Filtering techniques for big data and big data based uncertainty indexes

GEORGE KAPETANIOS,
MASSIMILIANO MARCELLINO, FOTIS PAPAILIAS

2017 edition

STATISTICAL
WORKING PAPERS

eurostat

Filtering techniques for big data and big data based uncertainty indexes

GEORGE KAPETANIOS,
MASSIMILIANO MARCELLINO, FOTIS PAPAILAS

2017 edition

Manuscript completed in November 2017

Luxembourg: Publications Office of the European Union, 2017

PDF ISBN 978-92-79-74460-0 ISSN 2315-0807 doi: 10.2785/880943 KS-TC-17-007-EN-N

Abstract

This work is concerned with the analysis of outliers detection, signal extraction and decomposition techniques related to big data. In the first part, also with the use of a numerical example, we investigate how the presence of outliers in the big unstructured data might affect the aggregated time series. Any outliers must be removed prior to the aggregation and the resulting time series should be checked further for outliers in the lower frequency. In the second part, we explore the issue of seasonality, also continuing the numerical example. Seasonal patterns are not easily identified in the high frequency series but are evident in the aggregated time series. Finally, we construct uncertainty indexes based on Google Trends and compare them to the corresponding Reuters-based indexes, also checking for outliers and seasonal components.

Keywords: Big Data, Unstructured Data, Time Series Conversion, Data Features, Filtering, Decomposition, Signal Extraction.

Acknowledgement:
This work has been carried out by George Kapetanios*, Massimiliano Marcellino† and Fotis Papailias‡ for Eurostat under a contract with GOPA. The Eurostat project manager was Dario Buono**. Special thanks to Lisa Bosotti for taking care of the layout and to Sabine Bankwitz for the backstopping support.

*george.kapetanios@kcl.ac.uk

†massimiliano.marcellino@unibocconi.it

‡fotis.papailias@kcl.ac.uk

Table of content

List of tables

List of figures

1 Introduction

This paper focuses on irregularities in big data, such as outliers or seasonal patterns. This topic has attracted considerable interest in the literature, but typical studies focus on univariate time series at monthly or quarterly frequency, while with big data we have the need to jointly analyze a vast amount of time series, possibly at very high frequency. A quick literature review can be however useful, to cast the problem in context.

The time series literature has been concerned with the identification of temporal patterns at least since the beginning of the 1980s. Pierce, Grupe and Cleveland (1984) is one of the main references in which seasonalities are analysed in ARIMA models. Harvey, Koopman and Riani (1997) and Proietti (2000) further extend the above results adding trigonometric expressions and formulation. Pedregal and Young (2006) contribute with a model which estimates a cyclical component too. Temporal patterns in forecasting are discussed in De Livera, Hyndman and Snyder (2011).

In the non-parametric setup, the X11 family of methods of the U.S. Census Bureau and the Seasonal and Trend decomposition using Loess by Cleveland, Cleveland, McRae and Terpenning (1990) have been mostly applied in practice. We refer to Ladiray and Proietti (2017) for an excellent and exhaustive analysis of seasonal patterns, also in high frequency. Instead, we aim to provide further insights on how specific temporal patterns, such as outliers and seasonalities, in big data, may affect the conversion to time series and, more specifically, what the applied researcher should do in order to avoid mistakes and the derivation of false stylised facts based on the aggregated series.

In the first part of this report, we investigate how outliers in the unstructured data might affect the resulting aggregated time series. To illustrate the theoretical problems, we use a simulated series of 720,000 observations with 1,698 positive outliers to construct a daily time series of 500 observations. We illustrate the results when either keeping or removing the outliers. Then, we check the resulting time series for any outliers at the lower frequency.

In the second part of the report, we discuss how recurrent temporal patterns (seasonality) in big data, and in particular in high frequency, can affect the resulting aggregated time series. As an illustration, and continuing the previous example, we show that seasonal patterns are difficult to be identified in high frequency. However, their presence becomes more obvious in the converted aggregated time series. Using techniques based on a previous project([1]), we analyse the trend and seasonal components, and discuss the remainder “cleaned” component.

In the third and final part of the report, we evaluate empirically the presence of outliers and seasonal patterns in uncertainty indexes based on textual data. We consider the Reuters-based uncertainty index developed by the same authors, and construct a similar index based on Google Trends. Then, we check both the Google and Reuters indexes for outliers and seasonalities, and compare the original and “cleaned” versions.

The rest of the report is organised as follows. Section 2 focuses on outliers detection. Section 3

([1]) In particular, Sections 2.1, 3.1 and 3.2 bear a resemblance with Task 4 in “Nowcasting with Big Data”, 2016, Eurostat project undertaken by the same authors.

discusses the identification and removal of seasonal patterns. Section 4 presents the Google Trends-based uncertainty indexes and their comparison to the corresponding Reuters-based ones. Section 5 summarizes the main results and concludes.

2 Outliers

2.1 Definition

An interesting issue about outliers is the identification of their origin, which is then related to the decision of whether to remove them or not. For example, using credit card transactions data, single transactions that are close to or equal to the maximum credit card limit (e.g., £10.000 in the UK), are likely outliers from a statistical point of view, as they are far away from the standard values, though they are interesting from an economic point of view. Similarly, scanner price food data collected after a particularly long freezing spell can return abnormal values for fresh vegetables, which are outliers from a statistical point of view but interesting to test market reactions from an economic point of view. In this case, the decision of whether to remove or not (and hence model) the outliers is not obvious, while if they are due to coding or measurement errors it is clear that they should be removed. A similar issue holds for missing observations, which can be absent either for economic reasons, e.g. a certain financial variable was not introduced before a certain date, or for other reasons, e.g., the temporary Government shut-down in the USA that prevented data collection for a few weeks. Let us assume for simplicity that we do want to remove the outliers and/or fill-in the missing observations.

It must be noted that, in all cases, we focus on big data in numerical format or numerical features which have been extracted by textual big data.

Chen and Liu (1993) provide a methodology which deals with the detection of five types of outliers:

1. Additive Outlier (AO)
2. Innovation Outlier (IO)
3. Level Shift (LS)
4. Temporary Change (TC)
5. Seasonal Level Shift (SLS).

Below we briefly provide the basic definitions for these types of outliers. More information can also be found in, e.g., Chen and Liu (1993) and Hyndman and Athanasopoulos (2014).

Let us assume that an unprecedented event occurs at time τ. In the case of an AO, the observed time series is given by

$$y_t = z_t + sI(t)$$

where $I(\cdot)$ is an indicator function,

$$I(t) = \begin{cases} 0, \text{for } t \neq \tau \\ 1, \text{for } t = \tau \end{cases}$$

and s is the shock caused by the outlier which causes the y_t series to deviate from the "true" value,

z_t.

An IO is more complex and can be defined, say in the case of ARMA models, as

$$y_t = \mu_t + \frac{\theta(L)}{\emptyset(L)}\left(\varepsilon_t + sI(t)\right)$$

where $I(\cdot)$ is the indicator function, as before, and $\theta\ (L)$ and $\emptyset\ (L)$ denote the MA and AR polynomials respectively.

An LS outlier causes a permanent change in the level of the series and is given by

$$y_t = z_t + sI_D(t),$$

where now

$$I_D(t) = \begin{Bmatrix} 0, \text{for } t < \tau \\ \gamma, \text{for } t \geq \tau \end{Bmatrix},$$

for a given γ.

A TC outlier is defined as

$$y_t = z_t + sD(t),$$

where

$$D(t) = \frac{1}{1-\delta}I(t)$$

Is a damping function and 0 < δ < 1.

Finally, an SLS is defined as

$$y_t = z_t + sI_S(t),$$

where

$$I_S(t) = \begin{Bmatrix} \gamma, \text{for } t = \tau + d \\ 0, \text{otherwise} \end{Bmatrix}$$

and d is the seasonal parameter.

The general framework by Chen and Lui (1993) can be applied to ARIMA time series. In general, consider a time series which is subject to m outliers (of any of the above types) given by

$$y_t = S_t + \frac{\theta(L)}{\emptyset(L)}\varepsilon_t,$$

where S_t contains the outliers, $\theta\ (L)$ and $\emptyset\ (L)$ denote the MA and AR polynomials and ε_t denotes the innovation term. Then, the presence of outliers is tested by means of t-statistics applied on

$$\pi(L)y_t = \pi(L)S_t + \varepsilon_t$$

where $\pi(L) = \sum_{i\_=0}^{l} \pi_i L^i$ is a proxy for $\emptyset^{-1}(L)\theta(L)$. In particular, all the t-statistics for each type of outlier and for every time point are computed recursively. Then, an outlier is identified when the corresponding t-statistic (in absolute value) exceeds a threshold.

The package "tsoutliers" offers detection of the above types of outliers, applying the Chen and Liu (1993) methodology along with an automated procedure for outliers detection (see López-de-Lacalle (2015)). This is a publicly available package distributed under the GNU license. The outliers are replaced using optimal interpolation in terms of in-sample MSE. The same procedure could be applied to the case of missing observations, which can be treated as outliers. Outlier detection and treatment procedures are also typically available in software for seasonal adjustment.

A simpler rule of thumb for the detection of outliers (and missing observations) is to identify as such all values in a given sample which are w standard deviations above or below the mean or the median of the series, where w is typically in the range 4-6. The outliers (and missing observations) are then replaced by averages of the adjacent observations. This "quick and dirty" way to handle outliers often proves quite useful in regression analysis.

Finally, it is worth noting that outliers are usually removed for modelling purposes but they are reintroduced at the end of the procedure unless they can be considered as errors or as not justifiable from an economic point of view.

2.2 Outlier detection algorithm

Big data often requires to be translated into a time series format suitable for further use in macroeconomics nowcasting. Outliers can pose a potential threat to this conversion process, in the sense that they can somewhat bias the outcome and therefore the ensuing nowcasting exercise.

In this subsection, we summarise the steps which need to be taken to deal with potential outliers in data

1. Given an unstructured big data, decide the frequency of aggregation and, there- fore, the resulting time series. The method of aggregation must also be set.
2. When organising the unstructured data to be aggregated, detect possible outliers and remove or replace them.
3. Then, proceed with the aggregation.
4. Repeat the above steps for all periods in the desired frequency and obtain the converted time series which is created based on the cleaned raw data.
5. Finally, check the resulting time series for outliers in the low frequency.

The above algorithm suggests a "double" outlier detection: (i) first in the raw unstructured data, and then (ii) in the converted time series. For example, let us assume that we have a tick by tick dataset of financial data which needs to be converted to daily time series using linear aggregation. Then, for a particular day, *t*, we first collect all the raw data which occurs on that day, detect and remove any outliers, and then aggregate it. Repeating this procedure for t = 1, 2, ..., T times results to the structured daily time series, x_t. Then, x_t is checked again for outliers at the daily frequency.

Of course, another possibility would be to first aggregate the series without doing any check in high frequency, and then remove the outliers directly in the aggregated time series. The results in this case can be different, depending on the magnitude of the outliers. To illustrate this point, in the next subsection we provide a detailed example based on simulated data.

2.3 An example of outliers in unstructured data

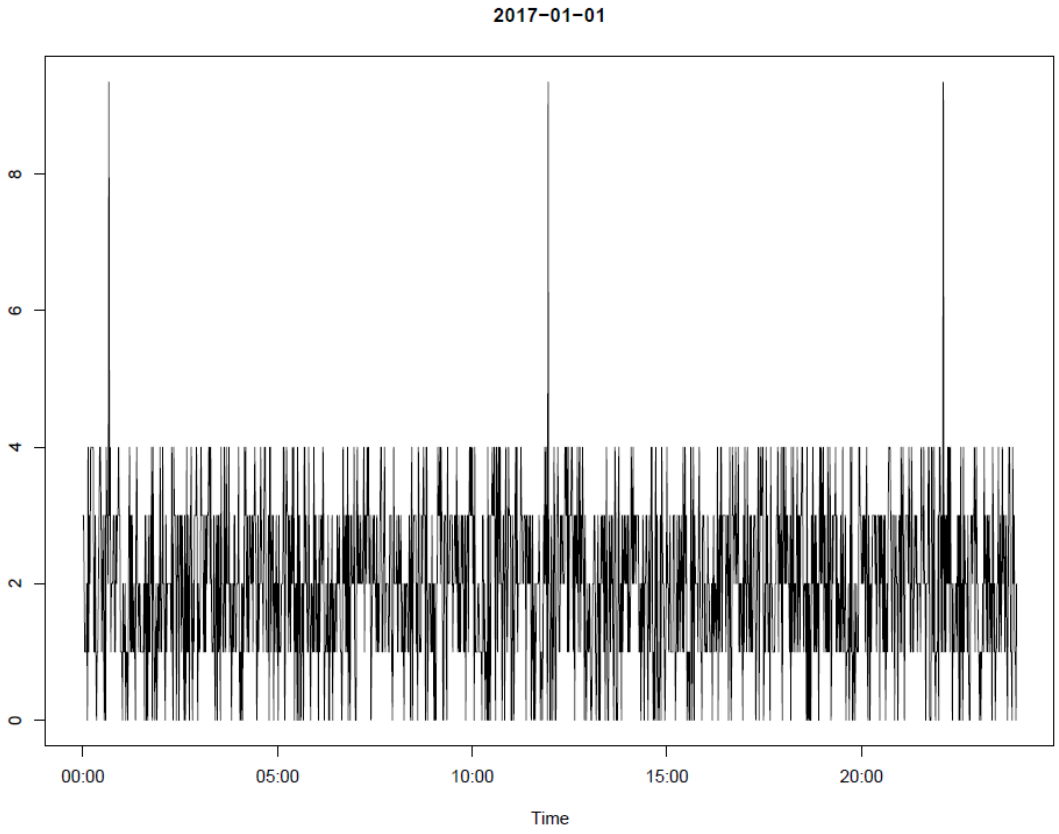
Analysis of various examples of unstructured big data conversion to time series has been already carried by the same authors, among others. However, there has not been a discussion regarding potential outliers, or other irregularities, which might exist in the raw data and pose threats to the resulting converted time series. Unfortunately, the lack of access to proprietary big data sources does not allow for a series-by-series comparison. Hence, in this subsection we provide a simulated example, using the "double" outliers algorithm introduced above and illustrating the difference outliers can make in the aggregation of the big data.

2.3.1 THE DATA

We generate a sample of 500 days with observations occurring every minute. This leads to 1,440 observations per day and N = 720,000 observations in total. In each minute, the variable can take the following integer values: {0, 1, 2, 3, 4}, drawn from a uniform distribution. In the case of categorical unstructured data, such as textual data, we can assume that these are values for a data feature, i.e., they correspond to the conversion of categorical to numerical data. For now, we assume that the researcher is not aware that the variable takes one of the above values, even if this would simplify the task to identify data errors.

Then, we impose a number of $2 \times [N^{0.5}]$ positive outliers. For the above case of 720,000 observations this leads to 1,698 outliers. The outliers are constructed as 6 standard deviations above the mean. Figure 1 displays the unstructured data during one day as a typical example of the data generating process. In the figure we can observe three outliers.

Figure 1: An intraday example of the generated sample. Three positive outliers are observed



Source: Based on author's calculations

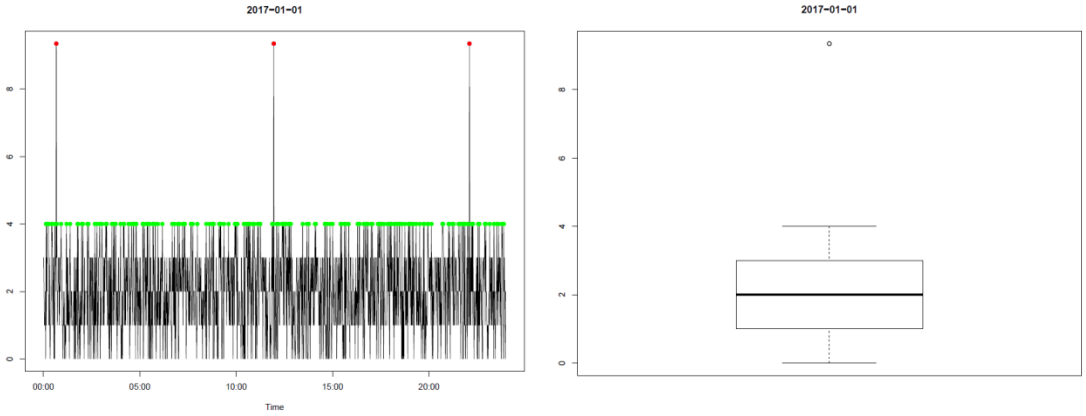
2.3.2. CONVERSION TO TIME SERIES

In this subsection we illustrate the effects that outliers might have in the conversion of unstructured data to time series. At first, the researcher must have a good understanding of the structure of the underlying data. For example, in the above generated sample we know that the observations of the unstructured data will be one value between 0 and 4. Therefore, any other value which is below 0 or above 4 should be considered as an outlier or a data error.

In order to automatically clean our data, we consider two methods to spot outliers: (i) assuming normality, we calculate the Z-scores of all observations and consider an outlier any value outside the ±1.96 and ±2.58 bounds which correspond to 5% and 1% levels([2]), respectively, and (ii) any value outside the outer fences of a standard boxplot. Figure 2 illustrates the two detection methods with the resulting detected outliers. It turns out that both 1% Z-score and the boxplot yield reliable results.

([2]) The outlier definition of any value 4 or 6 standard deviations away from the mean is also part of this definition.

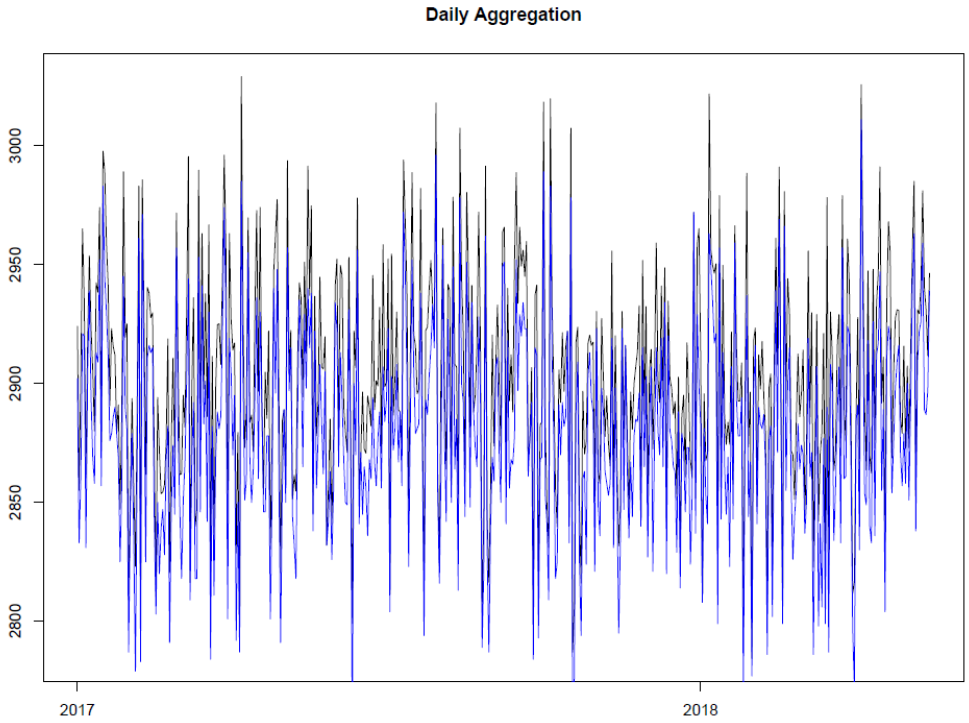
Figure 2: Examples of outliers detection. Green colour indicates 5% level, red colour indicates 1% level



Source: Based on author's calculations

The next step is to examine whether outlier detection in the unstructured data affects the aggregation and results in improved time series conversion. In this example, we linearly aggregate the unstructured data on a daily basis, with or without prior outlier detection and removal, resulting in two daily time series of 500 observations each. The outlier detection algorithm is run on an intraday basis replacing the outliers of each day with that day's median. In Figure 3 we plot the two series that turn out to be rather different, highlighting the importance of a careful pre-treatment of the outliers.

Figure 3: Effects caused by outliers. Black line indicates the original series which includes the outliers and the blue line indicates the cleaned series



Source: Based on author's calculations

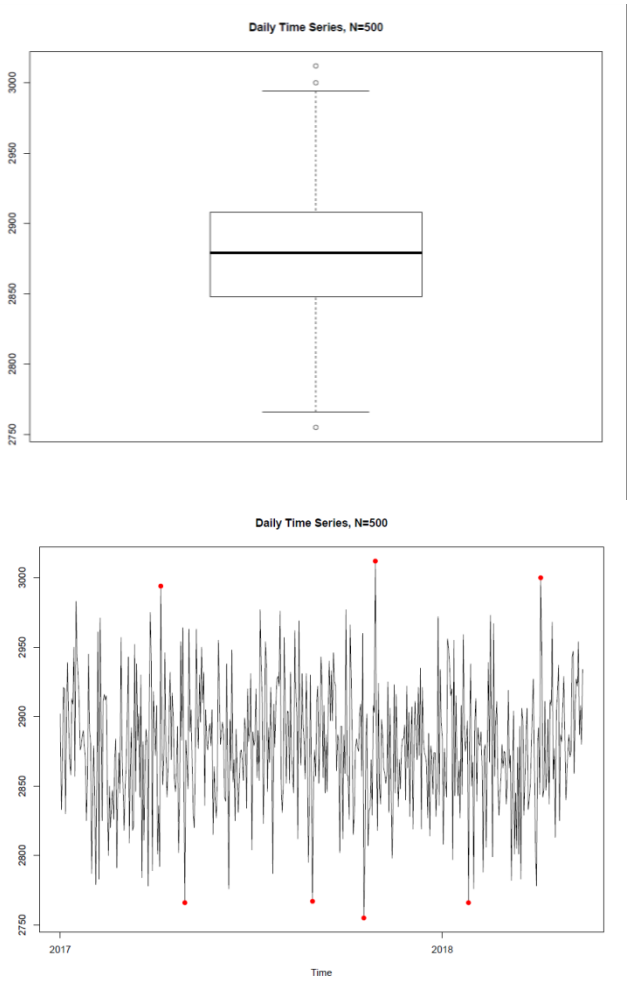
2.3.3 OUTLIERS IN THE CONVERTED TIME SERIES

The first step is to assess whether the converted daily time series still suffers from outliers. A simple

boxplot suggests a few remaining outliers. These are also identified using the Z-Scores at 1% level as before.

In theory, there should be no outliers in the daily series. However, some of them can emerge from the aggregation process, combined with the random features of the aggregated variables. Moreover, if we define outliers as observations 4 or 6 times away from the mean, then the resulting daily time series no longer presents outliers. Given the design of our DGP and the definition of outliers, any method can spot them. Therefore, once the outliers are removed the output is similar to that in Figure 3.

Figure 4: Outliers in the converted series



Source: Based on author's calculations

3 Seasonal patterns and signal extraction

3.1 Definition

Apart from outliers, repeating (predictable) patterns are also usually observed at seasonal frequencies, such as monthly or quarterly. These patterns are called seasonal variations or simply seasonalities. Seasonalities in a time series might be caused by a variety of factors, such as weather, fixed holidays, trading times, etc. In addition, calendar effects can also matter, such as those generated by moving holidays. As seasonal patterns and calendar effects can bias or hide the relationship between the big data indicators and the target variable of the forecasting exercise, it is preferable to remove them and work with adjusted indicators. On the other hand, care has to be exerted in the seasonal adjustment procedure, as spurious cycles can otherwise be generated, see for example Canova and Ghysels (1994), Stock and Watson (2002b). Moreover, common seasonal adjustment procedures are developed for monthly or quarterly time series, while with big data we may need to handle higher frequency time series and/or more general recurrent patterns.

There are various ways to try to take out the seasonal component from a time series, all of which can be also applied to indicators resulting from summaries of, possibly unstructured, big data. As mentioned, we focus on methods that are computationally simple and robust enough to be applied to a vast number of time series, possibly on a recursive manner. In this context, a leading possibility is to use a so-called structural approach, where each time series consists of signal plus noise and the signal is further decomposed into trend, seasonal and cyclical components. Seasonal adjustment aims at removing the seasonal component of a time series. The seasonally adjusted variables can then be used in the analysis of non-seasonal trends, over durations longer than the seasonal period.

An appropriate method for seasonal adjustment is chosen on the basis of a specific view about the decomposition of the time series, Y , into the “trend (T)”, “cyclical (C)”, “seasonal (S)” and, possibly, “irregular (I)” components([3]), and the interaction among them. In particular, typically the components are assumed to be independent and they act either additively or multiplicatively. In an additive time series model, the seasonal component could be extracted as

$$S = Y - (T + C + I),$$

based on specific parametric models for T , C and I.

In a multiplicative time series model, the seasonal component is instead expressed in terms of ratio and percentage as

$$\text{Seasonal Effect} = (T \times S \times C \times I)/(T \times C \times I) \times 100 = Y/(T \times C \times I) \times 100.$$

([3]) Irregular component includes patterns in the data which are not observed in a seasonal or cyclic manner but in a rather irregular form, such as outliers

The de-seasonalised time series data $(Y - S)$ will have only trend, cyclical and irregular components given by:

$$\text{Additive}: Y - S = T + C + I,$$

$$\text{Multiplicative}: Y/S \times 100 = (T \times C \times I) \times 100.$$

Empirically, one easy way to deal with seasonalities is to use the STL filtering approach based on local linear regression, *loess* method ("Seasonal and Trend decomposition using Loess"); see Cleveland, Cleveland, and Terpenning (1990). The R software has a built-in function called "stl" which performs the series decomposition.

