Distr. GENERAL

8 October 2012

WP. 4
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on New Frontiers for Statistical Data Collection
(Geneva, Switzerland, 31 October-2 November 2012)

Topic (i): New data sources

AN INVESTIGATION INTO USING GOOGLE TRENDS AS AN ADMINISTRATIVE DATA SOURCE IN ONS

Contributed Paper

Prepared by Daniel Ayoubkhani, Office for National Statistics, UK

I. Introduction

1. There is a vast array of measures of activity outside of official statistics that provide a picture of the economic climate. Some of these indicators are timelier than official statistics and so may be utilised to help predict official time series, or act as a means of quality assurance of these series.

2. One such source of data is Google Trends, which provides data on the volume of search terms entered into the Google search engine. This data source provides a clear account of what both consumers and firms are browsing the internet for. The Office for National Statistics (ONS) publishes official statistics on individuals' annual internet access; there has been a continued increase, with 0.9 million more individuals accessing the internet in 2009 compared to 2008, and 5million more in 2009 compared to 2006. The continued increase in internet access means that data on internet searches potentially have ever more strength as an indicator of individuals' economic intentions.

3. The internet is an expanding market place; ONS collects annual data on e–commerce which estimates the value of e–commerce sales to have increased by almost 25 per cent from 2008 to 2009. Moreover, it was estimated that by the end of 2009 approximately 15 per cent of businesses were using a website to sell their goods and services. An even stronger argument for the internet as an expanding and representative market place is that 52 per cent of firms purchased goods or services via the internet.

4. Therefore, with such an abundance of indicators of internet search activity available soon after the end of a reference period, there is strong motivation for an investigation into whether such data can be used for: (i) quality assuring; (ii) nowcasting (forecasting the current state); and (iii) supplementing/replacing official time series data in the UK.

5. This paper reports the methods and findings of an investigation focusing on the UK Retail Sales Index (RSI) published by ONS. Nowcasting methods are not frequently used by ONS during the RSI production process due to the relative timeliness of the input data. This investigation therefore attempts to establish whether or not Google Trends data might be a useful source of information for *quality assuring* RSI outputs prior to publication, and adds to preliminary

research, using the same data sources, conducted by Chamberlin (2010). If strong relationships can be found between RSI and selected Google Trends time series, these relationships could be exploited by ONS in order to validate estimates of retail sales activity before such estimates are published.

6. In the remainder of this paper: sections 2 and 3 describe data and methods respectively; section 4 reports results; section 5 proposes caveats and considerations that should be acknowledged alongside the results; and section 6 provides conclusions.

II. Data

A. Google Trends data

7. Google Trends data do not report the raw levels of search queries processed by Google, but rather a query index (Choi and Varian 2009). The query index is constructed from a query share, the definition of which provided by Choi and Varian is "the total query volume for a search term in a given geographic region divided by the total number of queries in that region at a point in time". Once the data are transformed into this format they are normalised so that the query share is 0 for the first full week of January 2004. Values at later dates report the percentage deviation from the query share for this week.

8. Google Trends data are classified into categories, with each search query assigned to a particular category by natural processing methods. The queries are classified into 27 top level categories and 241 categories at the second level (Choi and Varian 2009). Chamberlin (2010) provides an example classification: the query "car tyre" would be placed in the "Vehicle Tyres" category, which in turn is a subset of "Auto Parts", which is part of the top level "Automotive" category.

9. When extracting data from Google Trends[1], the data are presented as a weekly query index which indicates the percentage deviation from the date to which the data are normalised, as discussed above. However, for the purpose of the analysis reported in this paper, the levels of the data, rather than the weekly growth rates, were required. Therefore, an index was constructed by setting the value for the level of search activity in the first full week of January 2004 to 100 and then applying growth rates to all periods after this.

10. The primary benefit of using Google Trends data is that they are a timely source of information, but this also poses a problem. The Google Trends data are available on a weekly basis, but the analysis presented in this paper required data of a monthly frequency; therefore the Google Trends data had to be aggregated temporally. The method of aggregation adopted in this paper is a weighted average – for a given search query or category, each week's index is weighted according to the proportion of the week that falls in the target month. Therefore a week that falls completely in the target month will be given the weight of 7 divided by the total number of days in the month. Weeks that straddle two months are weighted by the number of days that fall in the target month divided by the total number of days in the month. Obviously this makes a strong assumption: that search behaviour is constant across days of the week (for example, search behaviour on a Tuesday is the same as that on a Saturday). Chamberlin (2010) outlines a further potential drawback; the search query reports a relative volume, so that a falling query share might represent slower than average growth in a search term rather than a decline in the raw volume of the query. Hence, the constructed index will decline when relative search levels have fallen, which may disguise an increase in the raw levels of a search query.

