J. (2015) *A Practical Introduction to Index Numbers*. Chichester: Wiley.

R Development Core Team (2009) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Smets, F. and Wouters, R. (2007) Shocks and frictions in US business cycles: a Bayesian DSGE approach. *Am. Econ. Rev.*, **97**, 586–606.

UK Statistics Authority (2009) *Code of Practice for Official Statistics*. London: UK Statistics Authority.

West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, vol. 18. New York: Springer.