Loess ("locally-weighted scatterplot smoothing") uses local regression to remove "jaggedness" from data.

1. At first a window length is specified. The larger the window, the more smooth the output result.
2. Then, a regression line is fitted to the observations that fall within the window. The points closest to the centre of the window are being weighted to have the greatest effect on the calculation of the regression line.
3. The weighting is reduced on those points within the window that are furthest from the regression line. The regression is re-run and weights are again recalculated.
4. We thereby obtain a point on the loess curve. This is the point on the regression line at the centre of the window.
5. The loess curve is calculated by moving the window across the data. Each point on the resulting loess curve is the intersection of a regression line and a vertical line at the centre of such a window.

According to Hyndman and Athanasopoulos (2014): "*STL can handle any type of seasonality, not only monthly and quarterly data. The seasonal component is allowed to change over time, and the rate of change can be controlled by the user. The smoothness of the trend-cycle can also be controlled by the user. It can be robust to outliers (i.e. the user can specify a robust decomposition). So occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component*."

One disadvantage of STL is that it does not automatically handle trading day or calendar variation (which on the other hand is handled by the X13-ARIMA-SEATS or the X12), and it only provides facilities for additive decompositions. However, it is possible to obtain a multiplicative decomposition by first taking logs of the data, and then back-transforming the components[4].

The decomposition of a series into unobserved components can also be handled in a general state space formulation, which can be estimated via the Kalman filter. This class of models is reviewed in detail in Harvey (1989). Typically, this is an efficient procedure but it can be very computationally demanded when applied in a big data context[5].

3.2 X13-ARIMA-SEATS

An alternative approach to seasonal adjustment is based on ARIMA modelling and it is also capable of handling outliers. Its implementation is rather simple, the computational costs are limited and the results rather robust, which makes the procedure suited for application to vast datasets, also in an official statistics context. The method is implemented in the TRAMO-SEATS software, which is also

([4])I t is important to note that the removal of intra-monthly periodicity does not necessarily remove monthly or lower frequency seasonality. Further, seasonal adjustment is needed only when the target variable for the nowcasting exercise is also seasonally adjusted, otherwise seasonality could be directly handled in the regression model that links the target and the explanatory variables.

([5]) Another possibility is to model seasonal effects using sine and cosine functions, see, e.g., Ng (2016) for an application in a big data context.

included in the JDemetra+ statistical software package for seasonal adjustment used at Eurostat. JDemetra+ also includes the X-12-ARIMA software, with documentation provided at http://ec.europa.eu/eurostat/sa-elearning/.

An updated version, known as X13-ARIMA-SEATS and used by the U.S. Census Bureau, could be also useful, and we now briefly describe it.

According to the official documentation, available at https://www.census.gov/srd/www/x13as/, X13ARIMA-SEATS includes the following features:

- *Extensive time series modelling and model selection capabilities for linear regression models with ARIMA errors*
- *The capability to generate ARIMA model-based seasonal adjustment using a version of the SEATS procedure originally developed by V. Gómez and A. Maravall at the Bank of Spain as well as nonparametric adjustments from the X-11 procedure*
- *Diagnostics of the quality and stability of the adjustments achieved under the options selected*
- *The ability to efficiently process many series at once.*

X13-ARIMA-SEATS therefore combines the X12-ARIMA and TRAMO/SEATS methodologies. We briefly provide some comments on them.

Following Maravall (2008a), TRAMO stands for Time Series Regression with ARIMA noise, missing observations and outliers. The TRAMO methodology considers the following general model:

$$z_t = \beta y_t + x_t,$$

where $y_t$is the deterministic component and $x_t$is an ARIMA process (stationary or nonstationary). An example for the deterministic component could be $y_t = a + bt$, while x_t is described in the usual form:

$$\emptyset(L)\delta(L)xt = \theta(L)at + c,$$

where $\emptyset(L)$ is the stationary AR polynomial, $\delta(L)$ is the nonstationary AR polynomial and $\theta(L)$ is the MA polynomial, c is a constant and L denotes the lag operator. For seasonal time series, the polynomials have a multiplicative structure, e.g., $\delta(L) = (1-L)^d\ (1 - L^s)^{ds}$.

The regression variable(s) can be user-defined or TRAMO-generated or both. The TRAMO-generated variables are standard and include: trading days, Easter, intervention variables, outliers etc. TRAMO then:

1. Performs an exact ML estimation of the regression-ARIMA model
2. Runs diagnostic checking
3. Detects AO, TC, LS and IO outliers
4. Performs an optimal interpolation of missing observations
5. Computes optimal Forecasts (in the sense of min. MSE)
6. Finally, it automatically identifies the best model and returns the “corrected” series.

The SEATS part of the program uses TRAMO as a pre-adjustment step which extracts the stochastic part (linearised series - ARIMA output) and the deterministic part (regression effects). The SEATS

methodology is then applied on the stochastic part of the series where the trend, the seasonal, the transitory and irregular components are extracted. Filters used by SEATS include the Wiener-Kolmogorov filter and the two-sided filter. The final output is the seasonally adjusted series; see Maravall (2008b).

The website http://www.seasonal.website/ provides a used friendly interface for the X13-ARIMA-SEATS approach. All the calculations are based on the R software, which runs in the background.

Empirically, it would be convenient to incorporate the seasonal adjustment and outlier treatment into the same software that transforms the unstructured big data into time series, and then uses them for nowcasting or forecasting the target variables of interest. For this, it could be appropriate to use the free software R, which has a "seasonal" package which offers full access to most options and outputs of X-13, including X-11 and SEATS, automatic ARIMA model search, outlier detection and support for user defined holiday variables, such as Chinese New Year or Indian Diwali. The package is publicly available under the GNU license and could be directly used by researchers in official statistical offices.

To conclude, we should mention that, unfortunately, X-13 and Tramo-Seats are only implemented for monthly or quarterly time series, but not for higher frequencies, such as weekly or daily. The use of weekly or daily data adds some additional complications, such as the distribution of the weeks across the months or the presence of weekends and moving festivals. These can be handled by proper calendar effects, and then either the structural approach or the ARIMA modelling can be applied to the higher frequency series to remove any remaining seasonal patterns and outliers. Similar suggestions have been made by Ladiray and Proietti (2017), to whom we refer for additional details.

3.3 Signal extraction

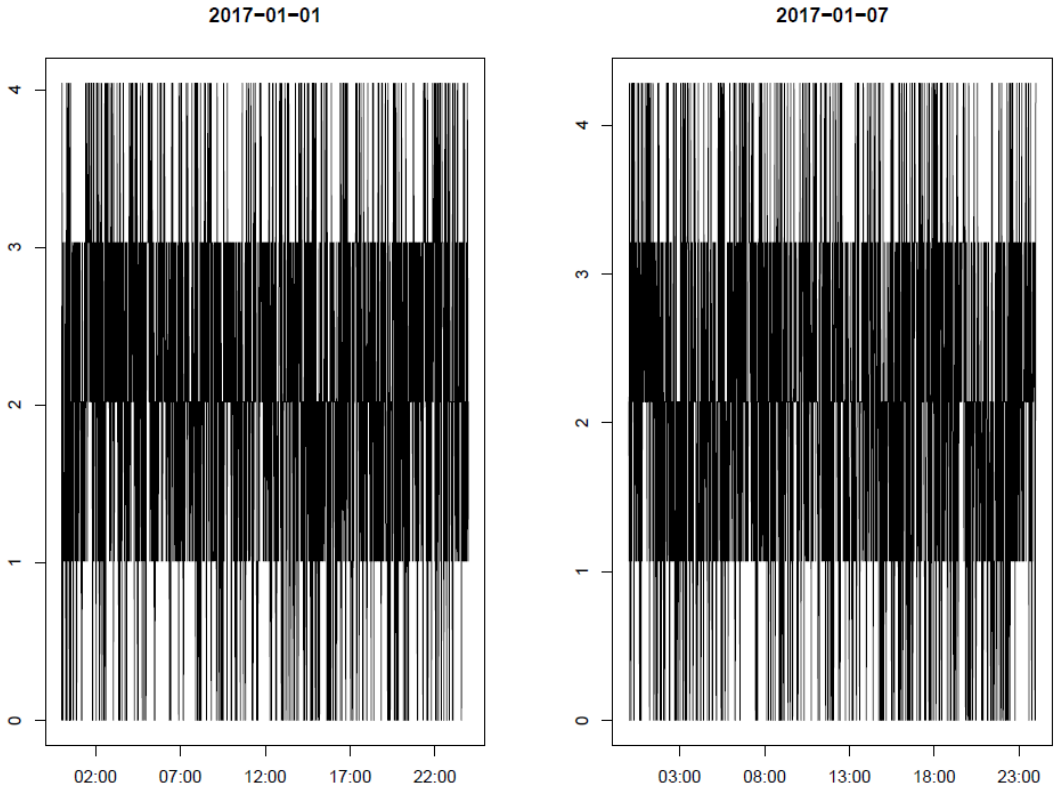
The seasonal effect is usually identified in the converted time series rather than in the original unstructured big data. To provide an illustration, we use again the example of Section 2.3, now imposing a seasonal effect in the daily time series, defined by the day of the week, with Sunday taking the lowest value and Saturday the highest. In particular, the big data, X generated in Section 2.3 is now transformed as $Y = X\,(1 + D/100)$, where D is a dummy variable indicating the day of the week taking values $D = 1, 2, \ldots, 7$, with 1 being a Sunday and 7 being a Saturday.

Figure 5 displays the unstructured data for two typical days: (i) 2017-01-01, which is a Sunday, and (ii) 2017-01-07, which is a Saturday. The data is generated as in Section 2.([6])

Apart from the slight diff in the maximum value, we cannot distinguish any seasonal pattern in high frequency. The seasonal pattern starts appearing in Figure 6, which displays all the raw data for the first two weeks. However, increasing the time frame results in incomprehensible figures.

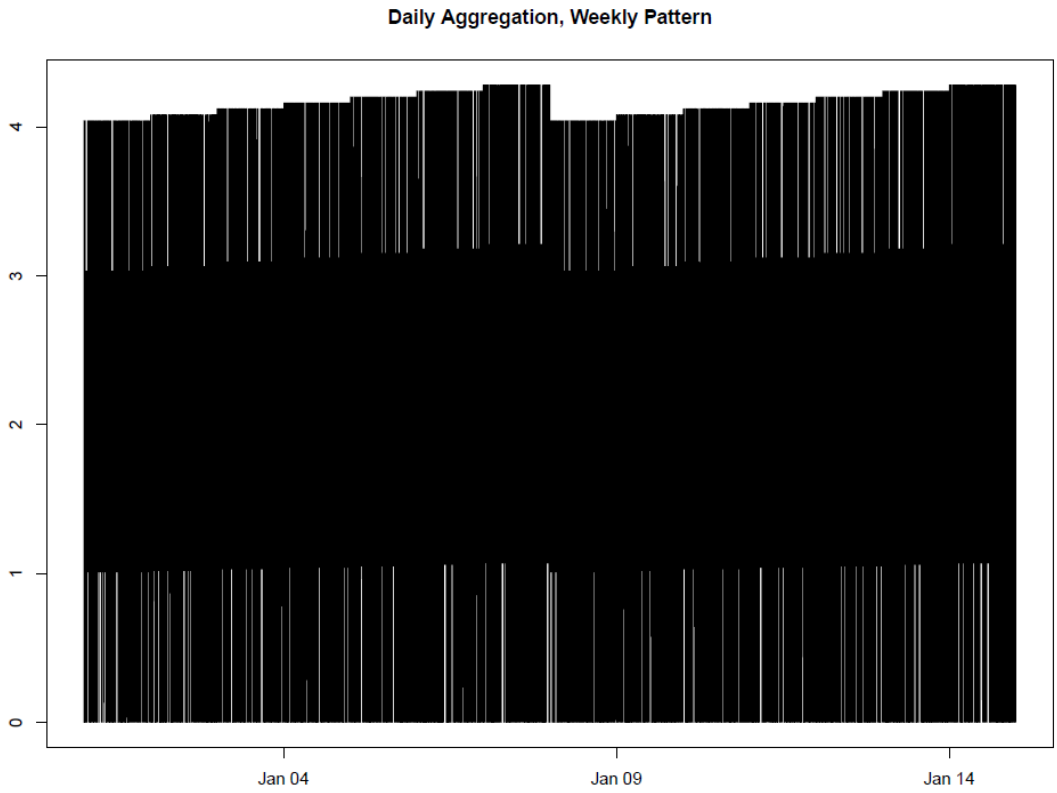
([6]) In this case, we impose larger values to the observations in order to create a seasonal pattern.

Figure 5: Unstructured data for two typical days. Seasonal effects exist but are not obvious in high frequency



Source: Based on author's calculations

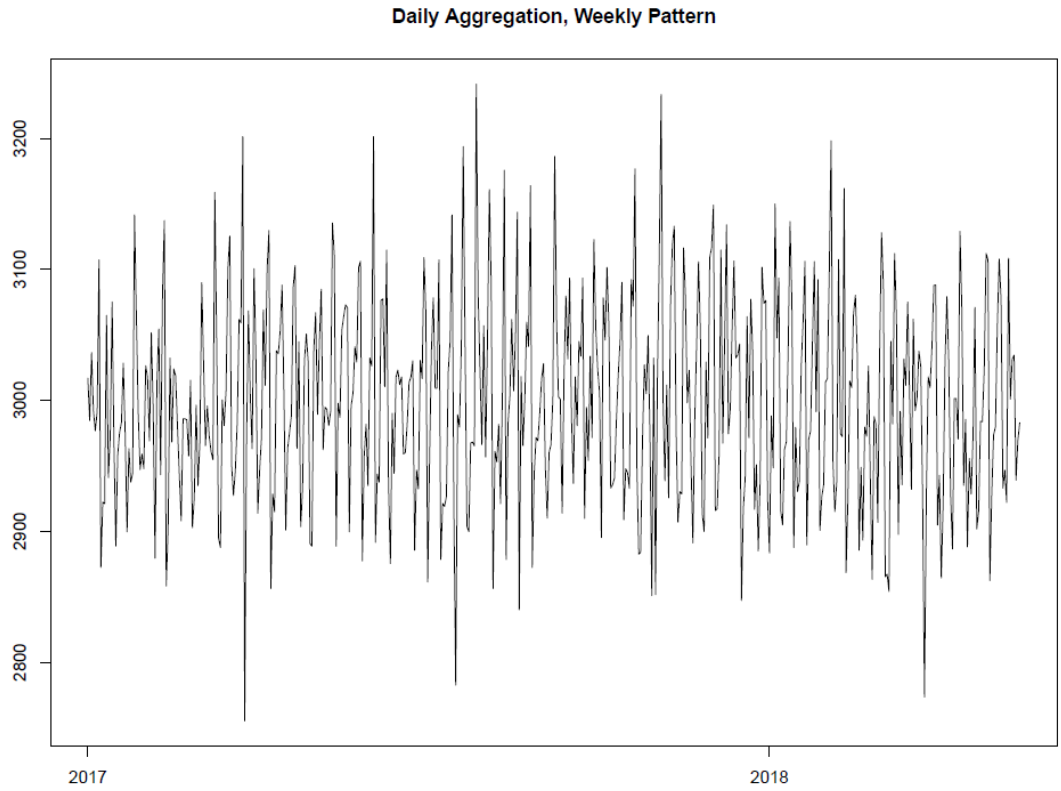
Figure 6: Unstructured data for the first 14 days. The weekly pattern becomes obvious looking at the maximum values



Source: Based on author's calculations

Aggregating at a daily frequency, as we did in Section 2.1, now leads to the series in Figure 7. This series can now be checked for outliers and seasonal patterns.

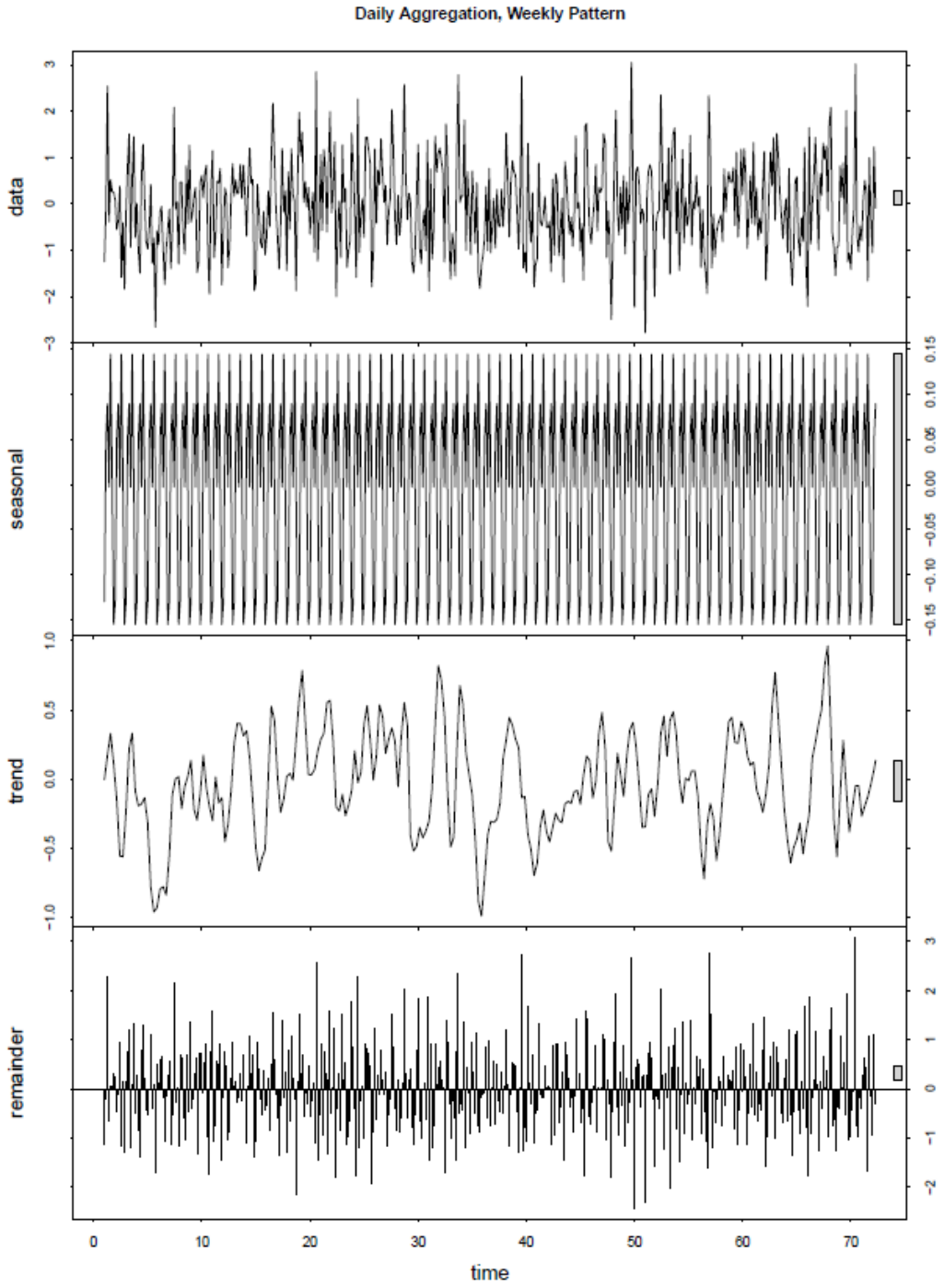
Figure 7: Converting the simulated unstructured data to daily time series. A seasonal pattern now seems more obvious



Source: Based on author's calculations

Given that the X13-ARIMA-SEATS method cannot treat daily time series, we use the STL decomposition method in our daily example with 7 days periodicity. Usually, the periodicity is known given that the researcher is aggregating at a desired frequency. However, one can also look at the spectral density to identify the most likely seasonal periodicity. Applying the STL method in this example leads to the (satisfactory) decomposition in Figure 8.

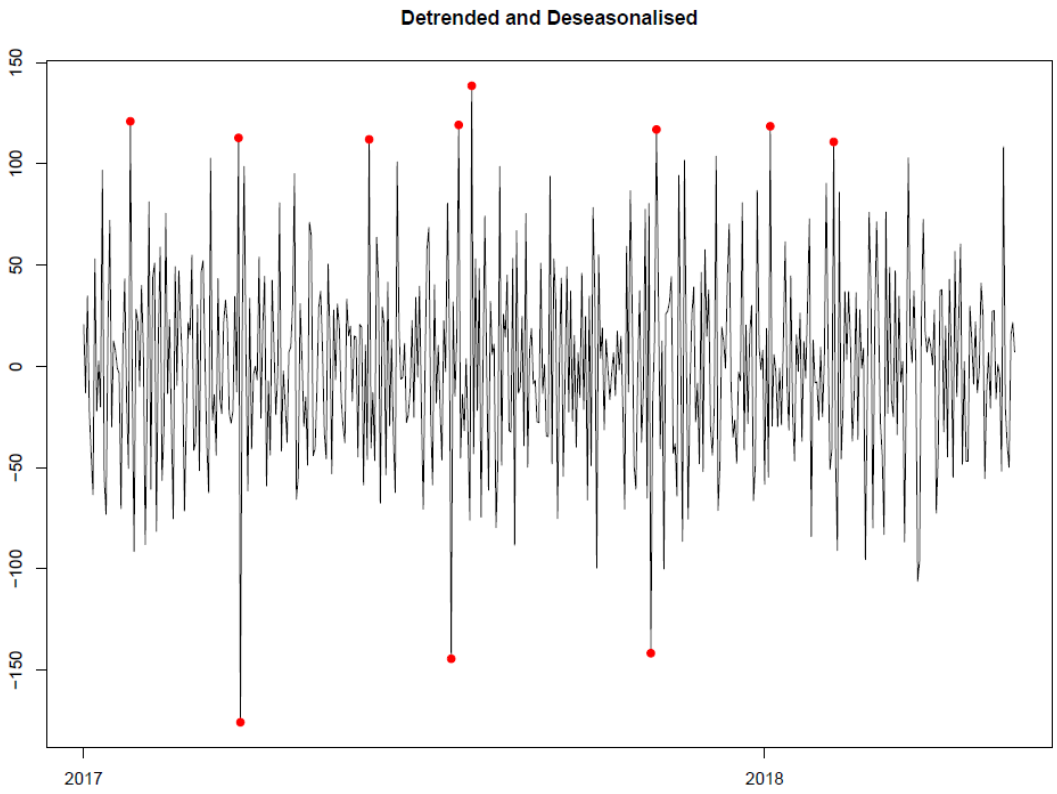
Figure 8: Decomposing the daily aggregated time series using STL



Source: Based on author's calculations

Finally, after removing the trend and seasonal components, we obtain the series in Figure 9, which can be further checked for outliers.

Figure 9: Cleaned series with some remaining outliers to be removed



Source: Based on author's calculations

4 From signal extraction to uncertainty indexes

In previous research, we have analysed the derivation of structured time series data from a big dataset of possibly unstructured data of various types. Particular attention has been paid to textual data obtained from web scraping, later used as an input for textual analysis software, whose output is one or more structured time series.

As an example, we focused on the extraction of information from news articles to construct measures of the extent of the uncertainty in the economy, which is quite relevant as uncertainty is an important determinant of the decisions of economic agents. For example, with higher uncertainty consumers and firms typically postpone their consumption and investment plans. Empirically, we downloaded a big set of about 3 million news articles from Reuters, searched each of them for a specific keyword (typically, "uncertainty" or a synonym), and finally aggregated the number of appearances at a daily or weekly frequency. This procedure resulted into a big data based uncertain index, that we labelled Reuters uncertainty index.

In this section, we first compare the Reuters uncertainty index to other indexes as well as volatility estimates. Then, we construct the corresponding Google Trends Uncertainty Indexes using similar keywords. Finally, we clean the resulting indexes by removing outliers or seasonal components, in order to make them ready for use as coincident or leading indicators in the nowcasting empirical analyses that will be conducted in the subsequent tasks.

4.1 Reuters uncertainty indexes

Table 1 summarises the dictionaries of keywords used in the construction of each Reuters Uncertainty Index (RUI)([7])

Table 1: Dictionaries of keywords by index

Index	At least one of	AND at least one of
Uncertainty	uncertain, uncertainty, uncertainties	
Risk	risk, risks, risky	
Italy, uncertainty	uncertain, uncertainty, uncertainties	Italy, Italian, Italians
Germany, uncertainty	uncertain, uncertainty, uncertainties	Germany, German, Germans
France, uncertainty	uncertain, uncertainty, uncertainties	France, French
UK, uncertainty	uncertain, uncertainty, uncertainties	UK, Britain, British, United Kingdom, Briton

Source: Based on author's calculations

In all subsequent comparisons, we focus on the general uncertainty index and risk index.

([7]) For more details see "Review of methods for feature extraction of big data sources to usable time-series for econometric modelling", Eurostat Working Paper.

4.1.1 ECONOMIC POLICY UNCERTAINTY INDEX

The Economic Policy Uncertainty Index (EPU) measures policy-related economic uncertainty based on newspaper coverage frequency by reporting the count of articles that contain words pertaining to uncertainty, the economy, and policy.([8]) Although monthly indices are provided for several countries, daily article counts are only available for the United States and the United Kingdom.