B. Retail Sales Index data

11. The analysis utilised a breakdown of the monthly RSI dataset for Great Britain, disaggregated in the same way as in Chamberlin (2010), in addition to the aggregate All Retail Sales time series:
    - Non–Specialised Food Stores

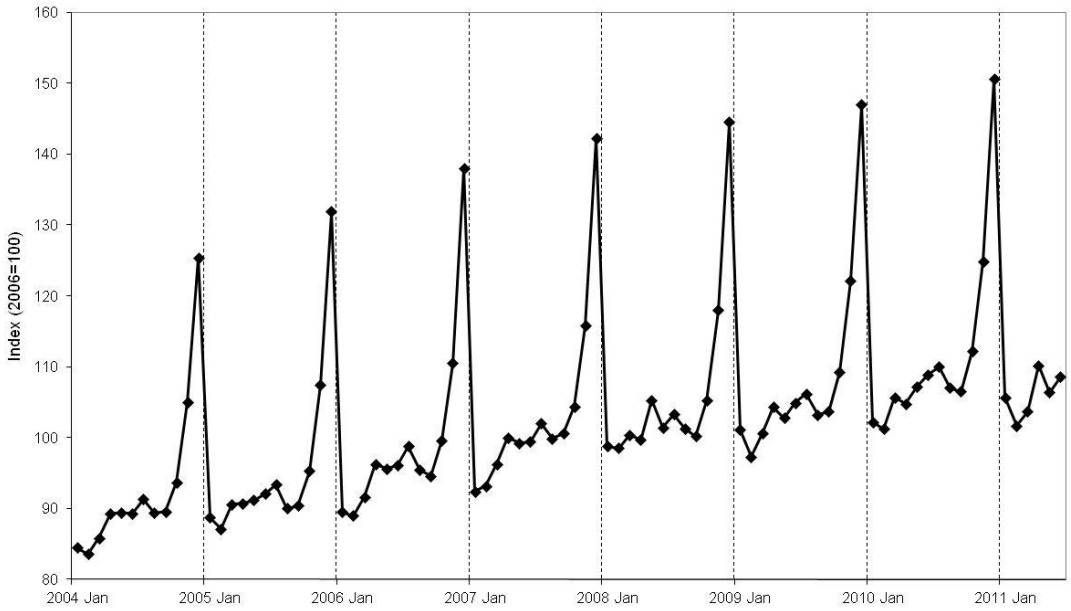
[1] http://www.google.com/insights/search

- Non–Specialised Non–Food Stores
- Textiles, Clothing and Footwear
- Furniture and Lighting
- Home Appliances
- Hardware, Paints and Glasses
- Audio and Video Equipment and Recordings
- Books, Newspapers and Stationery
- Computers and Telecommunications
- Non–Store Retailing

12. The RSI's 'large business index' was used for: Non–Specialised Food Stores; Non–Specialised Non–Food Stores; Textiles, Clothing and Footwear; Non–Store Retailing, as well as the aggregate All Retail Sales series. The analysis utilised non–seasonally adjusted chained volume data, presented as an index with 2006 set equal to 100. The data cover the span January 2004 (for consistency with the Google Trends data) to June 2011. All of the data are freely available via the ONS website[2].