Besides its innovative methodology that we have replicated and extended in our own analysis, there is extensive evidence that the EPU provides an accurate mea- sure of economic uncertainty. First, the EPU index correlates strongly with other measures of economic uncertainty (e.g. implied stock market volatility), as well as other measures of policy uncertainty. Moreover, using fi rm—level data, Baker, Bloom and Davis (2016) find that policy uncertainty is associated with greater stock price volatility and reduced investment and employment in policy—sensitive sectors. At the macro level, they also find that innovations in policy uncertainty foreshadow declines in investment, output, and employment in the United States and in other 12 major economies.

The EPU index is made available by major commercial data providers, including Bloomberg, FRED, Haver, and Reuters. In our analysis, we use monthly data from the Global EPU index.

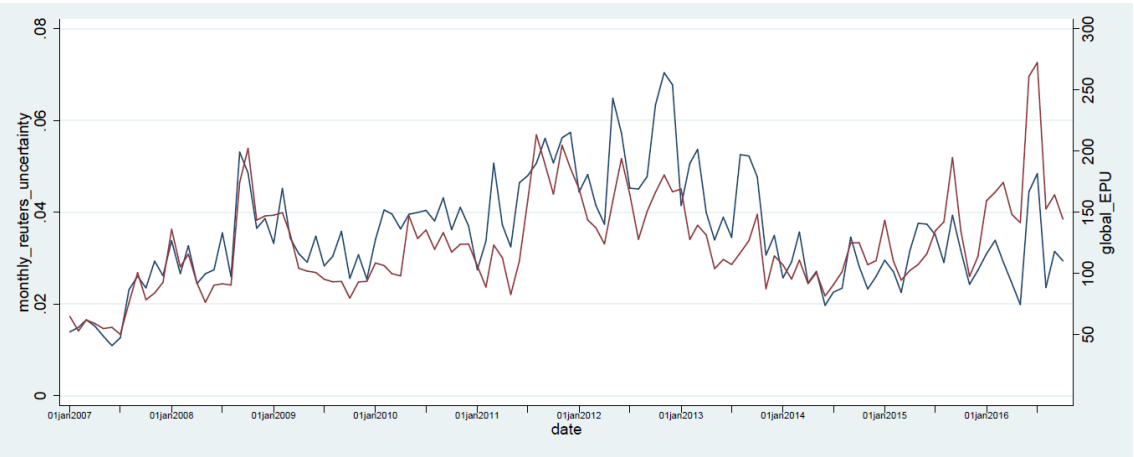
4.1.2 VIX

VIX is the ticker for the Chicago Board Options Exchange (CBOE) Volatility Index, a measure of market expectations of near-term volatility conveyed by S&P 500 Index (SPX) option prices.([9]) In fact, options prices reflect the market's perception of risk. When risk is perceived to be high, investors are willing to pay more for options than when perceived market risk is low. Moreover, options also reflect market expectations about stock prices variability: when the market expects large changes in stock prices, options prices tend to rise. As EPU, VIX is also made available by major commercial data providers. In our analysis, we use monthly aggregations from the daily VIX index.

4.1.3 A COMPARISON OF RUI, EPU AND VIX

Figure 10 shows a line plot of the EPU Global index and our index of uncertainty obtained from Reuters News article frequencies. Although the scales are different, co- movement is substantial, especially until 2012. The correlation coefficient computed on the entire period is 0.69.

Figure 10: Line Plot of EPU Global index (red) and Reuters Uncertainty index (blue), monthly data

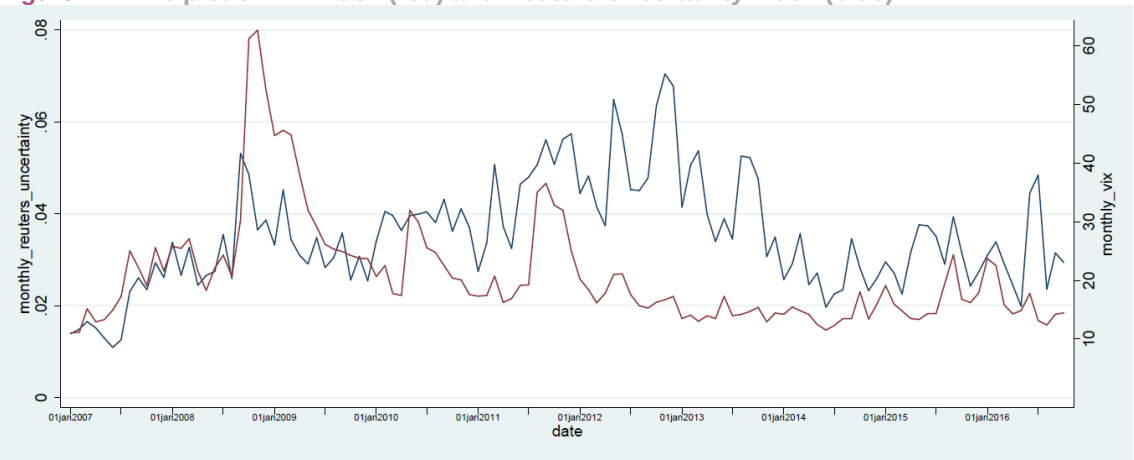


([8]) See Baker, Bloom, Davis (2016) for additional information.
([9]) The index was first proposed by Brenner and Galai (1987) and it was first published by CBOE in 1993. The present version of the VIX, updated in real-time, is the result of a 2003 methodology update.

Source: Based on author's calculations

Figure 11 shows a similar comparison between VIX and our uncertainty index. The two series commove strongly until the first quarter of 2012 when S&P volatility drops and remains subdued until the second half of 2015. The overall correlation coefficient is 0.21, and increases to 0.44 when restricting the sample to 2007-2012.

Figure 11: Line plot of VIX index (red) and Reuters Uncertainty index (blue)



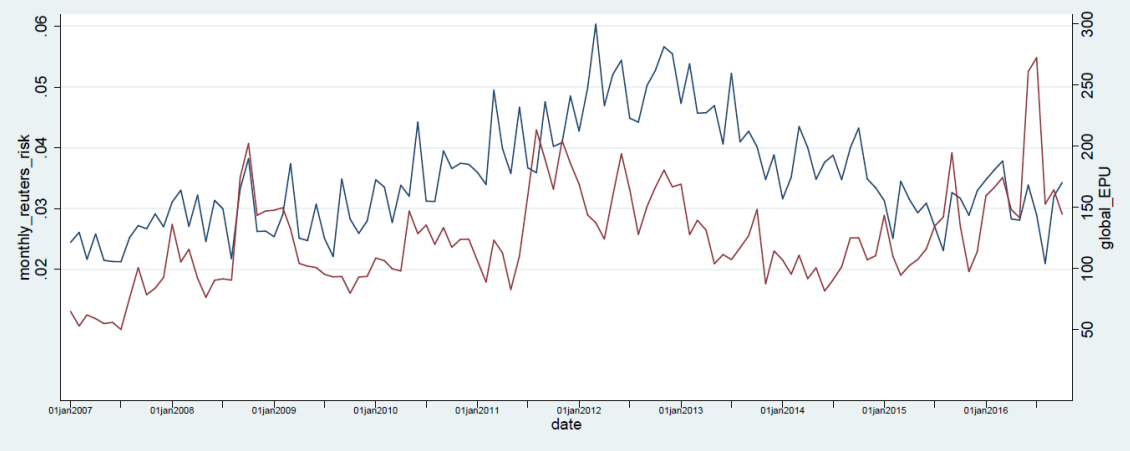
Source: Based on author's calculations

4.1.4 RISK INDEX

In economics, there is a clear distinction between uncertainty and risk. Knight (1921) discusses that uncertainty occurs when the likelihood of future events is indefinite or incalculable. In contrast, risk is present when future events occur with measurable probability. Therefore, we introduce risk indexes to complement the uncertainty indexes mentioned previously in order to take advantage of extra information which might be hidden under "risk".

The risk index computed from Reuters news displays a clear increasing trend until the height of the European debt crisis in 2012 and a subsequent reduction in the last three years. Correlation with the EPU Global index is moderate (correlation coefficient is 0.41) and peaks in the two series do not always coincide. For example, the Reuters risk index does not increase noticeably after the Brexit referendum; on the contrary, the EPU reaches its highest level since 2012.

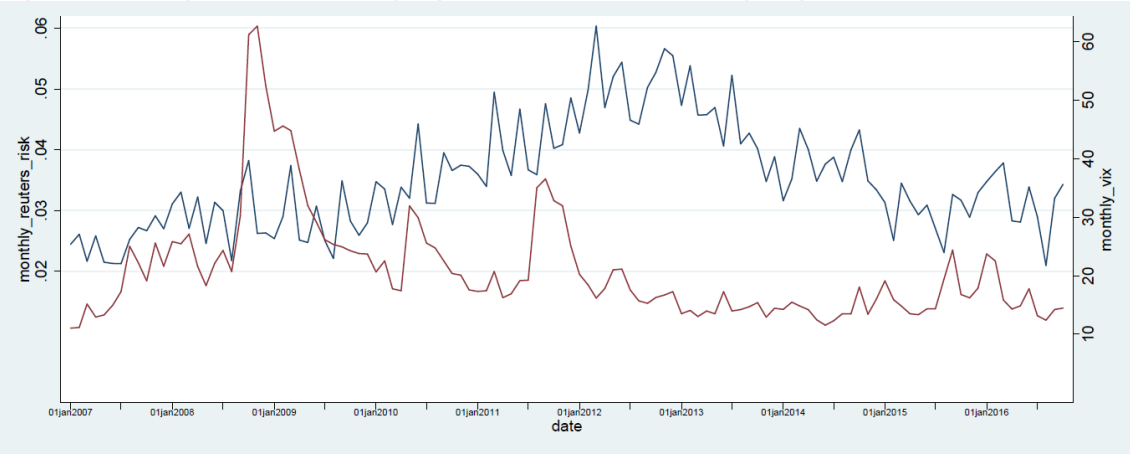
Figure 12: Line Plot of EPU Global index (red) and Reuters Risk index (blue), monthly data



Source: Based on author's calculations

Correlation between the Reuters risk index and VIX is slightly negative (-0.15), as the two series diverge after 2009: economic recovery in the US contributes to low volatility in the S&P 500, while the Reuters index captures the perjuring effects of the European debt crisis.

Figure 13: Line plot of VIX index (red) and Reuters Risk index (blue)



Source: Based on author's calculations

4.2 Google uncertainty indexes

Having constructed the Reuters Uncertainty Indexes, which compare well with other uncertainty estimates, it is now interesting to investigate the use of Google trends to derive alternative internet based uncertainty indicators.

For the general uncertainty and risk indexes, we consider four keywords, given that the searches are directed worldwide and the language barrier might jeopardise the result, and two keywords for the country-specific indexes using the domestic languages. In particular, we include the following Google Trends:

- Uncertainty: for the general Google Uncertainty Index we use four keywords across Worldwide web and news searches. These keywords are: "uncertainty",

- "economic uncertainty", "financial uncertainty" and "policy uncertainty".
- Risk: for the general Google Risk Index we use four keywords across Worldwide web and news searches. These keywords are: "risk", "financial risk", "political risk", "policy risk".
- France, Uncertainty: for the French Google Uncertainty Index we use the key-words "incertitude" and "risque" across web and news searches in the region of France.
- Germany, Uncertainty: for the German Google Uncertainty Index we use the keywords "unsicherheit" and "risiko" across web and news searches in the region of Germany.
- Italy, Uncertainty: for the Italian Google Uncertainty Index we use the key- words "incertezza" and "rischio" across web and news searches in the region of Italy.
- UK, Uncertainty: for the UK Google Uncertainty Index we use the keywords "uncertainty" and "risk" across web and news searches in the region of the UK.

For the separate countries we use both "uncertainty" and "risk" keywords in order to cover more user profiles and obtain a more robust result. Then, to construct each index we take the equally weighted average of the respective Google Trends. We also try to use data-dependent weights based on the spectral entropy of each series. Lower spectral entropy values indicate that the series is more predictable. However, the resulting uncertainty indexes do not change significantly and we omit the results from presentation.

In the next sections we compare each Google uncertainty index with the corresponding Reuters indexes. We use the General Uncertainty Index as our main example to briefly illustrate how the index is constructed, the presence of outliers, the trend and seasonal components and the final outliers-free, de-trended and de- trended seasonalised series. To save space, for all subsequent cases we present only the final indexes before and after cleaning. However, all figures for these cases are supplied in a separate Appendix.

4.2.1 GENERAL UNCERTAINTY INDEX

To construct the General Uncertainty Index based on Google Trends we first download the respective web-searched and news-searched trend series which are illustrated in Figure 14; each panel of the figure corresponds to a particular keyword, the left column depicts web searches whereas the right column depicts news searches.

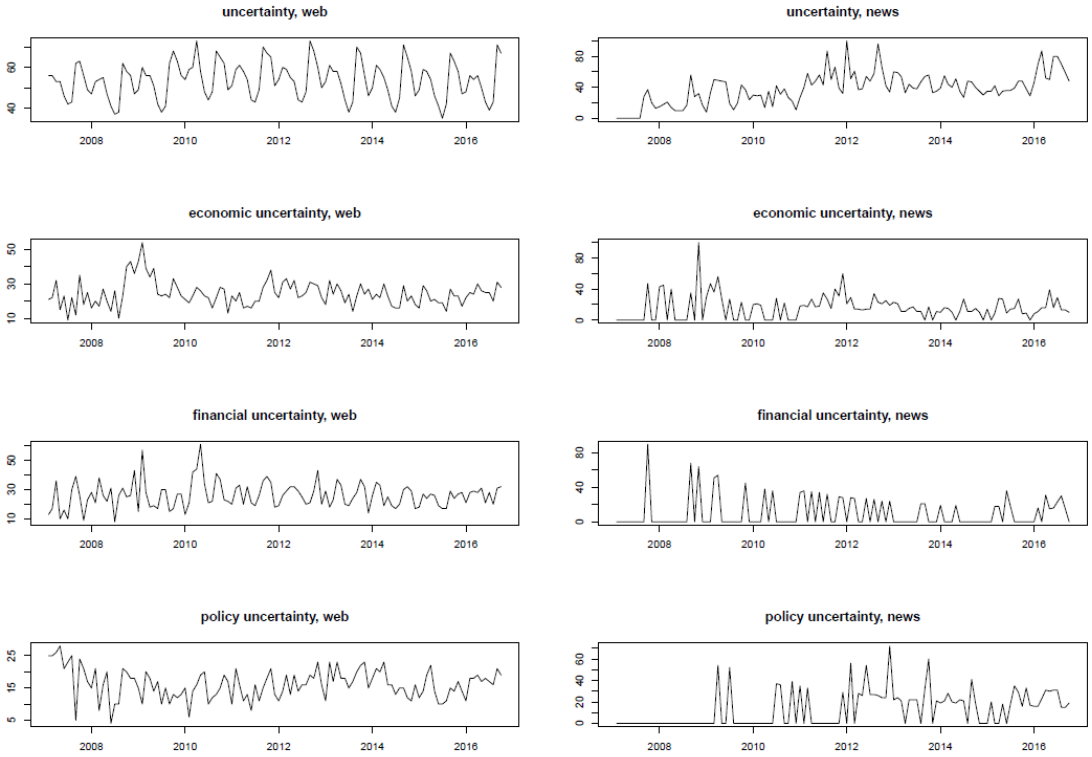
As mentioned above, the uncertainty index is constructed by taking the equal weighted average of all Google Trends. Figure 15 compares the resulting Google Uncertainty Index to the corresponding Reuters one. We see that, generally, the two indexes tend to move in the same direction. In particular, before 2012 the correlation coefficient of the two indexes is 0.53 and falls to 0.39 in the period after 2012. During the full sample, the two series are positively correlated at 0.48.

However, the above analysis is based on the raw series which have not been checked for outliers and seasonal patterns. Using the rule-of-thumb outlier detection rule, we define outliers as all values which are 4 standard deviations away from the mean. Actually, using this definition there are no outliers, therefore Figure 16 simply offers a time series plot of the series without any outliers indicated. However, as we show in the Appendix, other more specific indexes do present outliers.

Using the STL decomposition method, we then estimate the seasonal and trend components for the two series, as illustrated in Figure 17 and Figure 18. Finally, we compare de-seasonalised series in Figure 19. It turns out that the pre-2012 correlation is about 0.52, however the post-2012 correlation has now increased to 0.53, indicating that the lower correlation we detected before in this period was mainly caused by different seasonal components. The overall correlation of the seasonally adjusted RUI and Google indexes is also slightly higher than for their unadjusted counterparts, at 0.52.[10]

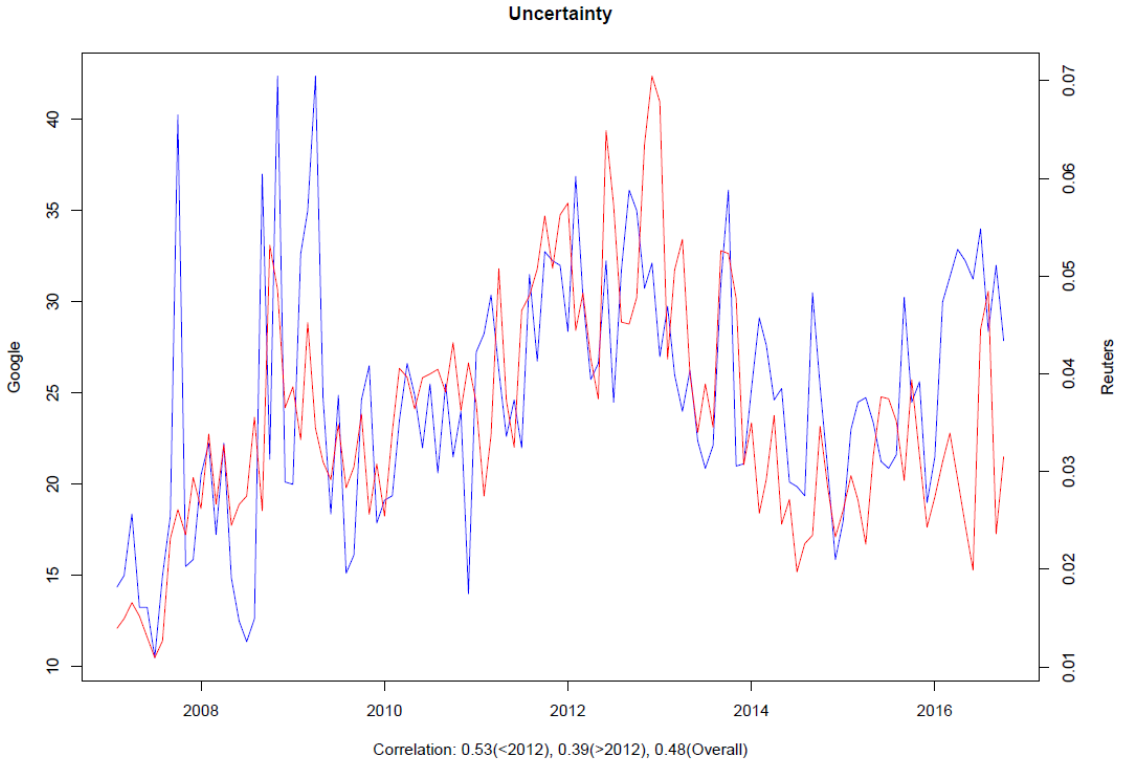
([10]) We also compare the Google index to EPU and VIX. We find that the correlation of Google and the Global EPU prior to 2012 is 0.45, increasing to 0.66 in the post-2012 period. Overall, the series are correlated at 0.59 which is lower than the correlation between Reuters and the EPU index. Using VIX, we see that the series seem to be uncorrelated. The figures can be found in the Appendix.

Figure 14: Google Trends to construct the General Uncertainty Index



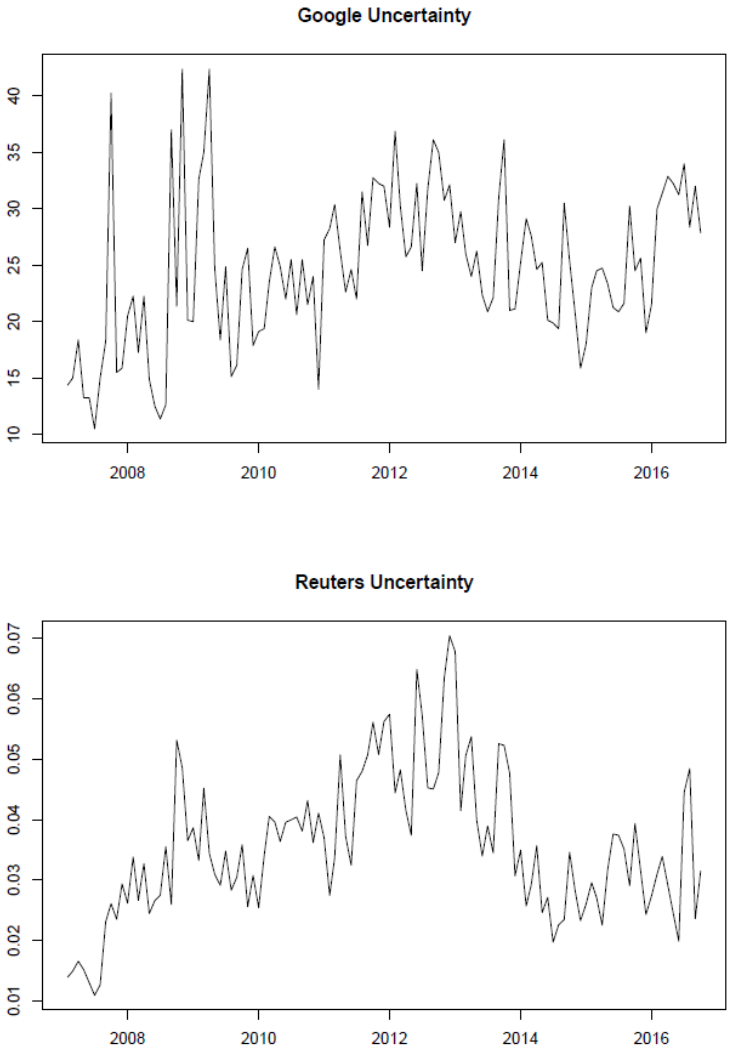
Source: Based on author's calculations

Figure 15: **Comparing the General Uncertainty Index of Google (left axis, blue colour) to the corresponding Reuters index (right axis, red colour). The correlations before 2012, after 2012 and during the whole sample are mentioned below the figure**



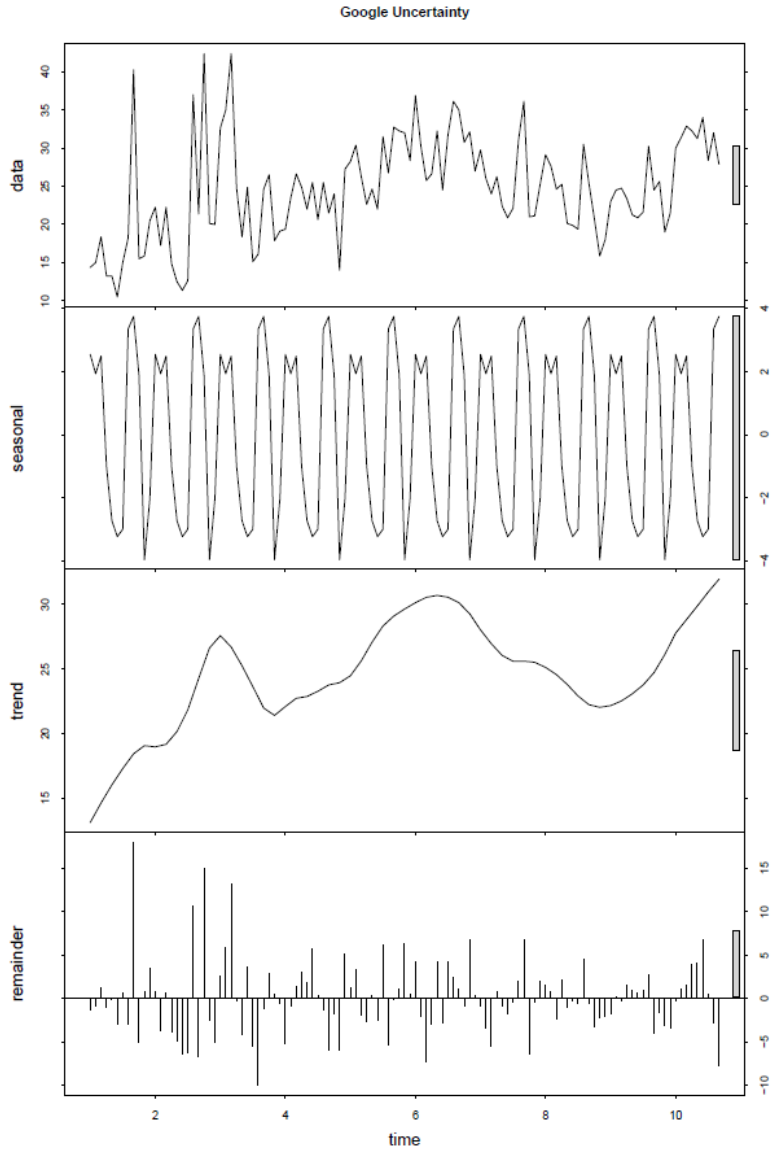
Source: Based on author's calculations

Figure 16: Detecting outliers



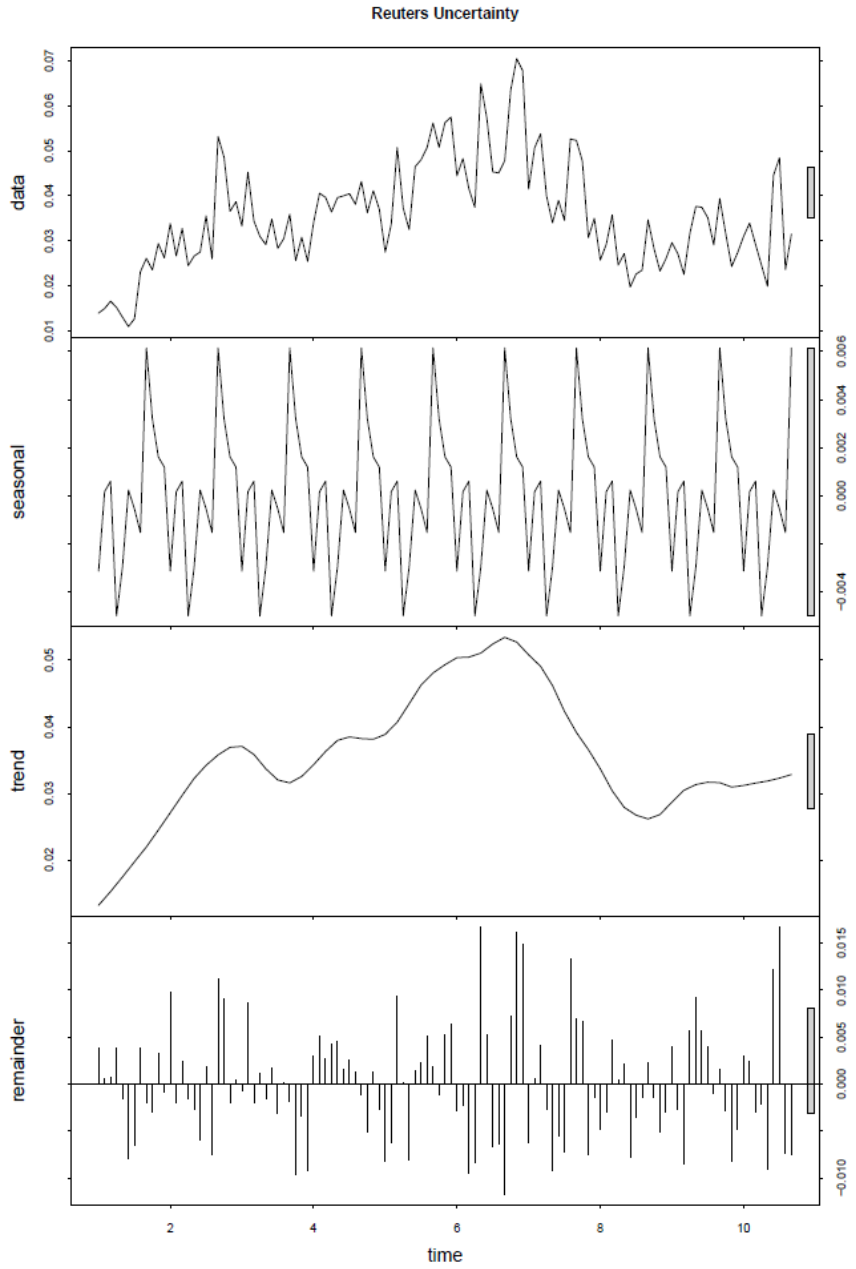
Source: Based on author's calculations

Figure 17: Google Uncertainty Index, STL decomposition



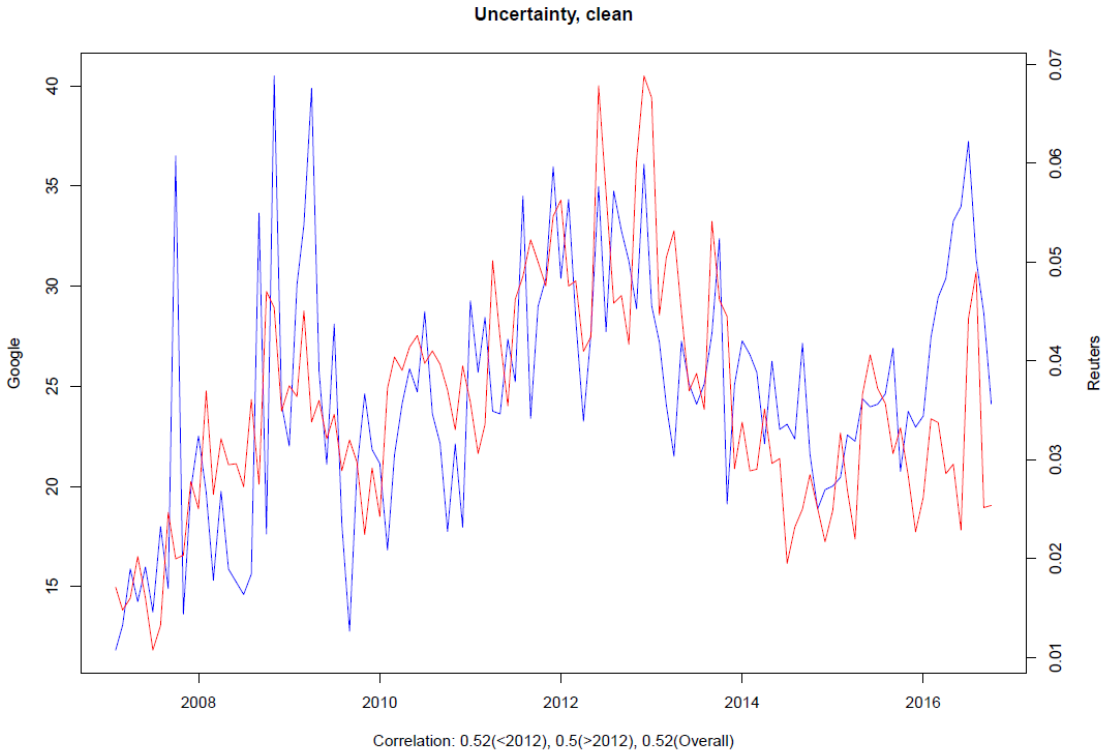
Source: Based on author's calculations

Figure 18: Reuters Uncertainty Index, STL decomposition



Source: Based on author's calculations

Figure 19: De-trended and de-seasonalised Uncertainty Indexes: Google (blue), Reuters (red)



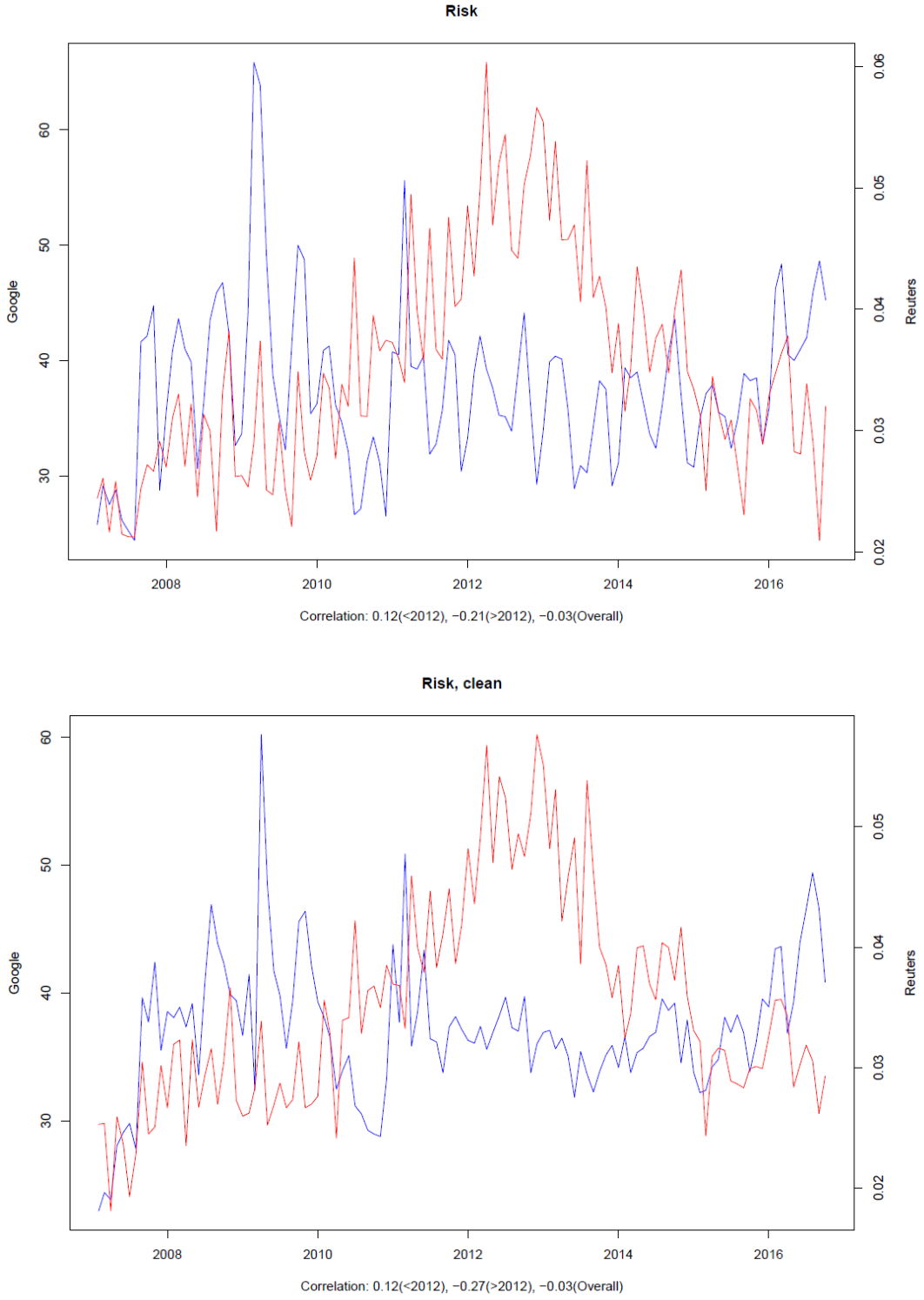
Source: Based on author's calculations

4.2.2 GENERAL RISK INDEX

As we mention above, for all subsequent cases we present only the final Reuters and Google risk indexes, before and after cleaning, while all detailed figures are in the Appendix.

In a similar manner to the General Uncertainty Index, we construct the General Risk Index searching worldwide for the keyword "risk" in web and news searches. From Figure 20, before the cleaning there is positive correlation between the Reuters and Google risk indexes in the pre-2012 period and negative correlation in the post- 2012 period. Overall, the series seem to be uncorrelated with a correlation coefficient equal to -0.03. In this case, the removal of the seasonal components has only minor effects.

Figure 20: General Risk Index based on Google Trends (blue) and Reuters (red). Top panel: the indexes before cleaning. Bottom panel: the outliers-free, de-trended and de-seasonalised series



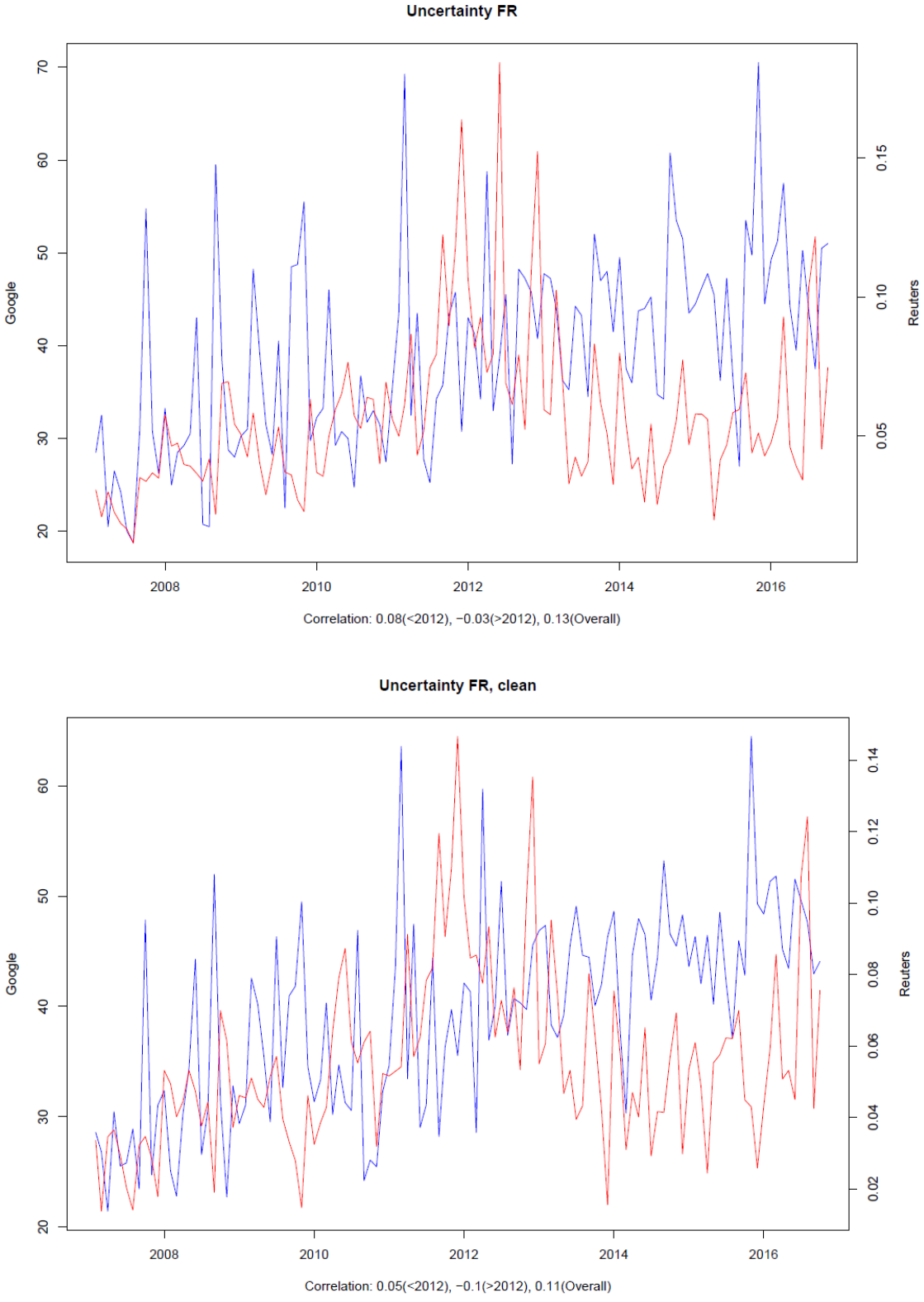
Source: Based on author's calculations

4.2.3 FRENCH UNCERTAINTY INDEX

To construct the French Uncertainty Index, as well as those for the other countries, in order to limit a proliferation of cases, we use keywords related to both uncertainty and risk. Specifically, for France, we use “incertitude” and “risque” and, as before, we include both web and news searches.

In Figure 21 we compare the resulting indexes with their Reuters counterparts. Before the cleaning, there is positive correlation at 0.08 in the pre-2012 period and negative correlation at -0.03 after 2012. Overall, the series are slightly positively correlated at 0.13. After the de-seasonalisation of the series, we observe a similar result in the pre-2012 period, and a more negative correlation coefficient equal to -0.1 after 2012. Overall, the correlation is now at 0.11.

Figure 21: France Uncertainty Index based on Google Trends (blue) and Reuters (red). Top panel: the indexes before cleaning. Bottom panel: the outliers-free, de-trended and de-seasonalised series



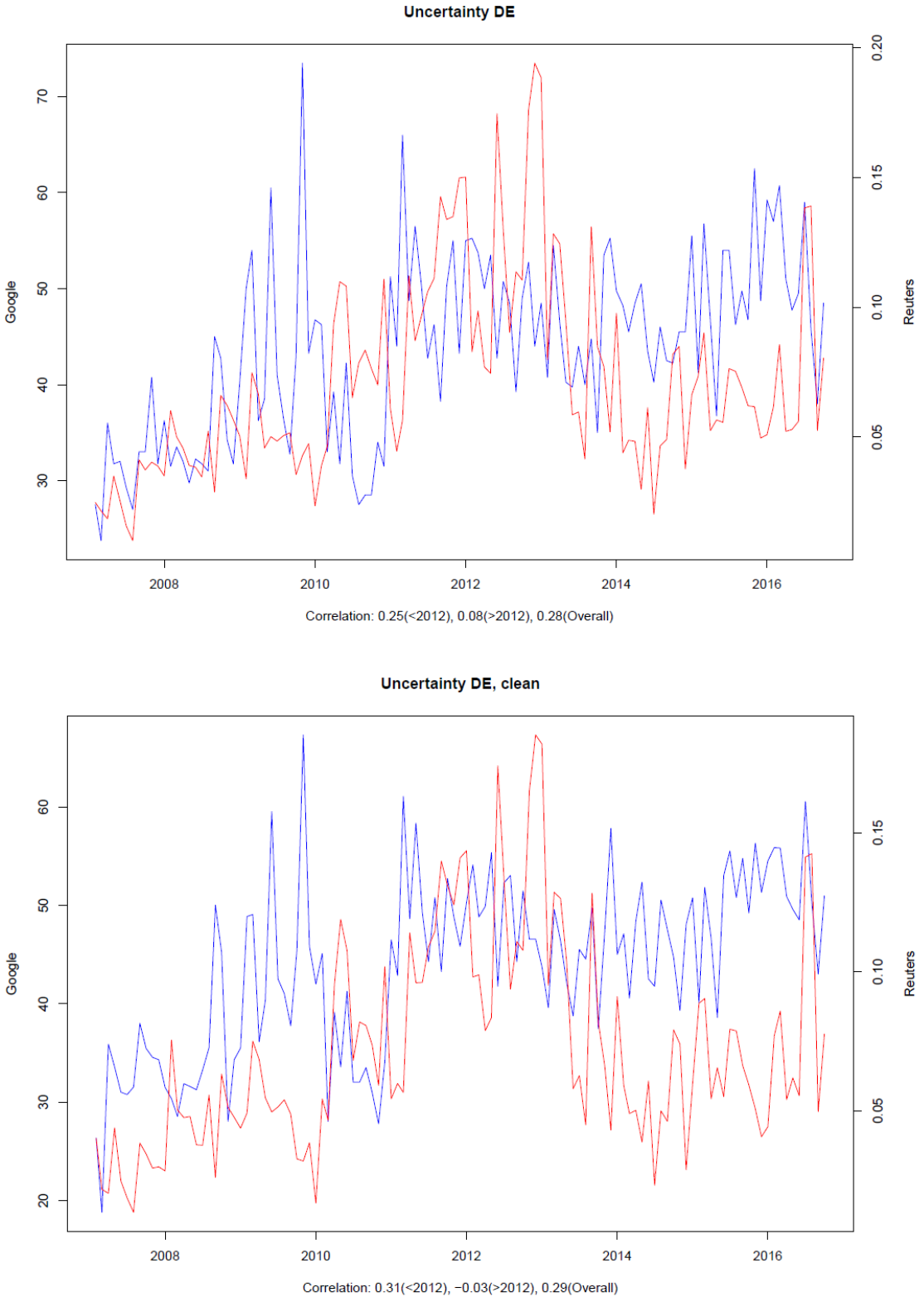
Source: Based on author's calculations

4.2.4 GERMANY UNCERTAINTY INDEX

To construct the German Uncertainty Index, we use “unsicherheit” and “risiko” as keywords and, as before, we include both web and news searches.

From Figure 22, we see that overall both Google and Reuters uncleaned indexes (top panel) move in the same direction. They both rise as we approach 2012 with a correlation of 0.25 in the pre-2012 period. Then, in the post-2012 period the correlation falls to 0.08. Overall, the indexes are positively correlated at 0.28 using the full sample. After the removal of the seasonal component, the indexes become more correlated before 2012 at 0.31, and uncorrelated in the post-2012 period at -0.03. Over the full sample, the correlation is similar to the unadjusted case, at 0.29.

Figure 22: Germany Uncertainty Index based on Google Trends (blue) and Reuters (red). Top panel: the indexes before cleaning. Bottom panel: the outliers-free, de-trended and de-seasonalised series



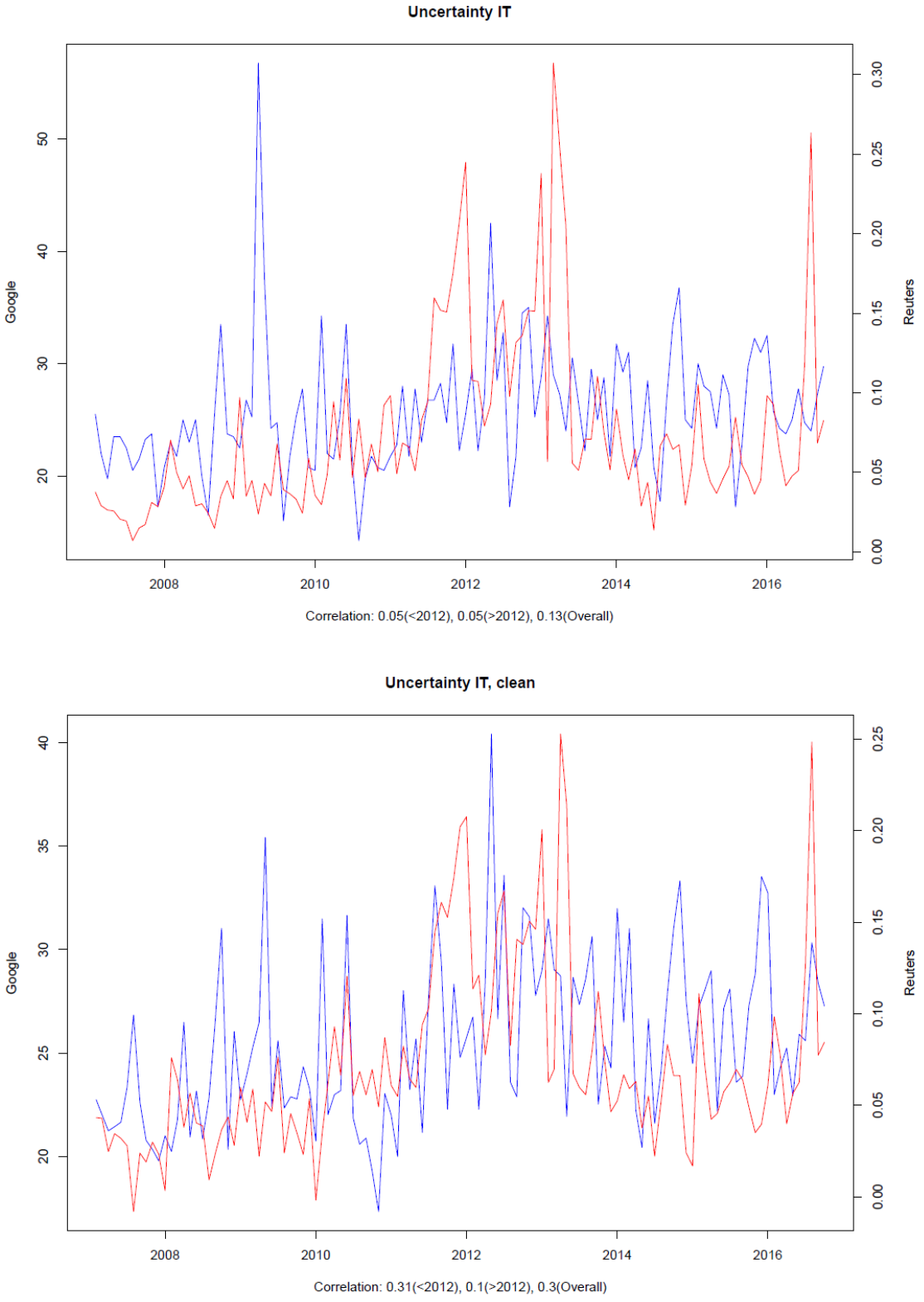
Source: Based on author's calculations

4.2.5 ITALY UNCERTAINTY INDEX

To construct the Italian Uncertainty Index we again use two keywords: “incertezza” and “rischio”, including both web and news searches.

From Figure 23, the overall correlation of the Reuters and Google indexes doubles from 0.13 to 0.30 after the removal of the seasonal component. Looking at the bottom panel of Figure 23, we see that before 2012 the correlation of the series is 0.31 (much higher compared to 0.05 which is the estimate before the cleaning) and 0.1 after 2012.