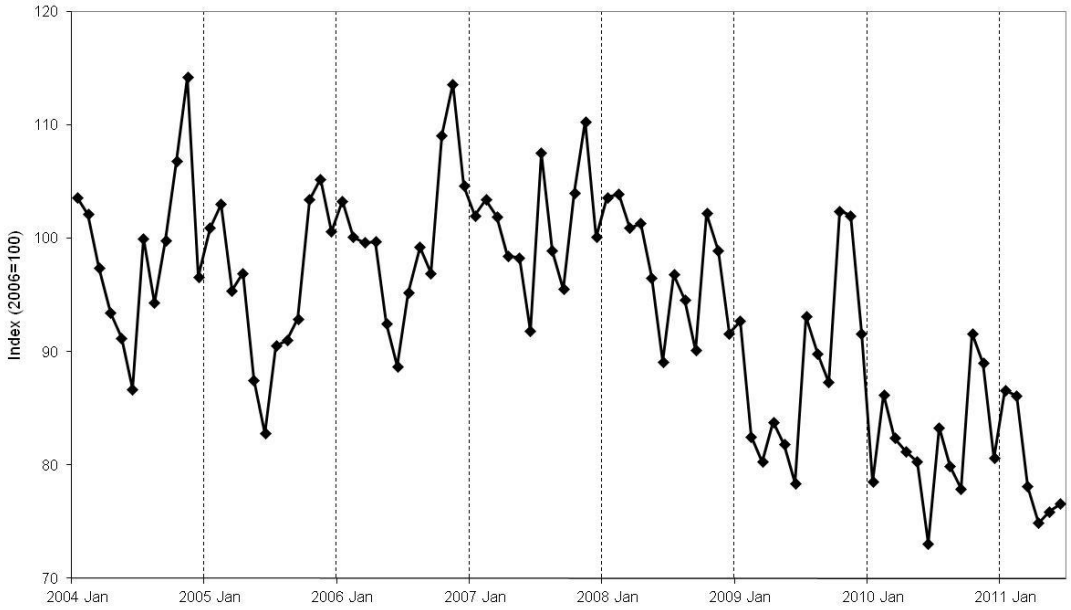
13. Many of the 11 RSI time series exhibit a clear trend and stable seasonal behaviour; for example, figure 2.1 illustrates the aggregate All Retail Sales series. Other series are more volatile, and may exhibit multiple turning points in their trend; for example, figure 2.2 illustrates the Furniture and Lighting series.

Figure 2.1 All Retail Sales (non–seasonally adjusted), January 2004 – June 2011



[2] http://www.ons.gov.uk/ons/rel/rsi/retail-sales/july-2011/tsd-retail-sales.html

Figure 2.2 Furniture and Lighting (non–seasonally adjusted), January 2004 – June 2011



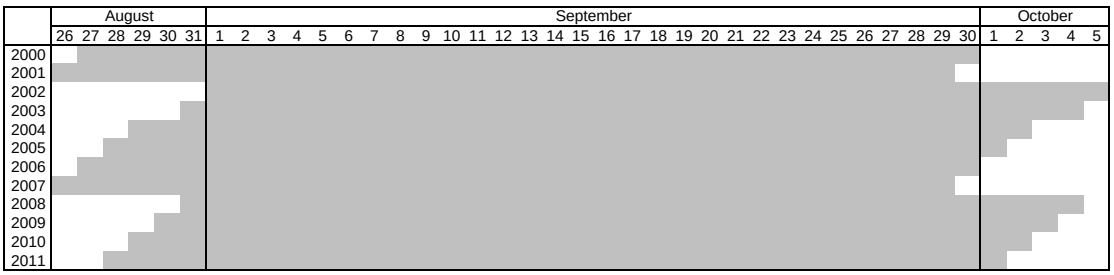
14. The analysis reported on in this paper used the same matching of RSI time series to Google Trends search categories as Chamberlin (2010), as well as the five most popular search queries (at the time of the analysis) within each of the categories. For example, the RSI time series Non–Specialised Food Stores was matched with the Google Trends search category Food Retailers, as well as the search queries "tesco", "asda", "morrisons", "sainsburys" and "tescos".

III. Methodology

A. Fitting the benchmark models

15. The calendar basis on which the RSI data are collected added a complication to the modelling procedure. RSI does not follow Gregorian calendar months, but rather four– or five–week long standard recording periods (SRPs). RSI is said to operate on a "4–4–5" basis; a repeating pattern consisting of a four–week long SRP, followed by another four–week long SRP, followed by a five–week long SRP. This design leads to 364 days per survey calendar year. Consequently, the disparity between the survey years and the Gregorian calendar years (and each of the months within them) evolves by one or two days every year; this evolving disparity is known as the "phase shift". The phase shift is cyclical in its behaviour, with each cycle assumed to last 28 years. In order to illustrate the phase shift, figure 3.1 below shows the movement of the September SRP, relative to the Gregorian calendar, between 2000 and 2011.

Figure 3.1 Arrangement of the september SRP, 2000 – 2011



16. The 4–4–5 structure of RSI can induce several calendar effects that may need to be accounted for when modelling the data. These effects may be categorised as being either direct or indirect. The direct effect arises from the fact that the composition of each SRP, in terms of the number of days from Gregorian calendar months, changes between successive years. The indirect effects are related to holiday periods in Great Britain and, more specifically, those holidays that "move" between SRPs from year–to–year due to the phase shift; these holidays are summarised in table 3.1 below. The direct effect is the result of seasonality, whilst the indirect effects are attributable to the fact that the presence of a holiday may affect the level of retail activity. More generally, estimates of retail activity will not be comparable between successive periods unless the effects of the phase shift, both direct and indirect, are accounted for.

Table 3.1 Holiday periods in Great Britain that move between SRPs

Holiday	Position of holiday	SRPs holiday can fall in
Easter	Good Friday and Easter Monday	March, April
Spring (late May)	Last Monday in May	May, June
Summer (late August)	Last Monday in August	August, September

17. For each of the RSI series, the direct and indirect effects of the phase shift were accounted for via the application of regARIMA (regression with Autoregressive Integrated Moving Average errors) modelling:

$$\varphi_p(B)\Phi_P(B^{12})u_t = \theta_q(B)\Theta_Q(B^{12})\varepsilon_t$$

where: B is the backshift operator (i.e. $Bz_t = z_{t-1}$, $B^2z_t = z_{t-2}$); $\varphi_p(B)$ is the non–seasonal AR operator of order p; $\Phi_P(B^{12})$ is the seasonal AR operator of order P; $\theta_q(B)$ is the non–seasonal MA operator of order q; $\Theta_Q(B^{12})$ is the seasonal MA operator of order Q; ε_t is an i.i.d. residual in month t with mean zero and variance σ^2; and u_t is the regression error in month t from the linear relationship:

$$(1-B)(1-B^{12})\ln y_t = \sum_i \beta_i(1-B)(1-B^{12})x_{it} + u_t$$

where: $(1-B)(1-B^{12})\ln y_t$ is an RSI time series after a log transformation and first order regular and seasonal differencing have been applied in order to achieve weak stationarity (see Box and Jenkins (1976)); $(1-B)(1-B^{12})x_{it}$ is the value of variable i (explained below) after first order regular and seasonal differencing have been applied; and β_i is the estimated regression coefficient corresponding to variable i.

18. The explanatory variables in the regression model are indicator variables constructed in such a way as to represent the direct and indirect effects of the phase shift. For example, the variable m_t – used to model the late May bank holiday effect – was constructed in the following way:

$$m_t = \begin{cases} 1 & \text{In May, in years where the bank holiday is in the May SRP} \\ -0.8 & \text{In June, in years where the bank holiday is in the May SRP} \\ 0 & \text{Otherwise} \end{cases}$$

19. The value –0.8 (rather than –1) was used because June is a five–week long SRP whilst May is a four–week long SRP.

20. In total, there were six direct phase shift variables (treated as a single effect), three Easter variables (treated as a single effect), and one variable each for the late May and late August bank holidays (treated as two separate effects). When the whole set of variables is considered, it allows for 15 combinations of effects plus one empty set. Each of these 16 regression models was fitted to each of the 11 RSI series. Simultaneously, the autocorrelation structure (the AR and/or MA orders) of the regression residuals was automatically identified for each model and each series using the 'automdl' procedure implemented in the U.S. Census Bureau's X–12–ARIMA software package (U.S. Census Bureau 2009). Model estimation was conducted via an iterative generalised least squares routine (Findley et al. 1988). Additive outliers and level shifts were also automatically identified in each of the affected series, and estimated in each of the relevant models, by the program. Finally, by selecting the model that minimises the value of Akaike's Information Criterion (the F–adjusted variant, AICC[3] (Hurvich and Tsai 1989)), a set of phase shift regression variables – with a specific autocorrelation structure of the model's residuals – was selected for each series. These are the *benchmark* models against which *alternative* models, adding Google Trends data, were compared.