Figure 23: Italy Uncertainty Index based on Google Trends (blue) and Reuters (red). Top panel: the indexes before cleaning. Bottom panel: the outliers-free, de-trended and de-seasonalised series



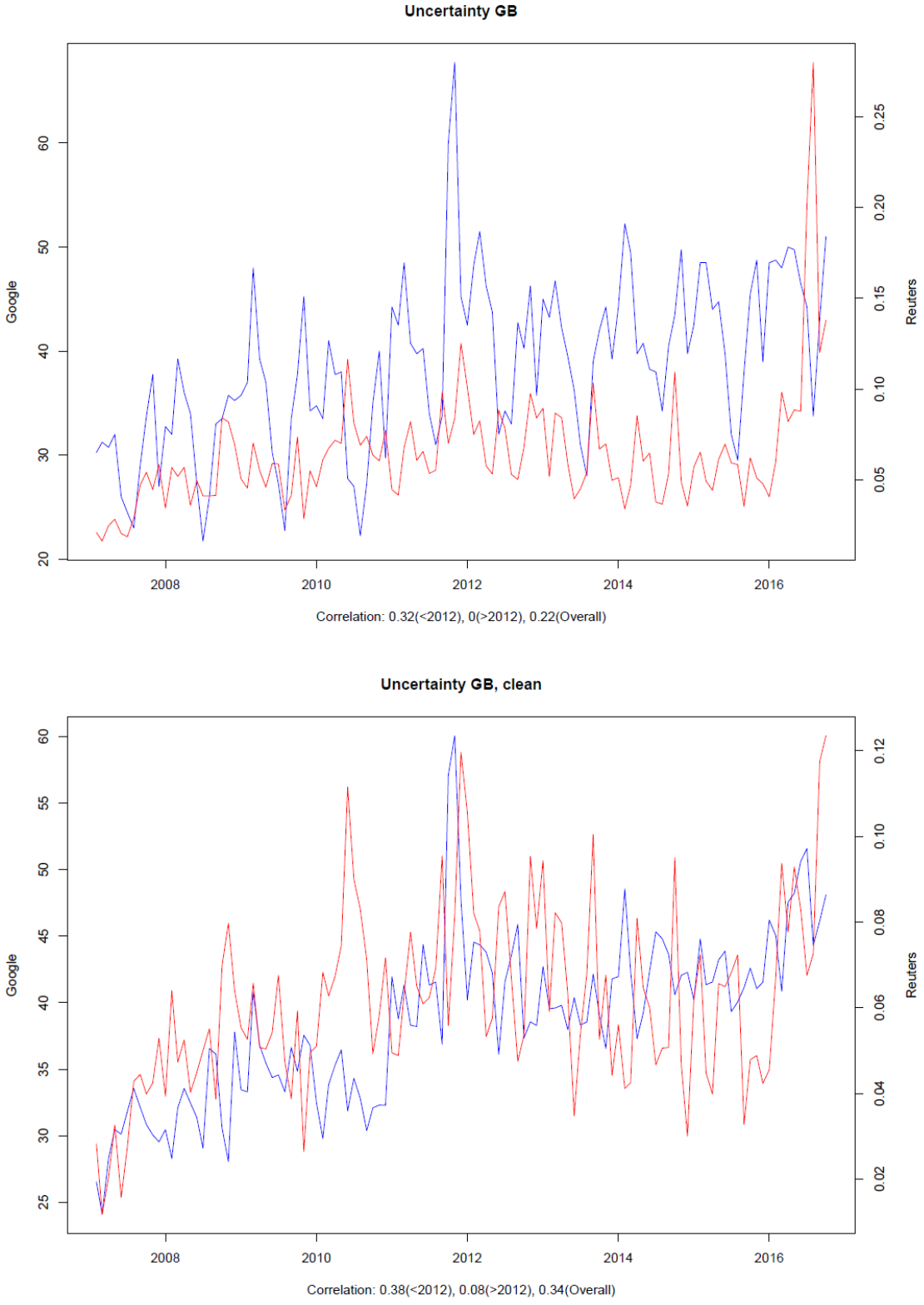
Source: Based on author's calculations

4.2.6 UNITED KINGDOM UNCERTAINTY INDEX

The Google based UK Uncertainty Index relies on the two keywords “uncertainty” and “risk”.

From the top panel of Figure 24, the Reuters and Google uncertainty indexes for the UK are positively correlated in the pre-2012 period, with a correlation coefficient equal to 0.32, but are uncorrelated in the post-2012 period. It is remarkable to see that the Reuters index spikes towards the end of the sample indicating a large volume of “uncertainty” keywords in news articles, which is not evident in the web searches before the cleaning. Moreover, looking at the bottom panel of Figure 24, we see that after the removal of the seasonal component, the correlation in the indexes has increased across all periods and, more importantly, both indexes spike from 2016 onwards.

Figure 24: United Kingdom Uncertainty Index based on Google Trends (blue) and Reuters (red). Top panel: the indexes before cleaning. Bottom panel: the outliers-free, de-trended and de-seasonalised series



Source: Based on author's calculations

5 Conclusion

In this report we discuss the importance of a proper treatment of outliers and seasonal patterns in the construction of indexes based on big data. With respect to standard time series analysis, the number of variables under evaluation and their possibly very high frequency complicates the analysis. Overall, we suggest to look at the variables one by one, and to apply robust methods for outlier detection and seasonality removal. The STL filtering approach based on local linear regression, loess method ("Seasonal and Trend decomposition using Loess"), seems particularly suited, as it is computationally very fast and can also handle high frequencies.

We have implemented this method first on simulated data and then on actual big data based uncertainty indexes. The analysis with simulated data has revealed that seasonal patterns are not easily identified in big data. For example, it is difficult to observe a daily or weekly seasonality when looking at minute by minute or more frequent data. However, these patterns emerge when aggregating the data to lower frequency, suggesting that the outlier detection and removal of seasonal patterns should be conducted first on the original data but then also on the aggregated/constructed data.

As a more practical application of the methods, we have constructed various Google based uncertainty indexes, and compared them with the previously developed Reuters indexes. We have "cleaned" all indexes for outliers and seasonal components, and compared the original and cleaned versions. In several cases, we found an increased correlated between the Google indexes and the Reuters indexes after removing the seasonal components. Moreover, some interesting patterns can be masked by the presence of temporal patterns, such as an increase in uncertainty in the UK from 2016 afterwards. This provides a clear indication of the importance of a proper pre-treatment of the big data based indicators prior to their use in economic analyses.

References

1. Baker, S., Bloom, N., Davis, S.J. (2016). “Measuring Economic Policy Uncertainty”, The Quarterly Journal of Economics, 131(4), 1593–1636.
2. Banbura, M., Giannone, D., Modugno, M., Reichlin, L. (2013). ”Now-Casting and the Real-Time Data Flow”, ECB Working Paper Series, No 1564.
3. Canova, F., Ghysels, E. (1994). “Changes in seasonal patterns : Are they cyclical?”, Journal of Economic Dynamics and Control, 18(6), 1143–1171.
4. Chen, C., Liu, L.-M. (1993). “Joint Estimation of Model Parameters and Outlier Effects in Time Series”, Journal of the American Statistical Association, 88(421), 284–297.
5. Cleveland, R. B., Cleveland, W.S., McRae, J.E., Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3–73.
6. Davidson, R., MacKinnon, J. G. (2004). “Econometric Theory and Methods”, New York: Oxford University Press.
7. De Livera, A.M., Hyndman, R.J., Snyder, R.D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association, 106(496), 1513–1527.
8. Elliott, G., Rothenberg, T.J., Stock, J.H. (1996). “Efficient Tests for an Autoregressive Unit Root”, Econometrica, 64(4), 813–836.
9. Giannone, D., Reichlin, L., Small, D. (2008). “Nowcasting GDP and Inflation: The Real-Time Informational Content of Macroeconomic Data Releases”, Journal of Monetary Economics, 55, 665–676.
10. Harvey, A. C., (1989). “Forecasting, Structural Time Series Models and the Kalman Filter”, Cambridge University Press.
11. Harvey, A., Koopman, S.J., Riani, M. (1997). The modeling and seasonal adjustment of weekly observations. Journal of Business & Economic Statistics, 15(3), 354–368.
12. Hyndman, R., Athanasopoulos, G. (2014). “Forecasting: Principles and practice”, Otexts.
13. Knight, F.H. (1921). “Risk, Uncertainty, and Profit”, Boston, MA: Hart, Schaff & Marx; Houghton Mifflin Co.
14. Ladiray, D., Proietti, T. (2017). “Eurostat Project on Seasonal Effects on Daily and Weekly Data”, Eurostat, European Commission.
15. Ladiray, D., Quenneville, B. (2012). Seasonal adjustment with the x-11 method (Vol. 158). Springer Science & Business Media.
16. Lo´pez-de-Lacalle, J. (2015). “Package ‘tsoutliers’”.
17. Maravall, A. (2008a). “Notes on programs TRAMO and SEATS: TRAMO part”, Technical Report, Bank of Spain.
18. Maravall, A. (2008b). “Notes on programs TRAMO and SEATS: SEATS part”, Technical Report, Bank of Spain.
19. McCracken, M.W., Ng, S. (2015). “FRED-MD: A Monthly Database for Macroeconomic Research”, Working Paper 2015-012B, Federal Reserve Bank of St. Louis: Research Division Working Paper Series.
20. Ng, S. (2016). Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data. Working Paper.

21. Pedregal, D.J., Young, P. C. (2006). Modulated cycles, an approach to modelling periodic components from rapidly sampled data. International Journal of Forecasting, 22(1), 181–194.

22. Pierce, D.A., Grupe, M.R., Cleveland, W.P.(1984). Seasonal adjustment of the weekly monetary aggregates: A model-based approach. Journal of Business & Economic Statistics, 2(3), 260–270.

23. Proietti, T. (2000). Comparing seasonal components for structural time series models. International Journal of Forecasting, 16(2), 247–260.

24. Said, S.E., Dickey, D.A. (1984). “Testing for Unit Roots in Autoregressive- Moving Average Models of Unknown Order”, Biometrika, 71(3), 599–607.

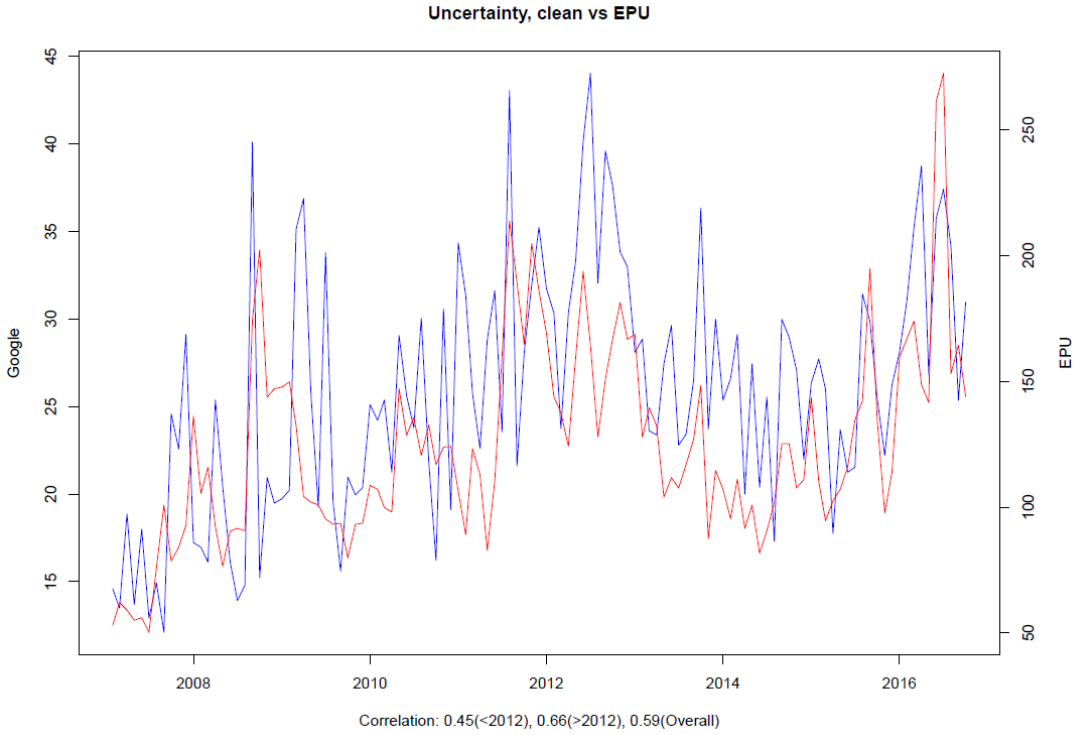
25. Stock, J.H., Watson, M.W. (2002a). “Forecasting using Principal Components from a Large Number of Predictors”, Journal of the American Statistical Association, 97, 147–162.

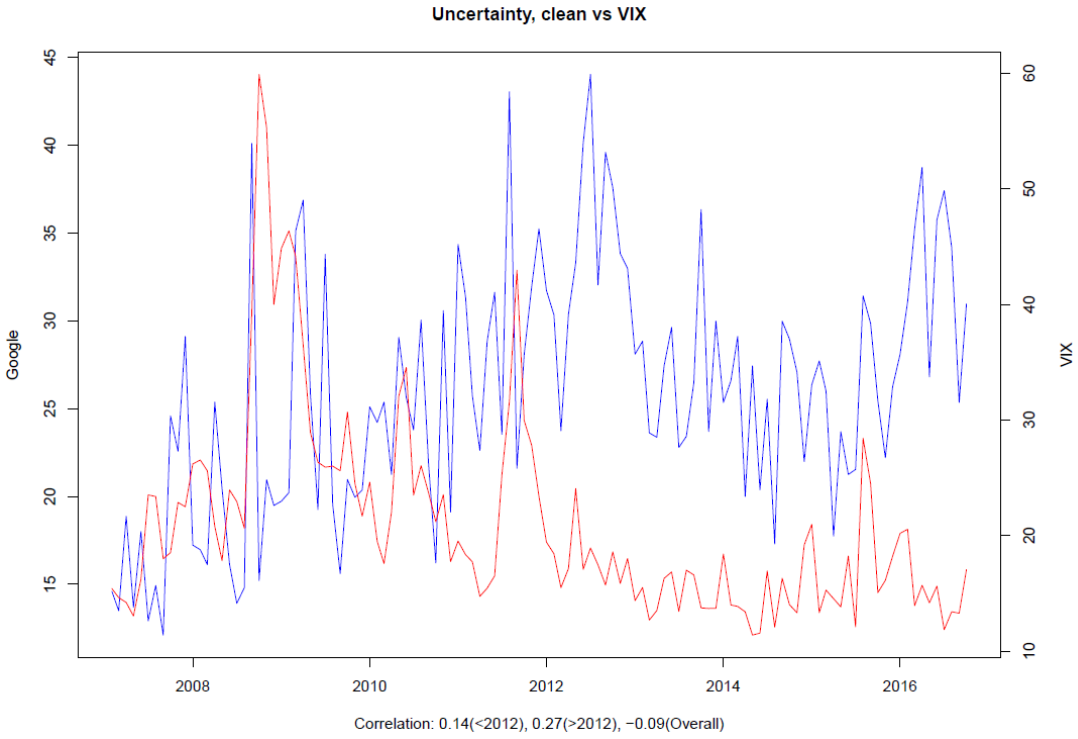
26. Stock, J.H., Watson, M.W. (2002b). “Macroeconomic Forecasting using Diffusion Indexes”, Journal of Business and Economic Statistics, 20, 147–162.

27. Stock J.H. and M.W. Watson (2006), ”Forecasting with Many Predictors”, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.) *Handbook of Economic Forecasting*.