B. Pre–whitening and cross–correlation

21. Rather than just assuming static relationships between the Google Trends and RSI data (that internet searches are contemporaneously related to retail activity), cross–correlations were estimated and plotted so that dynamic relationships could be identified (although models with "forced" static relationships were estimated for all series). However, before these estimates and plots could be produced, the RSI and Google Trends data had to be filtered in order to remove any autocorrelation within each of the time series. The presence of autocorrelation can result in spuriously large sample cross–correlations and hence misleading inferences and interpretations may be reached. The filtering process is often referred to as "pre–whitening". RegARIMA models were fitted to each of the time series and the residuals from these models were extracted. The model residuals were (approximately) free of autocorrelation and had properties that are similar to those of white noise – hence the term "pre–whitening".

22. Once the modelling procedure was completed, each of the pre–whitened RSI series could be cross–correlated with each of its corresponding pre–whitened Google Trends series. Significant positive cross–correlations at positive lags of Google Trends were of primary interest – such correlations suggest a positive relationship between the RSI and Google Trends data, where the latter is a leading indicator of the former. Significant negative correlations were ignored (as these were deemed to be infeasible for the purposes of this analysis), as were significant correlations at negative lags (as these suggest that the RSI data lead the Google Trends data – not a useful result in the context of this analysis). Sample cross–correlations were estimated up to a lag of three months; significant correlations beyond this were deemed to be infeasible for the purposes of this analysis.

C. Fitting the alternative models

23. For each of the 11 benchmark models, a number of competing alternative models were specified. For a given RSI series, the phase shift regressor set in the benchmark model was fixed in each of the alternative models. Then, for the given RSI series and a given alternative model, the relevant Google Trends series was added to the regressor set. Each of the Google Trends series has at least one alternative model associated with it, modelling its static relationship with the corresponding RSI series. Some Google Trends series are also associated with alternative models describing lagged relationships. For example, the sample cross–correlation plot of the Furniture and Lighting RSI series against the "garden" Google Trends search term series indicates significant correlations at lags two and three. Therefore, this Google Trends series is associated with four alternative models: one describing the static relationship between the series; one where the Google Trends series leads the RSI series by two months; one where the Google Trends series leads the RSI series by three months; and finally one where the Google Trends series leads

[3] AICC penalises the log likelihood by the number of parameters whilst accounting for series length.

the RSI series by both two and three months (so that two Google Trends explanatory variables are included in the model).

24. Some of the RSI series have more than once Google Trends search category associated with them. A multiple regression model was constructed for each of these series, including variables for the (fixed) set of phase shift effects and all Google Trends search categories associated with the RSI series (plus any automatically identified additive outliers and/or level shifts). Each of these models was then subjected to a process of backward deletion, whereby insignificant Google Trends terms were sequentially removed from the model, starting with the least significant, until all remaining Google Trends terms were found to be significant. The number of alternative models compared to each of the 11 benchmark models was therefore determined by: (i) the number of lagged relationships identified; and (ii) the number of Google Trends search categories utilised.

25. Note that, like the RSI dataset, all of the Google Trends series were log transformed and subjected to first order regular and seasonal differencing. The autocorrelation structure of the regression residuals, and any additive outliers and/or level shifts, were automatically re–identified and re–estimated for each series using the same methods within X–12–ARIMA as those used for the benchmark models.

D. Comparison metric

26. AICC was used to compare the goodness of fit of the alternative models with those of the benchmark models; this criterion penalises model complexity as well as rewarding goodness of fit. Modelling observed data will nearly always result in lost information and, ideally, this information loss should be minimal. A useful criterion for selecting a model from two or more candidate models might be to minimise information loss. Whilst it is not possible to calculate this quantity with certainty (because the true underlying data generating process is unknown), it is possible to estimate how much more (or less) information is lost by fitting one model when compared to another. Suppose we fit two models, *i* and *j*, and model *i* results in a smaller AICC value than model *j*; then the quantity $\exp([AICC_i - AICC_j] / 2)$ tells us how many more times as likely model *j* is to minimise information lost than model *i* (and will be less than zero in this case).

IV. Results

27. Table 4.1 summarises results for each of the 11 RSI time series. There appears to be some potential for using Google Trends data for quality assuring estimates of retail sales activity for some components of RSI. Furniture and Lighting is particularly noteworthy, where over 90 per cent of the fitted alternative models are more likely to minimise information lost than the benchmark model. However, the results are less promising for other components, such as Non–Specialised Food Stores and Non–Specialised Non–Food Stores. The benchmark models are more likely to minimise information lost than any of the fitted alternative models for these components of RSI, as well as for the aggregate All Retail Sales series. A large proportion of Google Trends terms are statistically significant for some components of RSI, but do not necessarily add value in terms of reducing information lost, relative to the benchmark models. For example, all six Google Trends terms are statistically significant for the Books, Newspapers and Stationary component, but only one of the six fitted alternative models leads to an AICC value which is lower than that of the benchmark model.