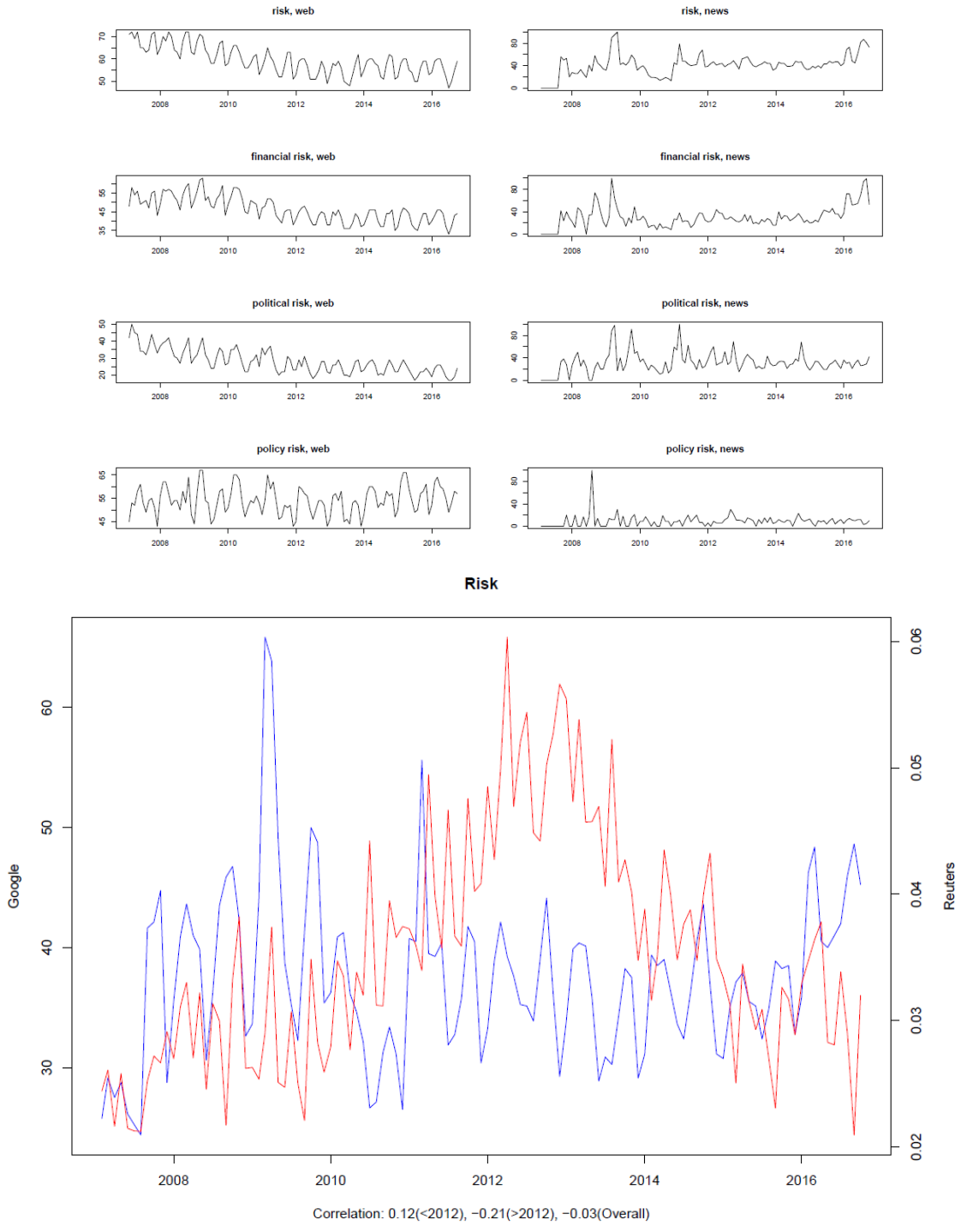
Appendix: All figures

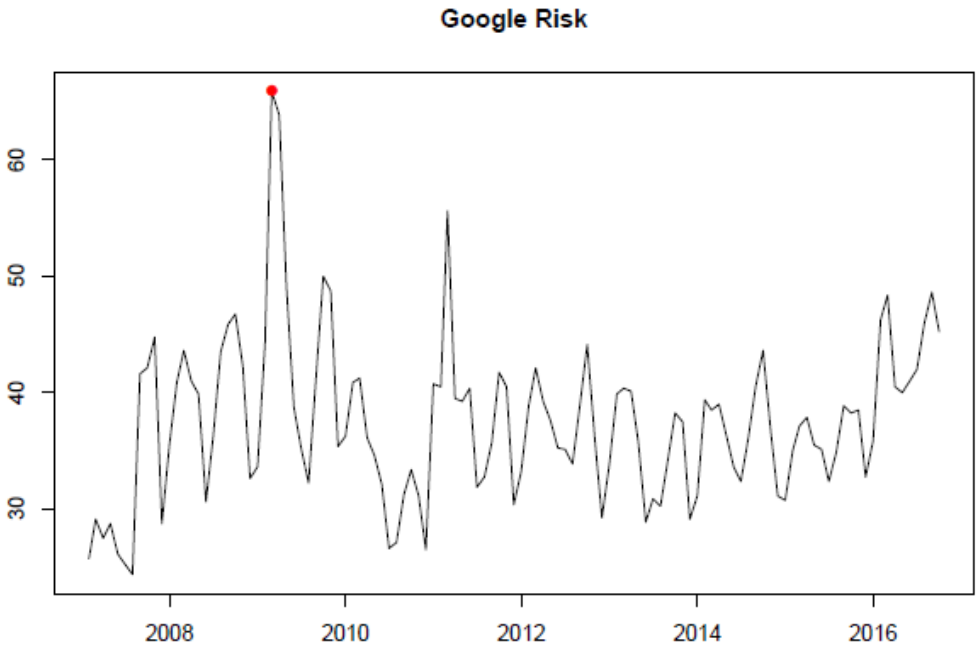
Google uncertainty vs EPU and VIX

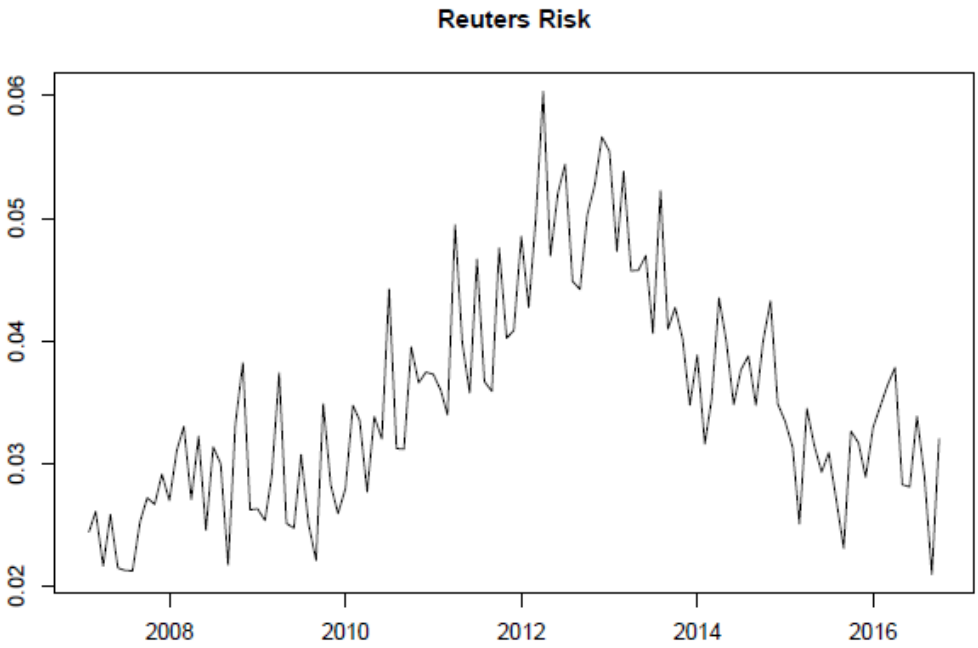


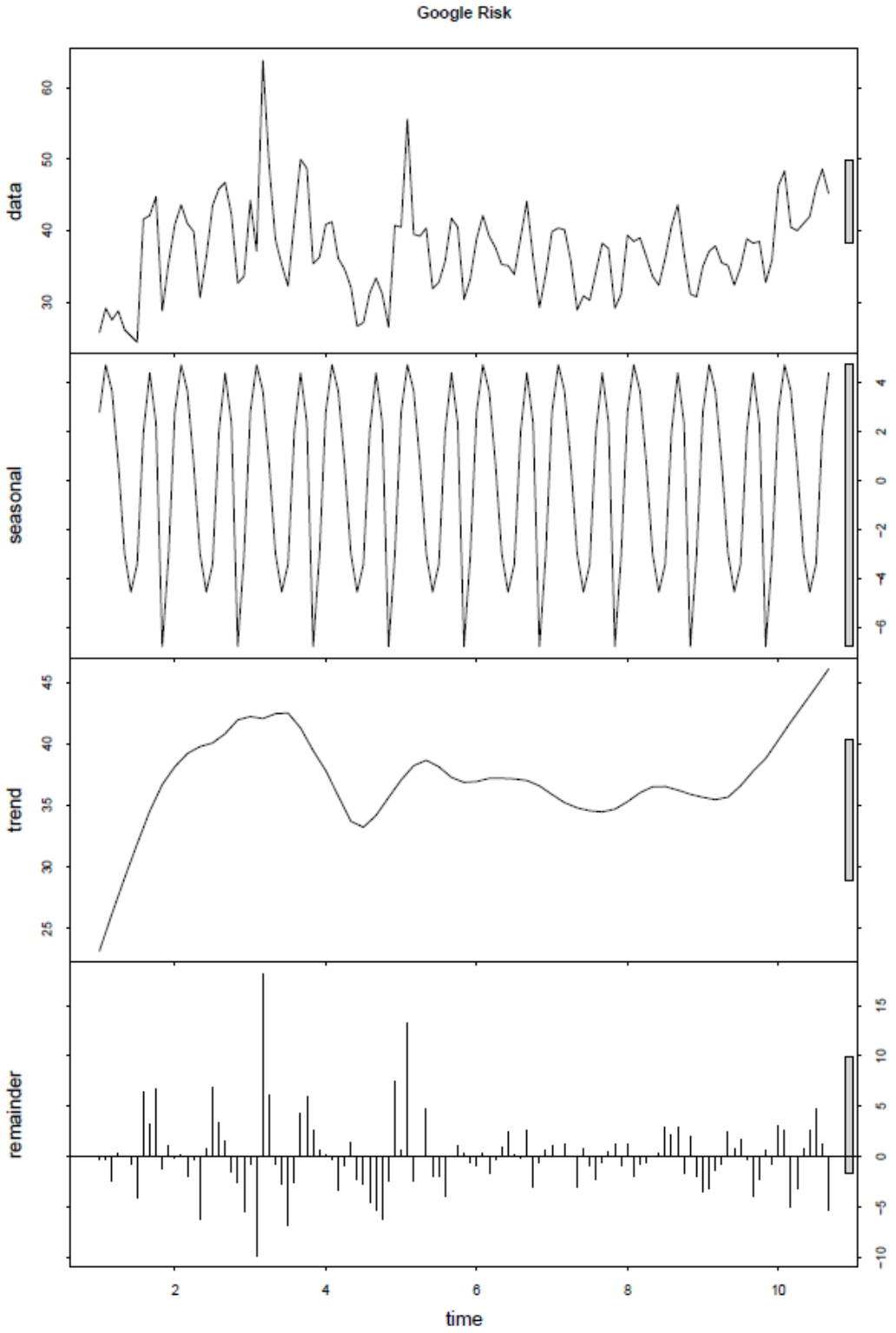


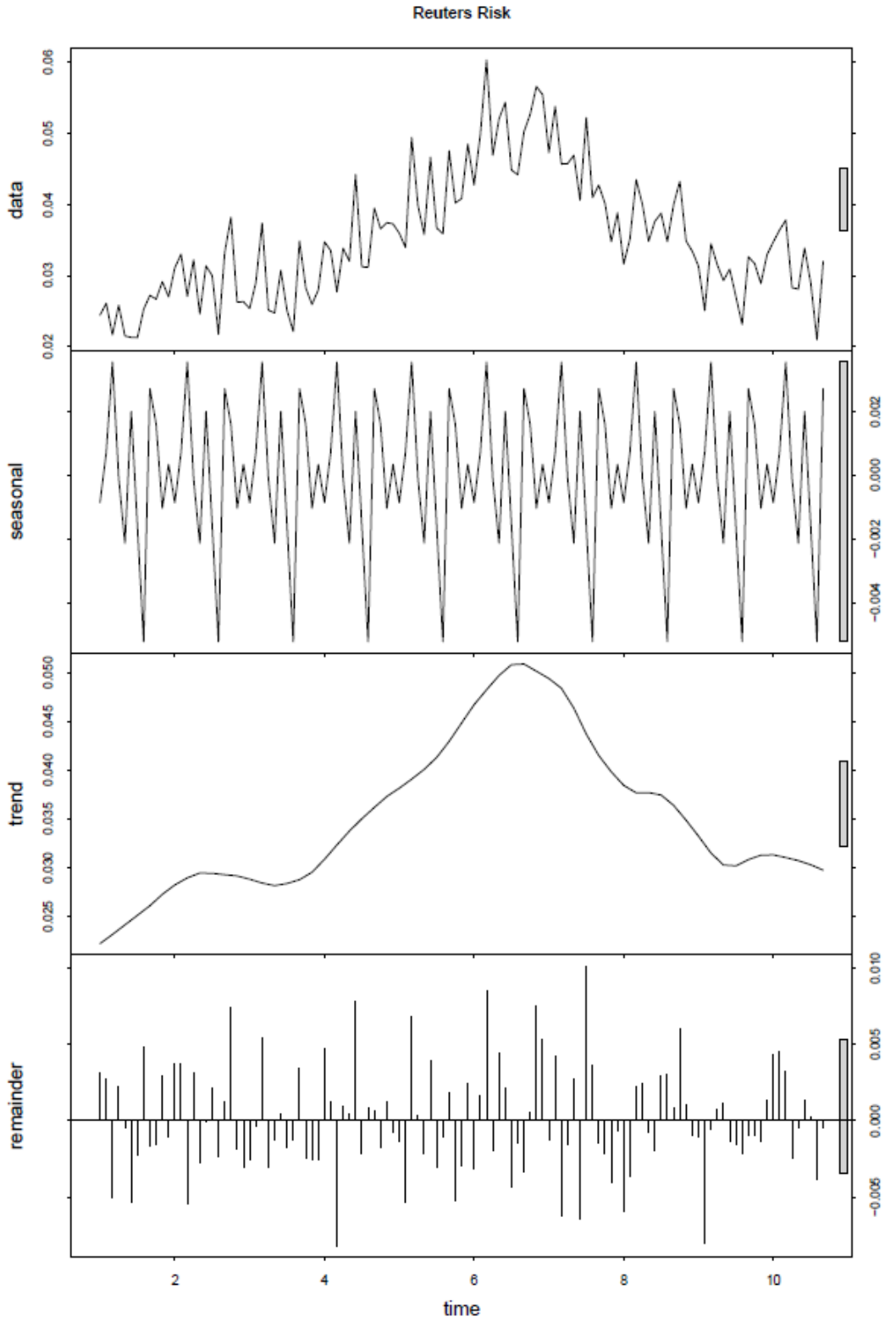
General risk index

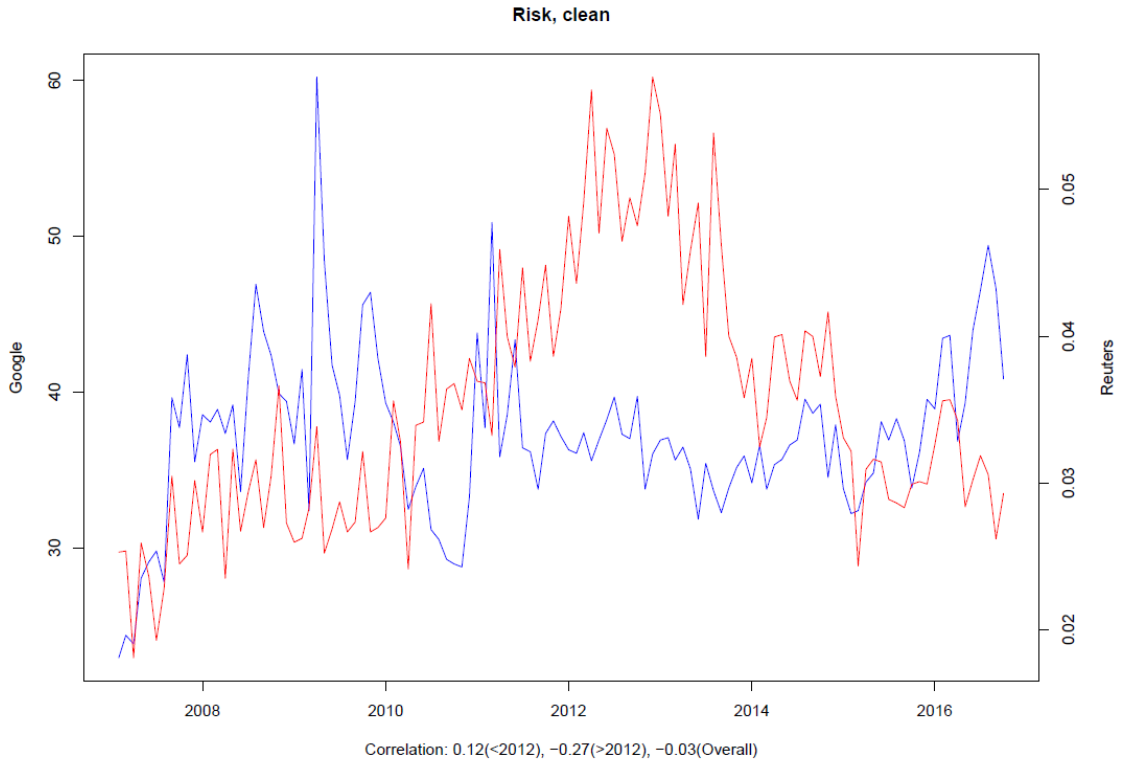



Google Risk
60
50
40
30
2008
2010
2012
2014
2016

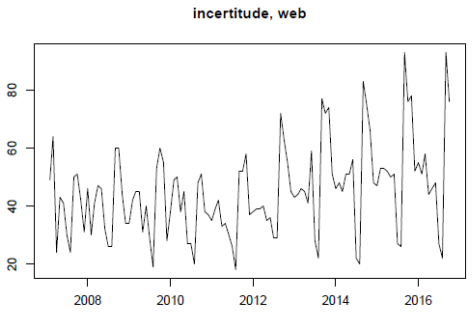

Reuters Risk
0.06
0.05
0.04
0.03
0.02
2008
2010
2012
2014
2016

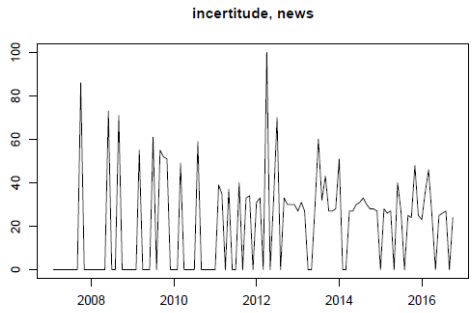

Google Risk
data
seasonal
trend
remainder
time
2
4
6
8
10

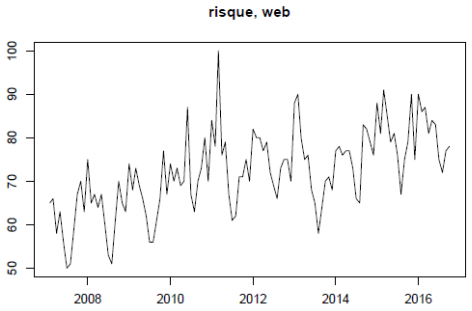

Reuters Risk
data
seasonal
trend
remainder
time
2
4
6
8
10

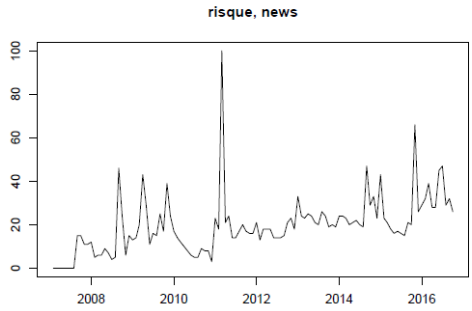

Risk, clean
Google
Reuters
60
50
40
30
0.05
0.04
0.03
0.02
2008
2010
2012
2014
2016
Correlation: 0.12(<2012), −0.27(>2012), −0.03(Overall)

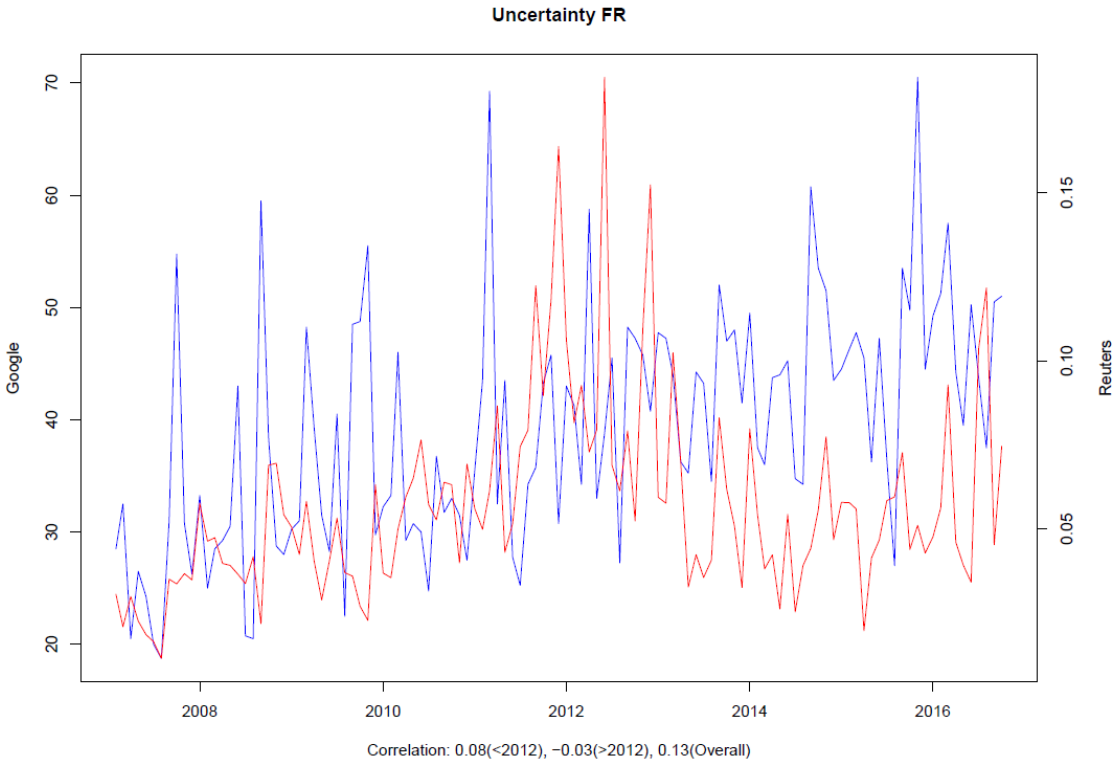
France

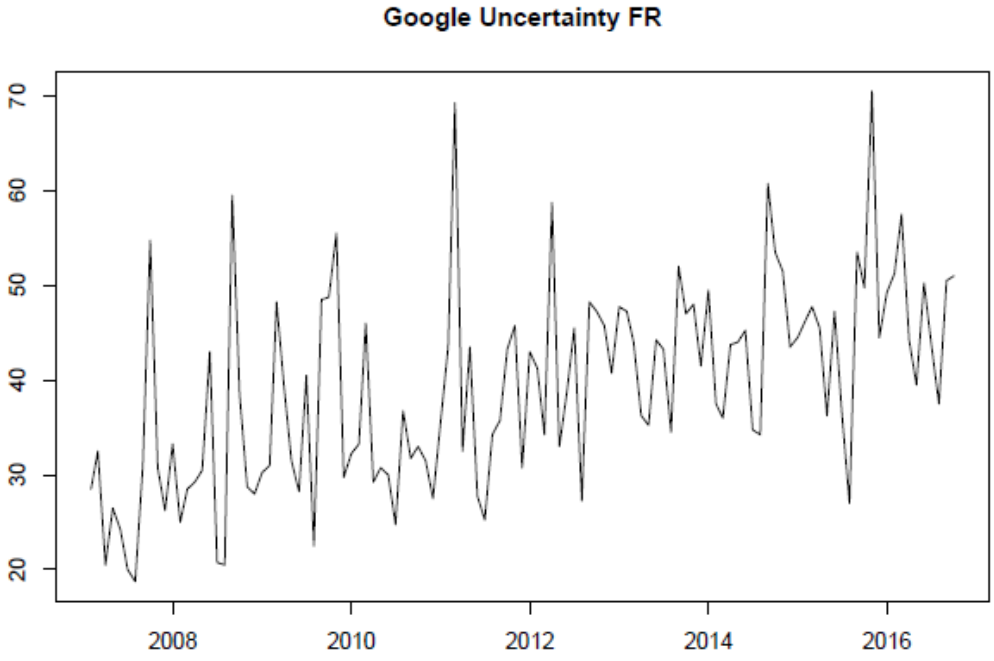

incertitude, web
20
40
60
80
2008
2010
2012
2014
2016

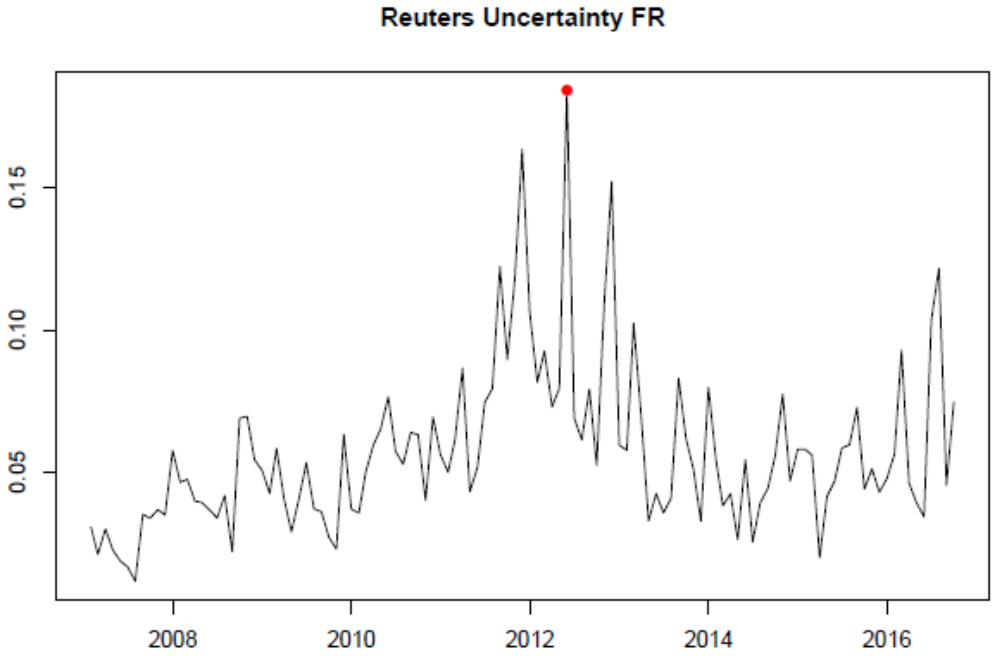

incertitude, news
0
20
40
60
80
100
2008
2010
2012
2014
2016

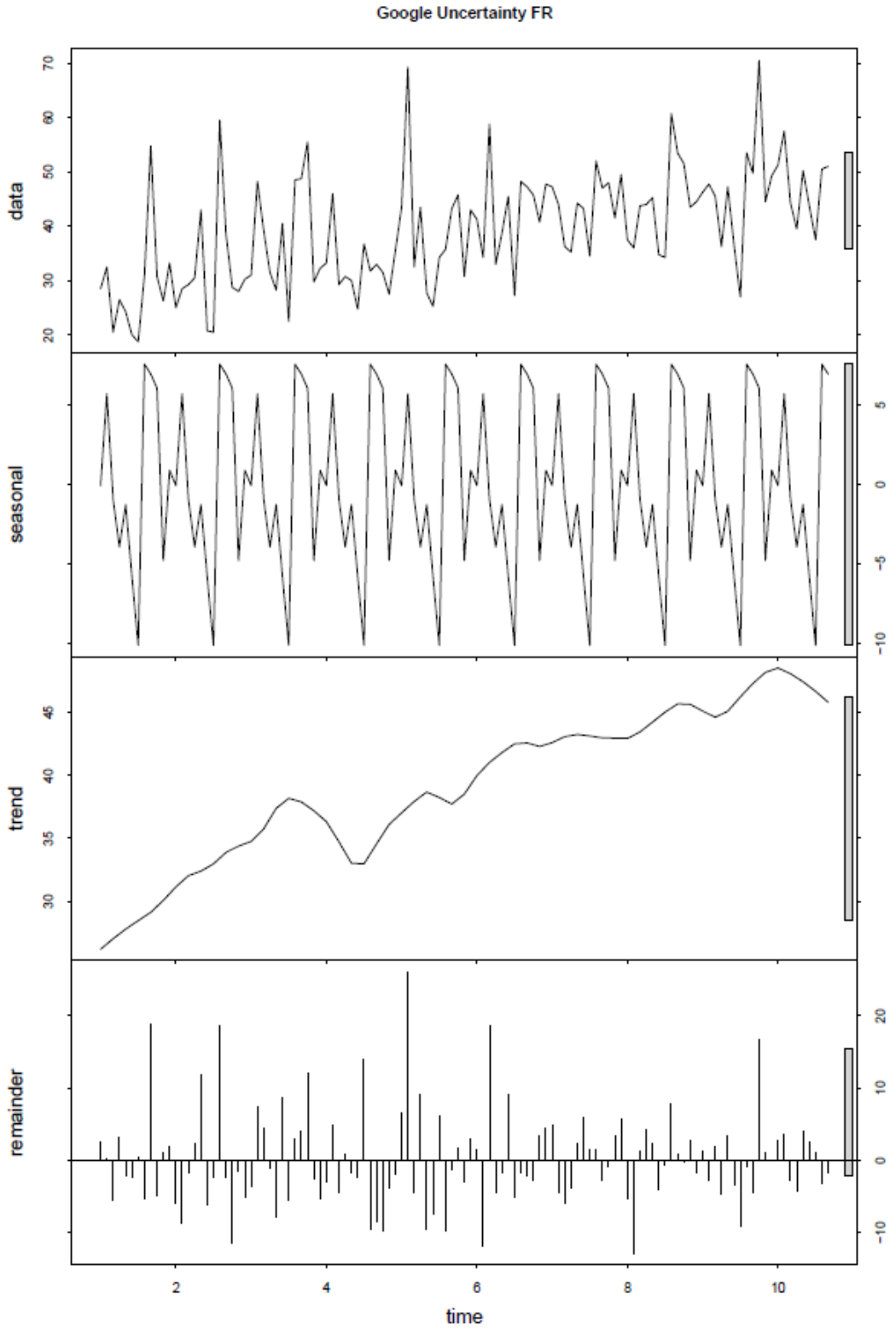

risque, web
50
60
70
80
90
100
2008
2010
2012
2014
2016

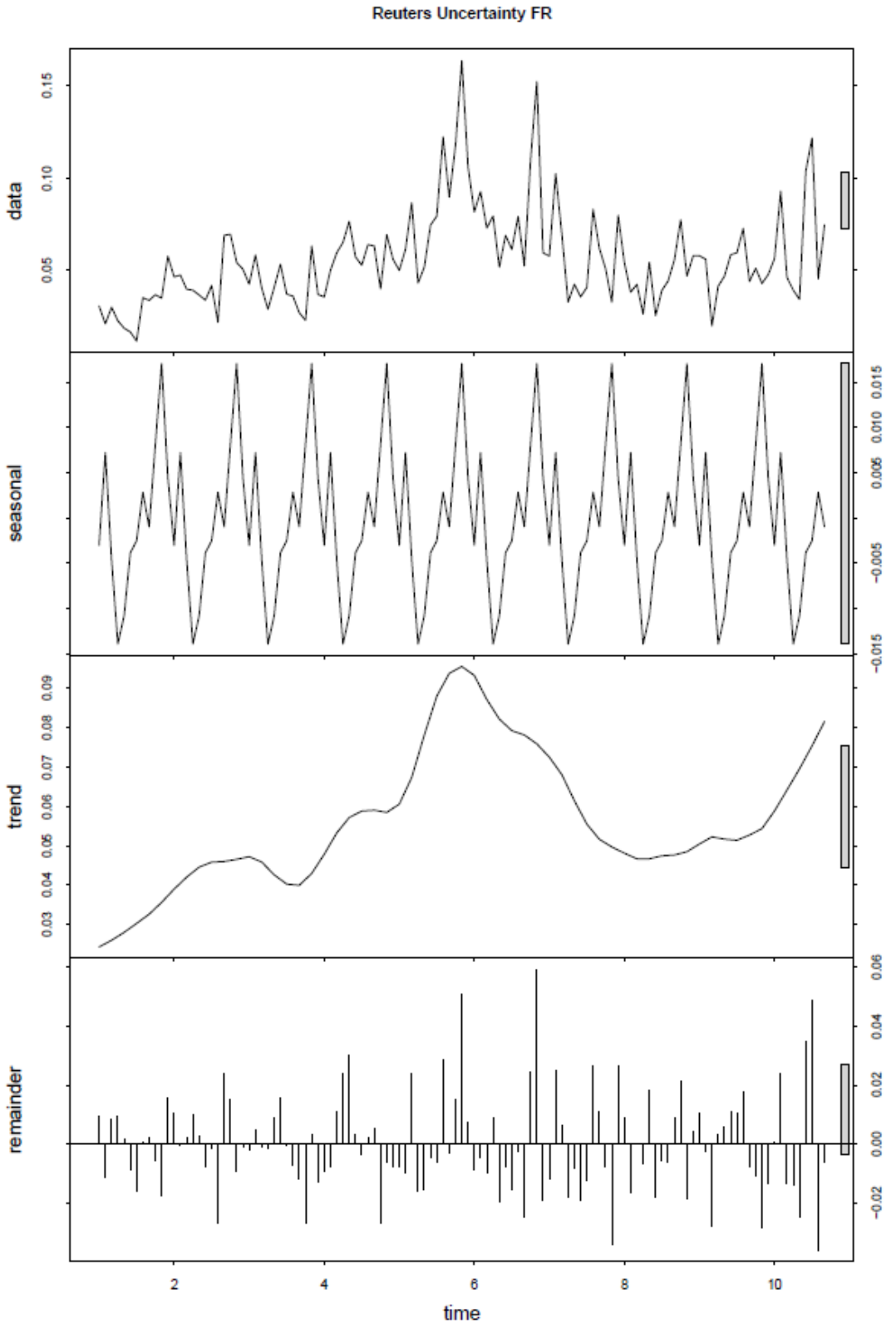

risque, news
0
20
40
60
80
100
2008
2010
2012
2014
2016

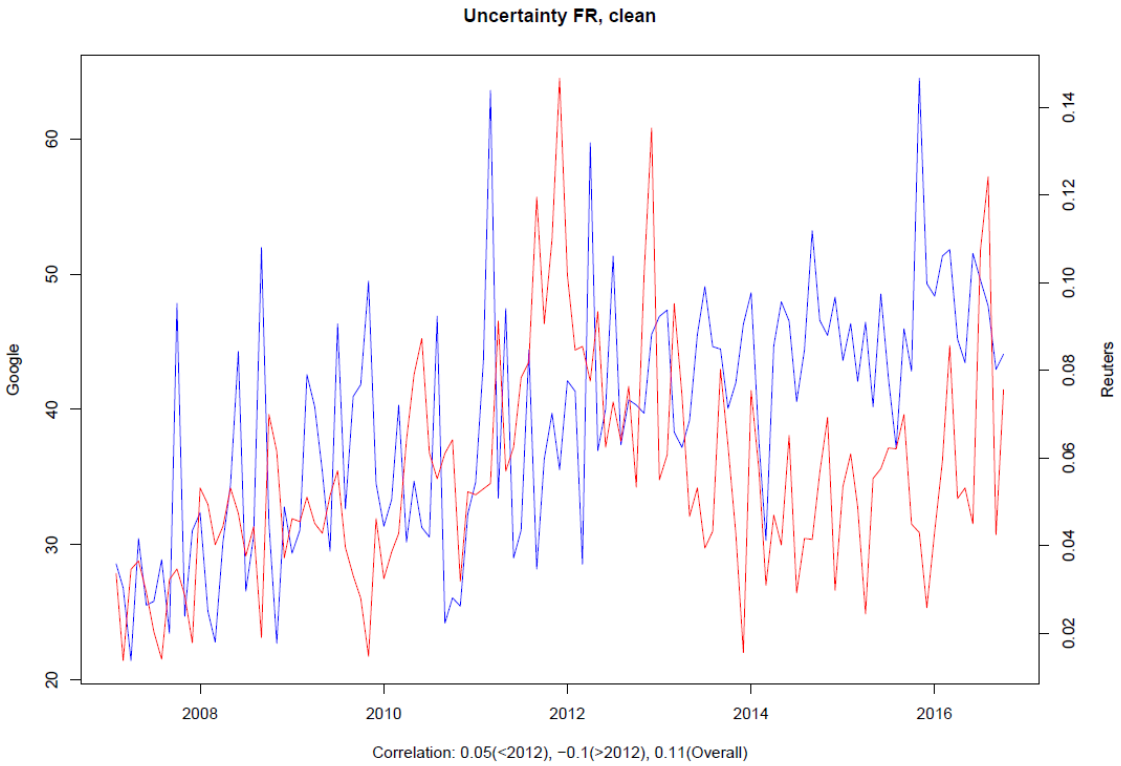

Uncertainty FR
Google
Reuters
20
30
40
50
60
70
0.05
0.10
0.15
2008
2010
2012
2014
2016
Correlation: 0.08(<2012), −0.03(>2012), 0.13(Overall)


Google Uncertainty FR
70
60
50
40
30
20
2008
2010
2012
2014
2016

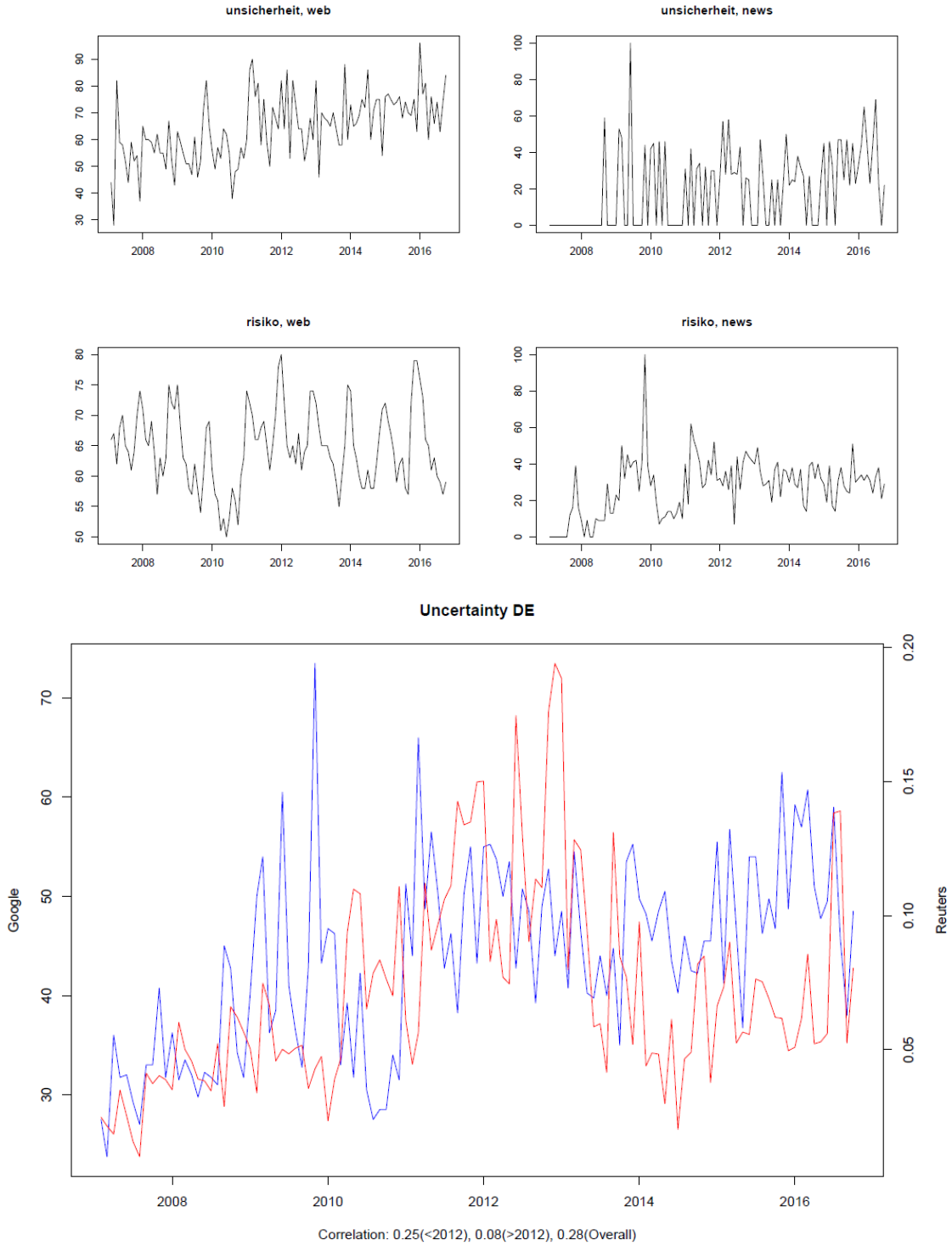

Reuters Uncertainty FR
0.15
0.10
0.05
2008
2010
2012
2014
2016

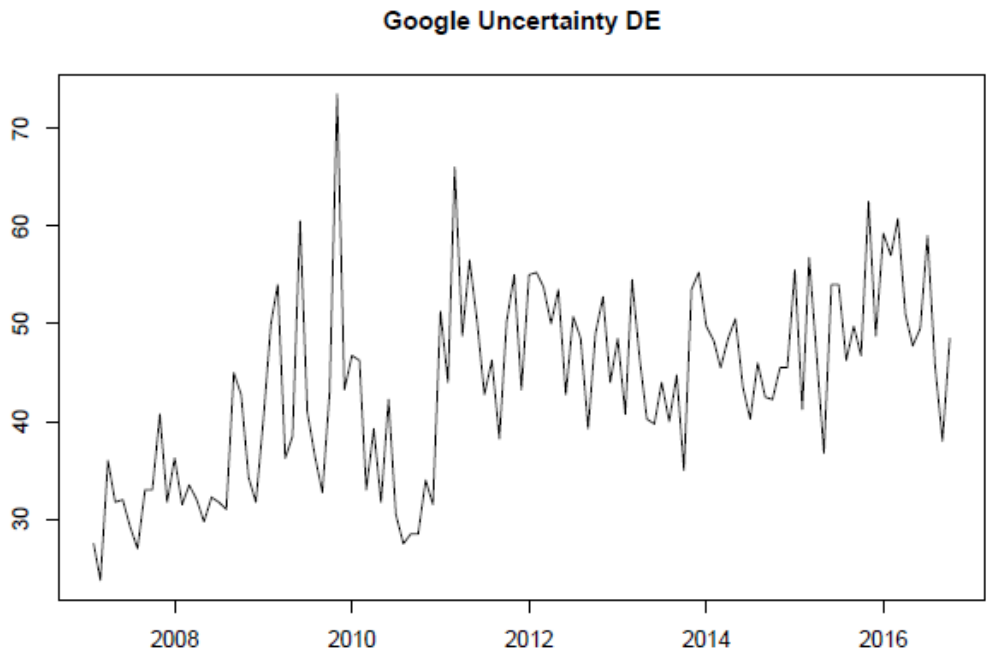

Google Uncertainty FR
data
seasonal
trend
remainder
time
20
30
40
50
60
70
-10
-5
0
5
30
35
40
45
-10
0
10
20
2
4
6
8
10

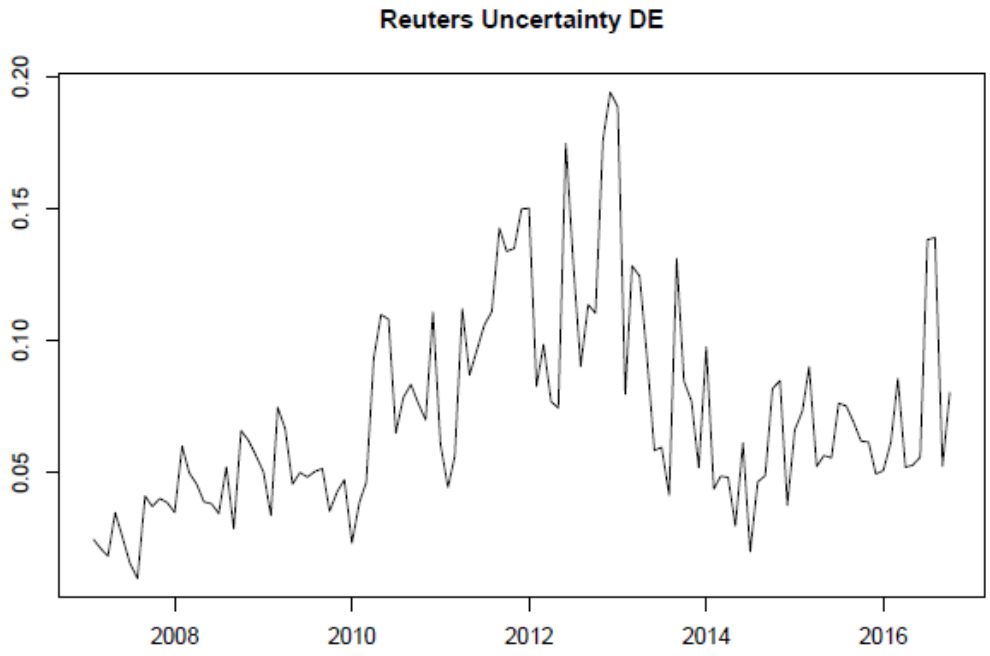

Reuters Uncertainty FR
data
seasonal
trend
remainder
time
2
4
6
8
10

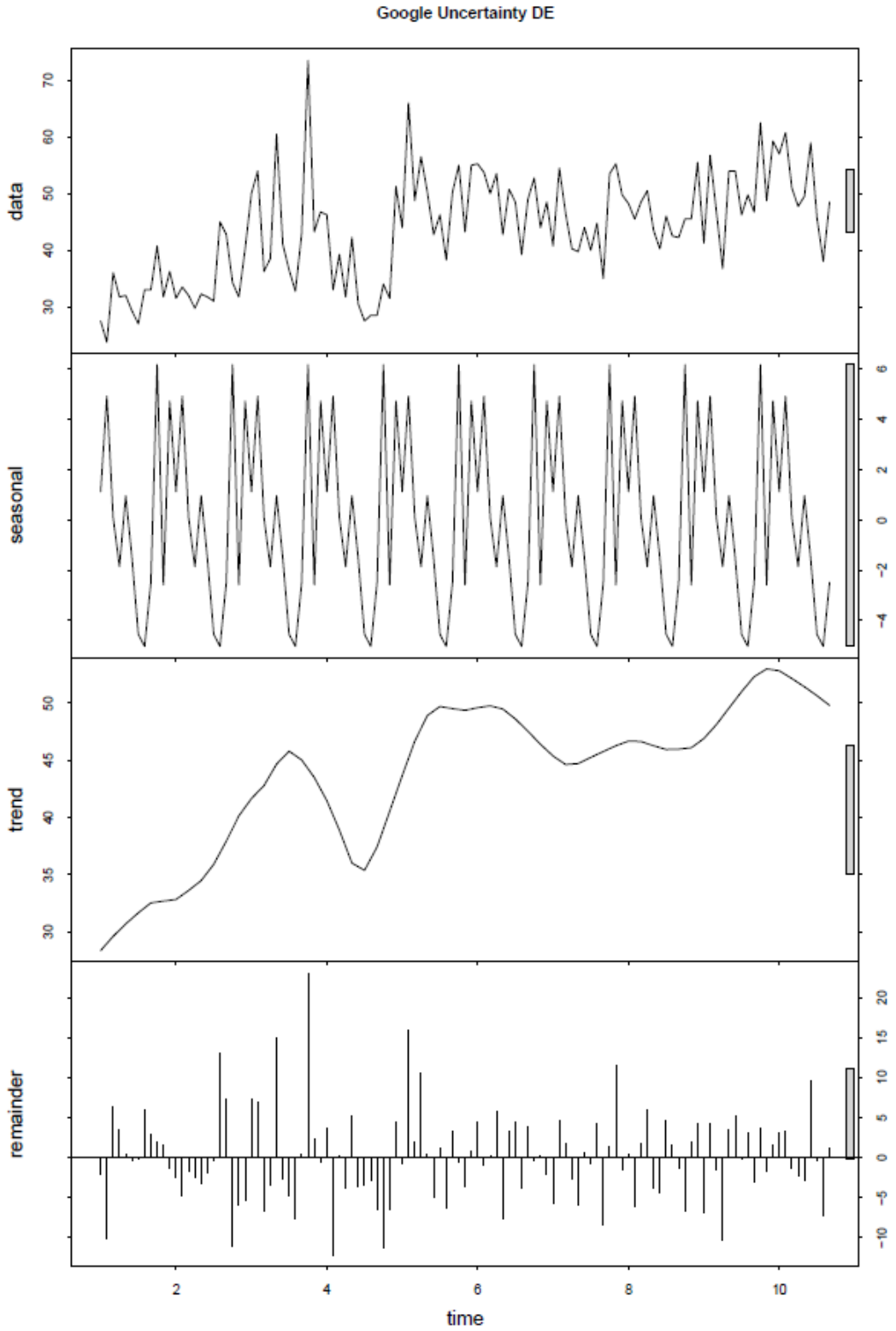

Uncertainty FR, clean
Google
Reuters
20
30
40
50
60
0.02
0.04
0.06
0.08
0.10
0.12
0.14
2008
2010
2012
2014
2016
Correlation: 0.05(<2012), −0.1(>2012), 0.11(Overall)

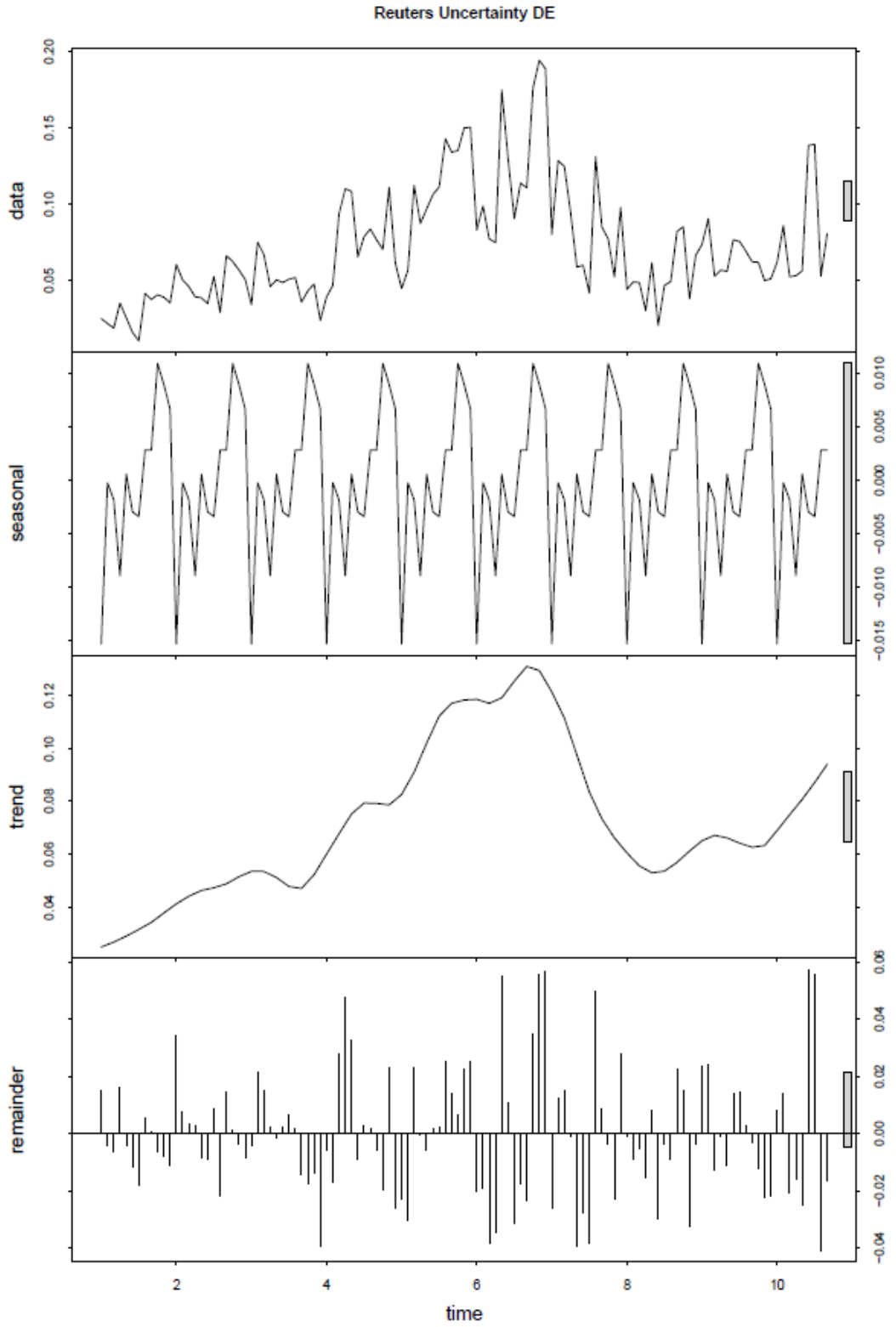
Germany

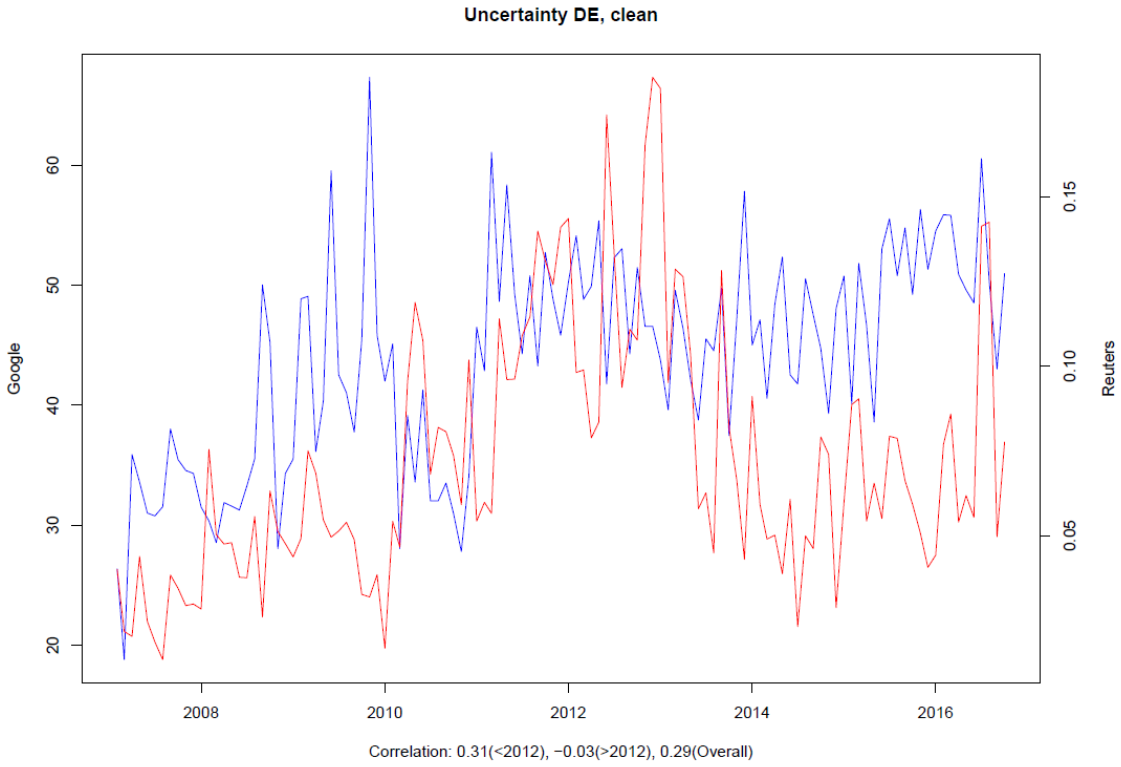

unsicherheit, web
unsicherheit, news
risiko, web
risiko, news
Uncertainty DE
Google
Reuters
Correlation: 0.25(<2012), 0.08(>2012), 0.28(Overall)


Google Uncertainty DE
70
60
50
40
30
2008
2010
2012
2014
2016

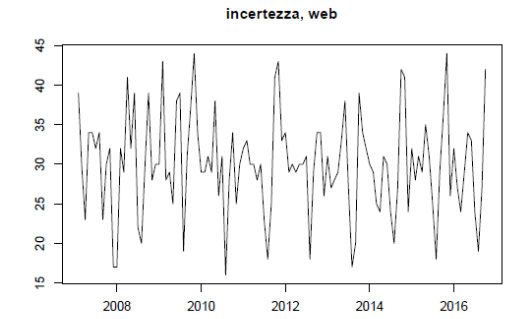

Reuters Uncertainty DE
0.20
0.15
0.10
0.05
2008
2010
2012
2014
2016

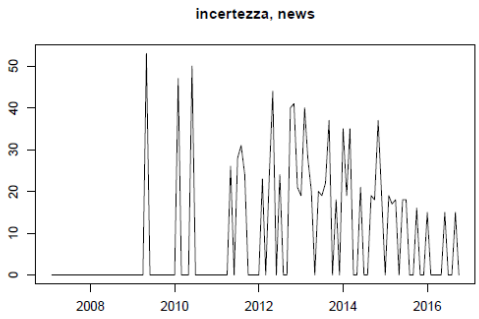

Google Uncertainty DE
data
seasonal
trend
remainder
time

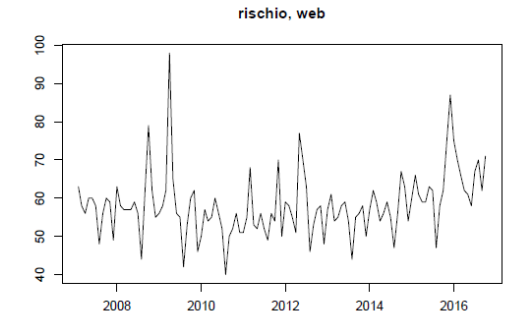

Reuters Uncertainty DE
data
seasonal
trend
remainder
time

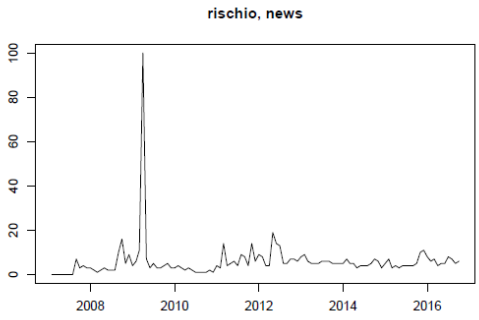

Uncertainty DE, clean
Google
Reuters
20
30
40
50
60
0.05
0.10
0.15
2008
2010
2012
2014
2016
Correlation: 0.31(<2012), −0.03(>2012), 0.29(Overall)