Table 4.1 Results for each of the RSI time series

Component of the Retail Sales Index (*number of alternative models fitted*)	**Percentage of alternative models with lower AICC than benchmark**	**Percentage of Google Trends terms significant at 5% level**
All Retail Sales (*8*)	0.0	37.5
Non–Specialised Food Stores (*6*)	0.0	0.0
Non–Specialised Non–Food Stores (*6*)	0.0	83.3
Textiles, Clothing and Footwear (*23*)	30.4	36.0
Furniture and Lighting (*31*)	90.3	78.8
Home Appliances (*7*)	14.3	0.0
Hardware, Paints and Glass (*6*)	50.0	100.0
Audio and Video Equipment (*44*)	43.2	51.0
Books, Newspapers and Stationery (*6*)	16.7	100.0
Computers & Telecommunications (*31*)	9.7	15.2
Non–Store Retailing (*7*)	42.9	42.9

28. The remainder of this section further explores the results for the Furniture and Lighting RSI series; according to AICC, Google Trends data appear to have most potential for quality assurance of this RSI component.

29. The Google Trends search categories that matched the Furniture and Lighting RSI series most closely were: Lighting; Home and Garden; Homemaking and Interior Décor; and Home Furnishing. Only six of 33 specifications returned an insignificant relationship between the Google Trends and RSI data at the 10 per cent level. 18 of the Google Trends regressors were significant at the 1 per cent level whilst 26 were significant at the 5 per cent level. Table 4.2 details the three alternative models with the lowest AICC values.

Table 4.2 Furniture and Lighting: three alternative models with the lowest AICC values

Google Trends Term	**Lag(s)**	**Google Trend Category**	**AICC**
lighting	0	Home Furnishings	412.47
curtains curtains curtains	0 & 1	Homemaking and Interior Decor	414.76
lights	0	Lighting	415.63
Benchmark			**432.29**

30. Inclusion of the Google Trends search term "lighting" (within the Home Furnishings category) minimised the AICC and resulted in a model which is over 20,000 times more likely to minimise information lost than the benchmark model. No significant phase shift effects were identified for the Furniture and Lighting series and the regression errors were modelled as an ARMA(1,0)(0,1) process, so the "best" fitted alternative model, according to AICC, can be expressed as:

$$(1-B)(1-B^{12})\ln y_t - \beta(1-B)(1-B^{12})\ln x_t = \frac{(1-\Theta B^{12})}{(1-\varphi B)}\varepsilon_t$$

31. The parameter estimate associated with the Google Trends term was found to be highly significant ($p<0.001$). The same model resulted in the lowest mean absolute percentage error (MAPE) of out–of–sample, one–step–ahead predictions over the final 12 months (July 2010 to June 2011) of the Furniture and Lighting series, as indicated by table 4.3 and figure 4.1. In total, 28 of the 31 fitted alternative models outperformed the benchmark model in terms of out–of–sample predictive performance. Hence, if Google Trends data had been used to predict the official Furniture and Lighting time series for quality assurance purposes for observations between July 2010 and June 2011, more accurate predictions would have been obtained than if only past observations and the arrangement of the survey calendar had been used.

Table 4.3 Furniture and Lighting: three alternative models with the lowest MAPE values

Google Trends Term	Lag(s)	Google Trend Category	MAPE
lighting	0	Home Furnishings	2.38
lighting	0	Lighting	2.49
Home Furnishings	0	N/A	2.51
Benchmark			**3.87**

Figure 4.1 Furniture and Lighting: predictions over the final 12 months of the series

V. Caveats and considerations

A. Changes over time

32. Many of the Google Trends search queries involved in the analysis are associated with specific companies or products. For many of these queries, the Google Trends time series have been shown to be correlated with the corresponding official RSI data. For example, for the Audio and Video Equipment and Recordings RSI series there are a number of alternative models that outperform the benchmark model and have significant parameter values associated with Google Trends terms. Empirically, there is an argument for considering such models for use in routine production activity in ONS. However, search queries associated with specific companies or products may be highly transitory; a product which is popular today may command far less interest one year from now, especially for electronics, computing, telecommunications and clothing products. Google Trends time series for such products may not be useful for prediction in the long run, as their relationship with the corresponding RSI series is not likely to be stable over time. In extreme cases, the observed Google Trends time series may start only very recently, or end in the very near future, or both (as search data are not provided if they fall beneath a certain threshold). Therefore, models would need to be frequently updated with new Google Trends terms; this is unlikely to be practical or sustainable.

B. Length of Google Trends time series

33. The fact that Google Trends data are only available from 2004 means that models incorporating Google Trends terms can only be constructed using a relatively short span of data. The benchmark models in this analysis were constructed using RSI data from 2004 onwards, for consistency with the alternative models. However, RSI data are available for periods pre–dating 2004 and so the benchmark models could be re–identified and re–estimated using longer time

series. This may lead to improved predictive performance, relative to the corresponding alternative models, and would provide a better representation of current ONS practice in general.

C. Assumptions

34. Only significant positive cross–correlations at positive lags were considered during the model building stage – such correlations suggest a positive relationship between the RSI and Google Trends data, where the latter is a leading indicator of the former. Significant negative correlations were ignored and sample cross–correlations were estimated up to a lag of three months. However, there is an argument for also considering significant cross–correlations at negative lags (suggesting a negative relationship between the RSI and Google Trends data) and lags greater than three months (suggesting a relatively long delay between search and purchase activities).

D. Explanatory variable selection

35. The final – but perhaps most import – consideration is variable selection. Google Trends provides a bank of millions of possible explanatory variables, constrained only by the letters and words that are entered into the Google search engine. Relatively poor performance of an alternative model against a benchmark model, for example in the case of Non–Specialised Food Stores, could be the result of including the "wrong" Google Trends term(s) in the alternative model, rather than Google Trends being a bad data source in general. However, given the huge array of search terms from which explanatory variables may be chosen, there is clearly a need for automation of the variable selection procedure; developing such a procedure will require careful consideration.

VI. Conclusions

36. This paper has reported the methods and findings of an investigation into the potential for using Google Trends data in ONS, with particular focus on RSI. The investigation has attempted to establish whether or not Google Trends data might be a useful source of information for quality assuring RSI outputs prior to publication, and adds to initial research, using the same data sources, conducted by Chamberlin (2010).

37. The results of the investigation are mixed. The series for which Google Trends data appear to have most potential for quality assurance of retail sales are those which are relatively volatile and/or contain turning points in their trend, and are therefore less easily forecast using only past observations and the arrangement of the survey calendar. Generally speaking, these series describe non–staple, "luxury" purchases by consumers, for goods whose consumption is relatively elastic to changes in prices and income. The best examples of such series include: Furniture and Lighting; Hardware, Paints and Glass; and Audio and Video Equipment and Recordings. These series may be contrasted with those that describe "necessity goods", for example Non–Specialised Food Stores, which exhibit stable trends and clear seasonal behaviour. Alternative models, which include Google Trends terms as predictors, appear to add little value to the benchmark models (whose predictions are based only on past values of the series and the arrangement of the survey calendar) for these latter series.

References

Box G E P and Jenkins G M (1976) *Time Series Analysis: Forecasting and Control*, Second edition, Holden Day: San Francisco.

Chamberlin G (2010) 'Googling the Present', *Economic and Labour Market Review* (Dec 2010), available at: http://www.statistics.gov.uk/cci/article.asp?ID=2621&Pos=5&ColRank=1&Rank=1

Choi H and Varian H (2009) *Predicting the Present with Google Trends*, Google Inc, available at:

http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/papers/initialclaimsUS.pdf

Findley D F, Monsell B, Otto M, Bell W and Pugh M (1988) *Toward X–12–ARIMA*, Research report number 88/27, Statistical Research Division, U.S. Census Bureau, available at: http://www.census.gov/srd/papers/pdf/rr88-27.pdf

Hurvich C M and Tsai C L (1989) 'Regression and Time Series Model Selection in Small Samples', *Biometrika* (Vol. 76, No. 2).

U.S. Census Bureau (2009) *X–12–ARIMA Reference Manual*, Version 0.3, available at: http://www.census.gov/srd/www/x12a