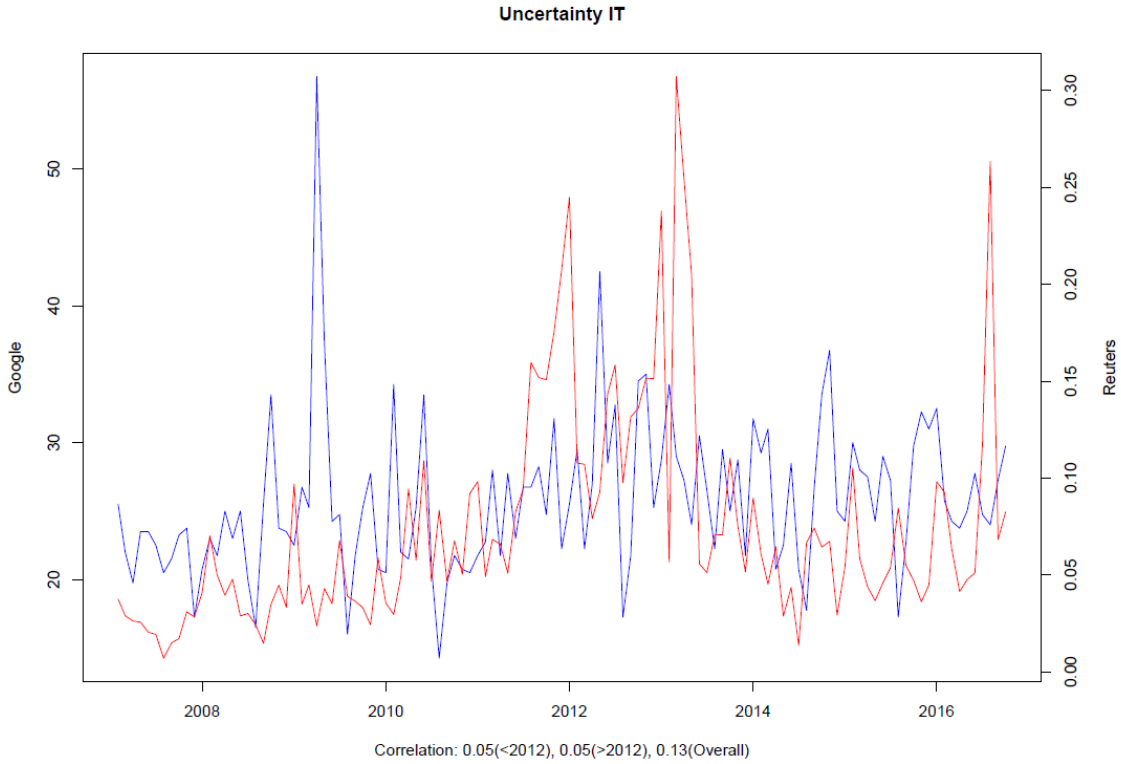
Italy

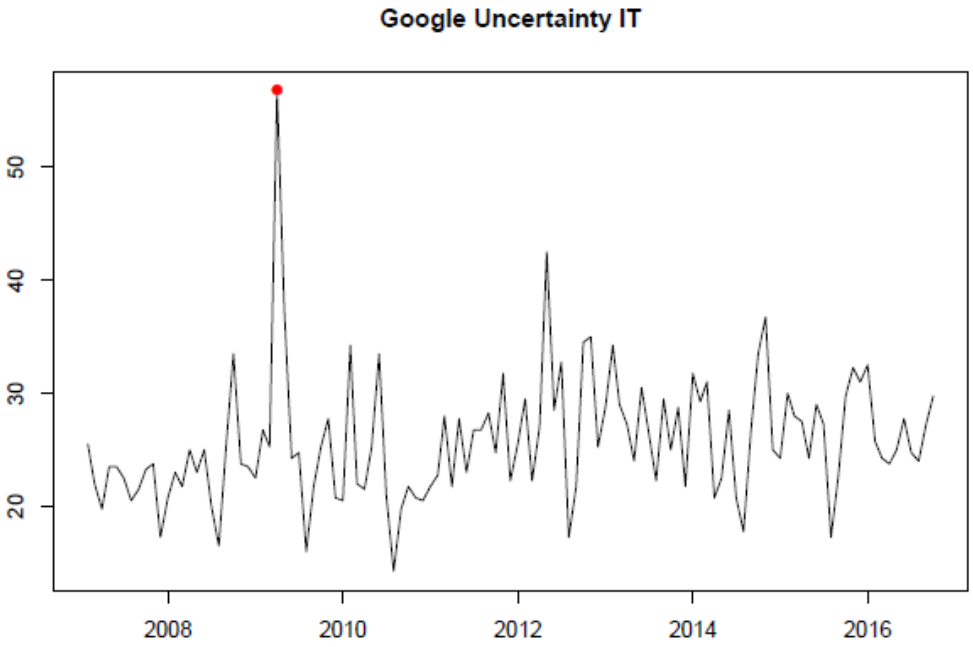

incertezza, web
45
40
35
30
25
20
15
2008
2010
2012
2014
2016

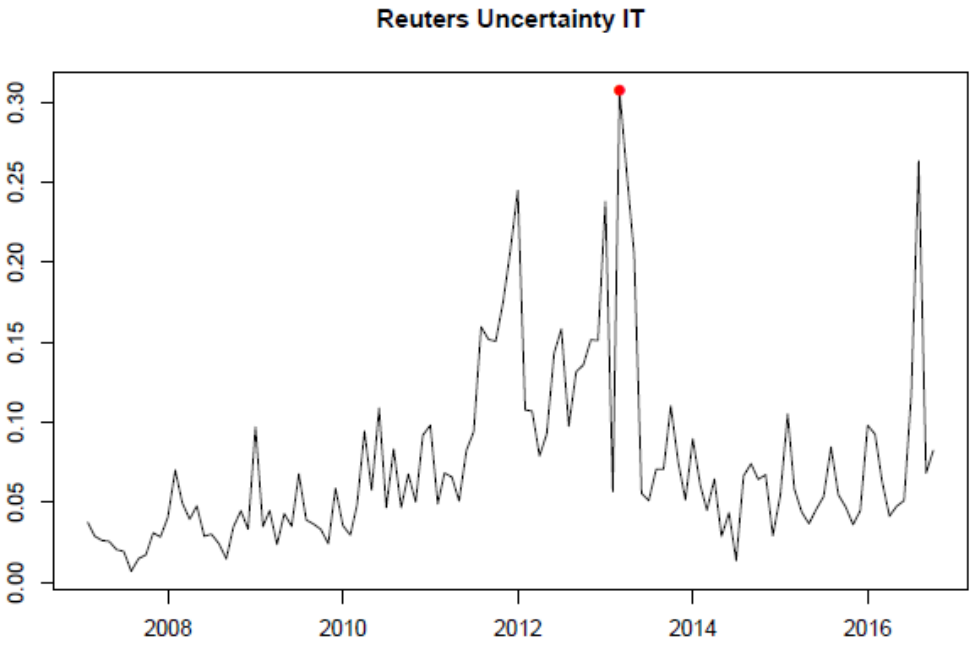

incertezza, news
50
40
30
20
10
0
2008
2010
2012
2014
2016

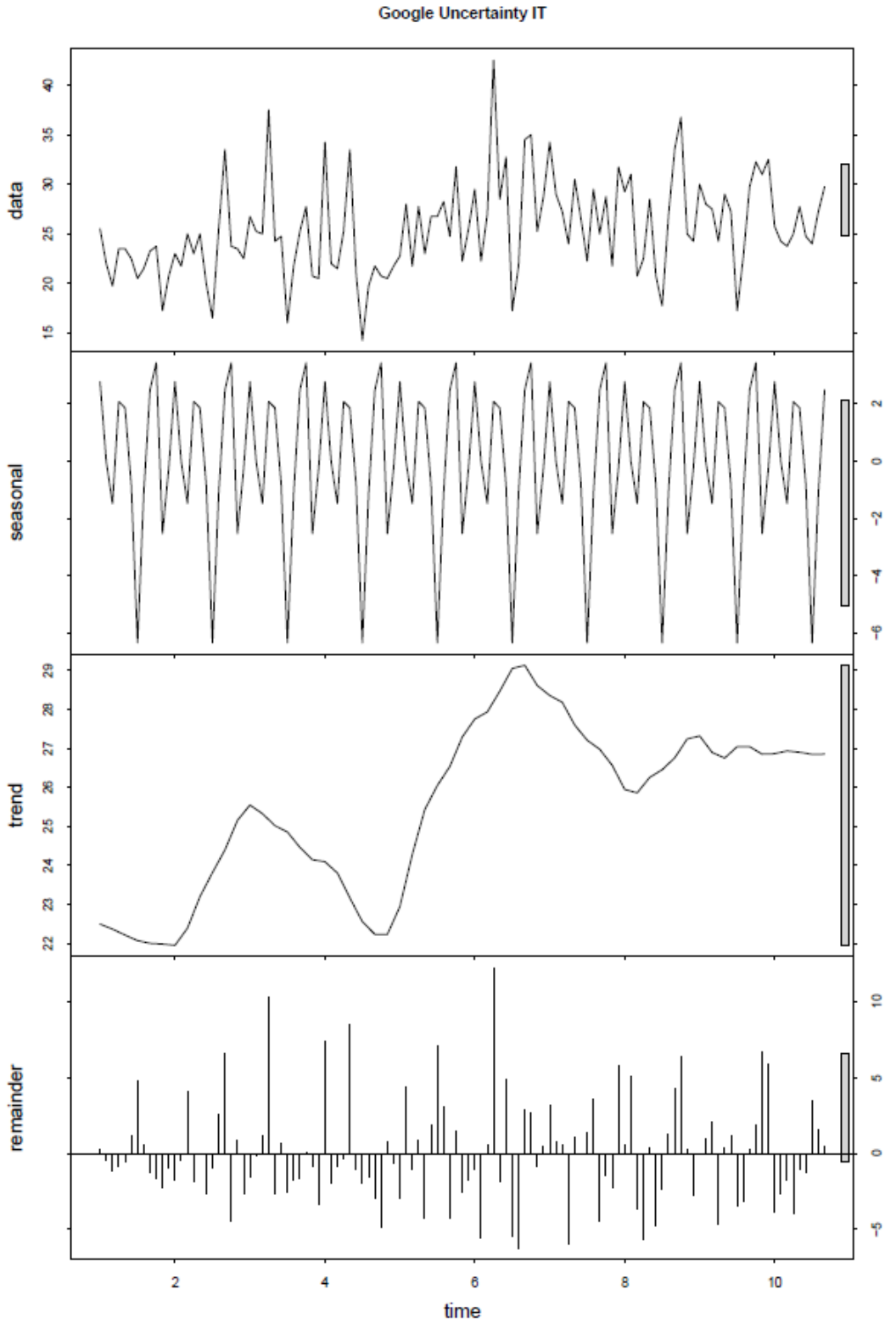

rischio, web
100
90
80
70
60
50
40
2008
2010
2012
2014
2016

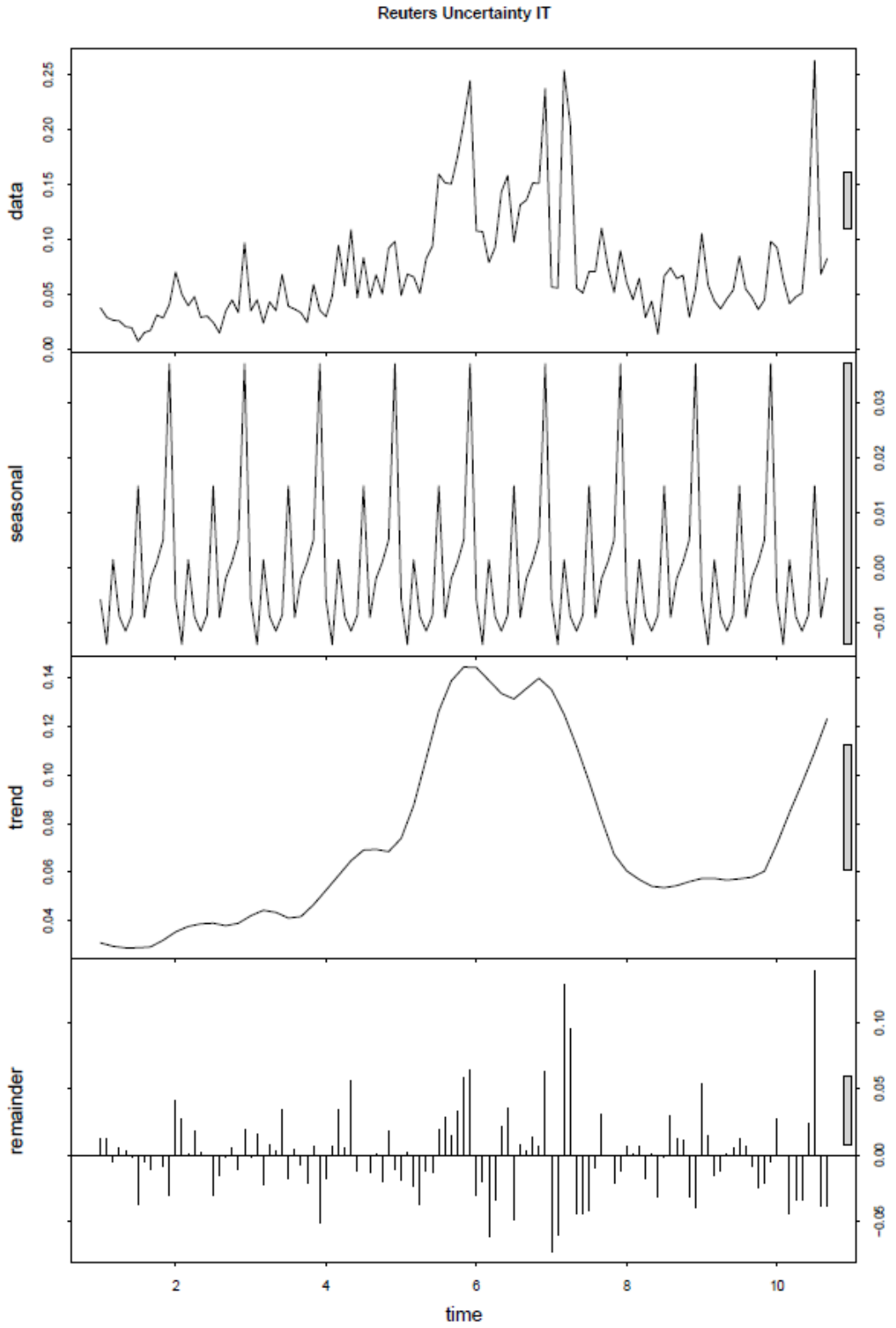

rischio, news
100
80
60
40
20
0
2008
2010
2012
2014
2016

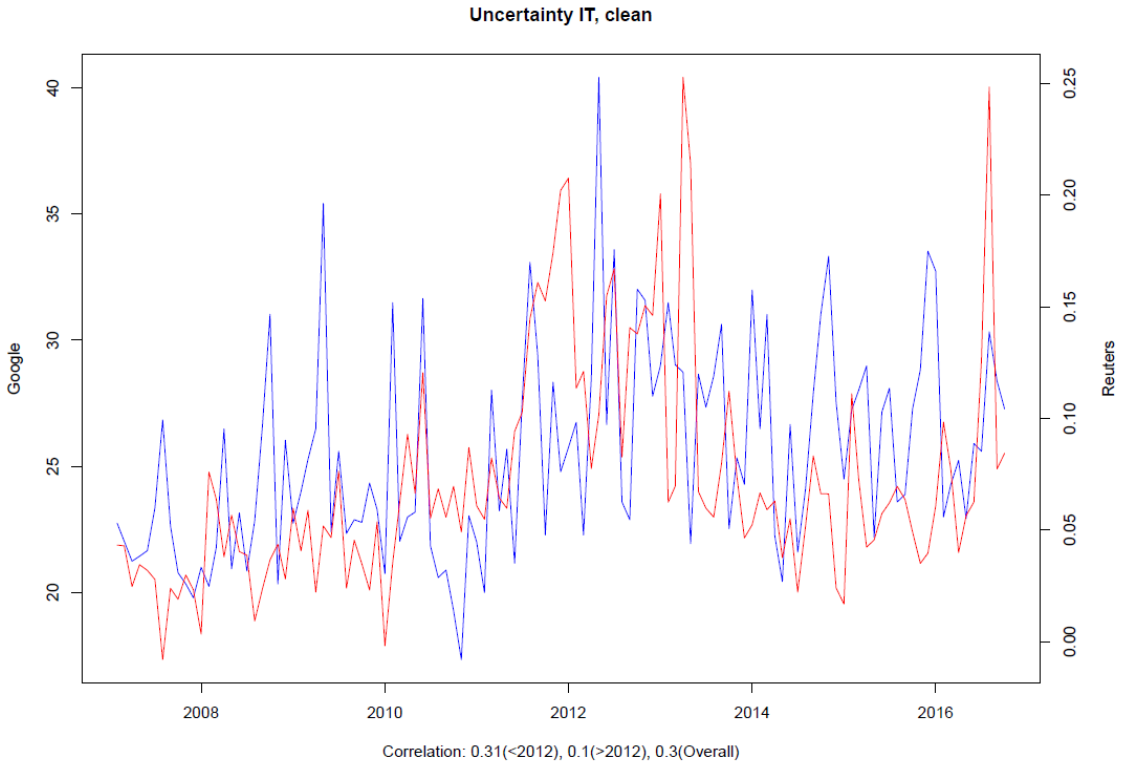

Uncertainty IT
Google
Reuters
50
40
30
20
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2008
2010
2012
2014
2016
Correlation: 0.05(<2012), 0.05(>2012), 0.13(Overall)


Google Uncertainty IT
50
40
30
20
2008
2010
2012
2014
2016

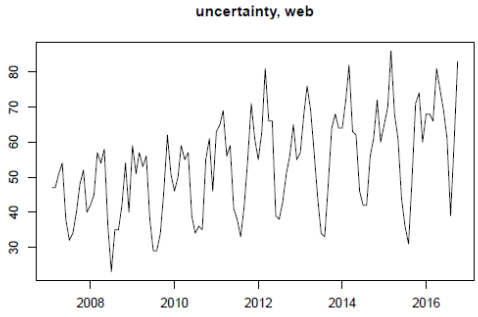

Reuters Uncertainty IT
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2008
2010
2012
2014
2016

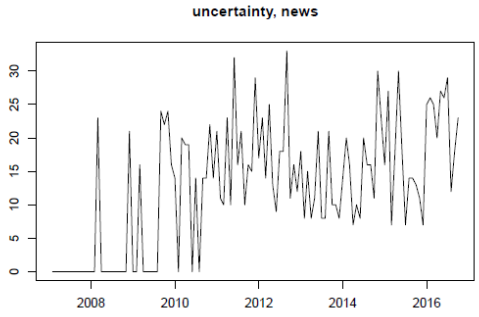

Google Uncertainty IT
data
seasonal
trend
remainder
time
2
4
6
8
10

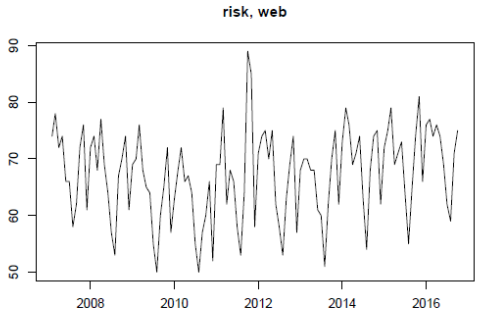

Reuters Uncertainty IT
data
seasonal
trend
remainder
time
2
4
6
8
10

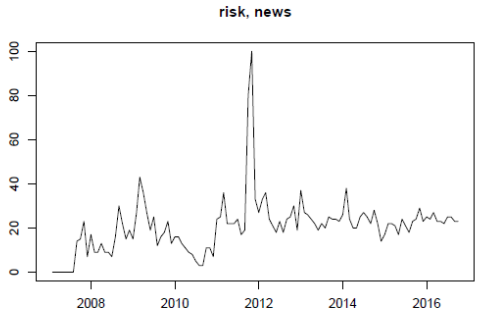

Uncertainty IT, clean
Google
Reuters
40
35
30
25
20
0.25
0.20
0.15
0.10
0.05
0.00
2008
2010
2012
2014
2016
Correlation: 0.31(<2012), 0.1(>2012), 0.3(Overall)

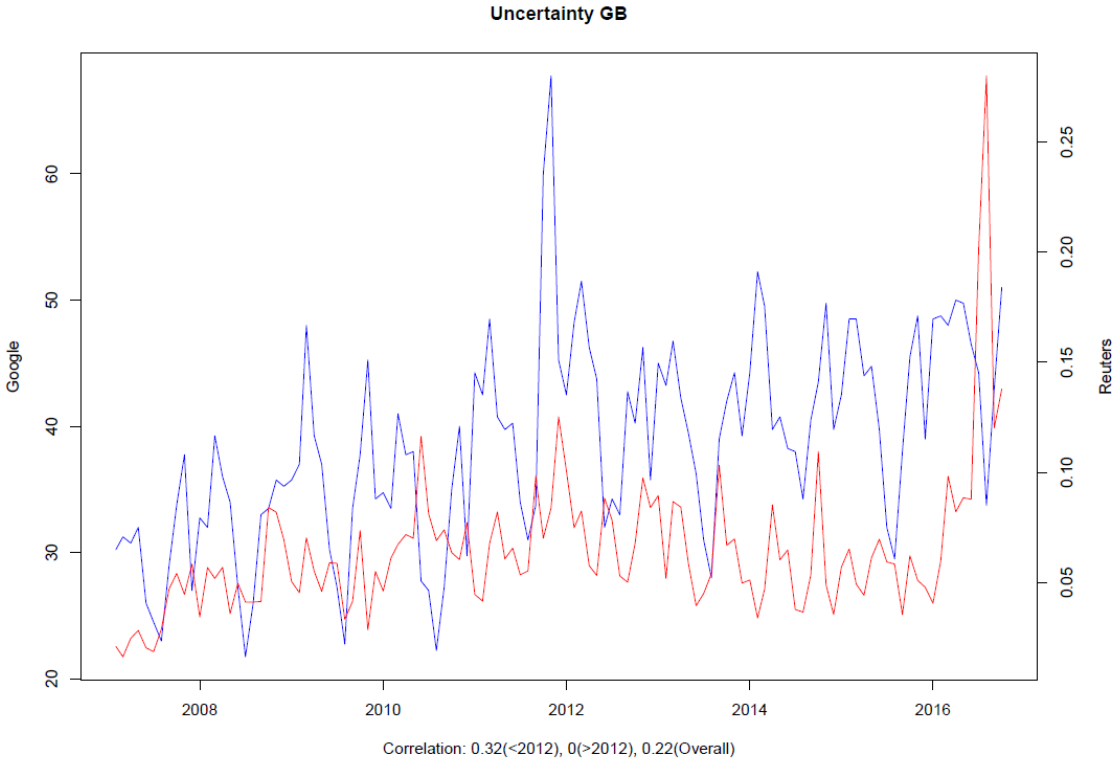
United Kingdom

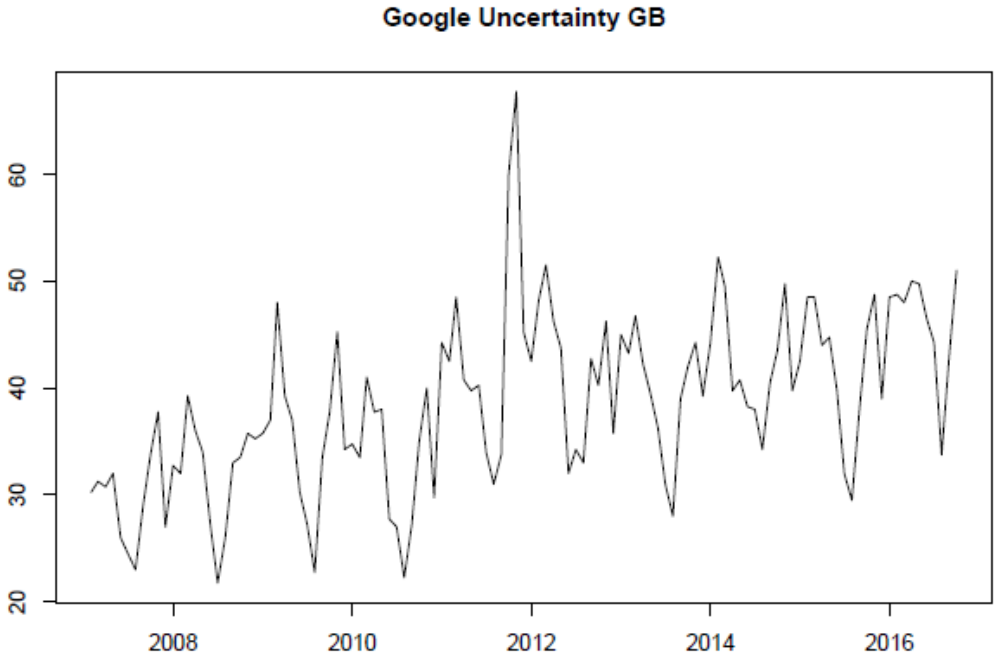


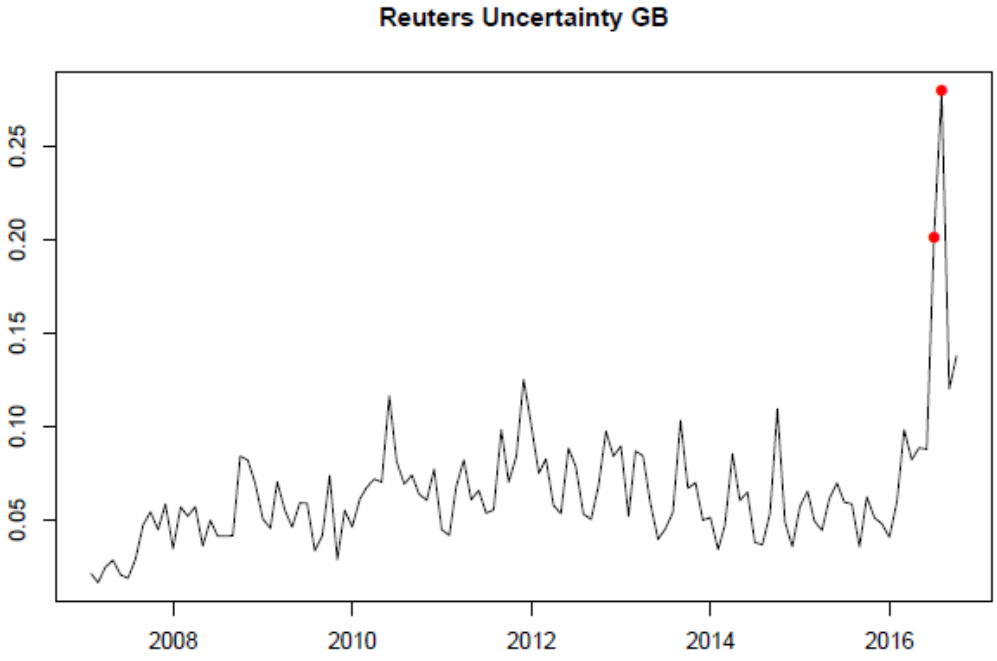


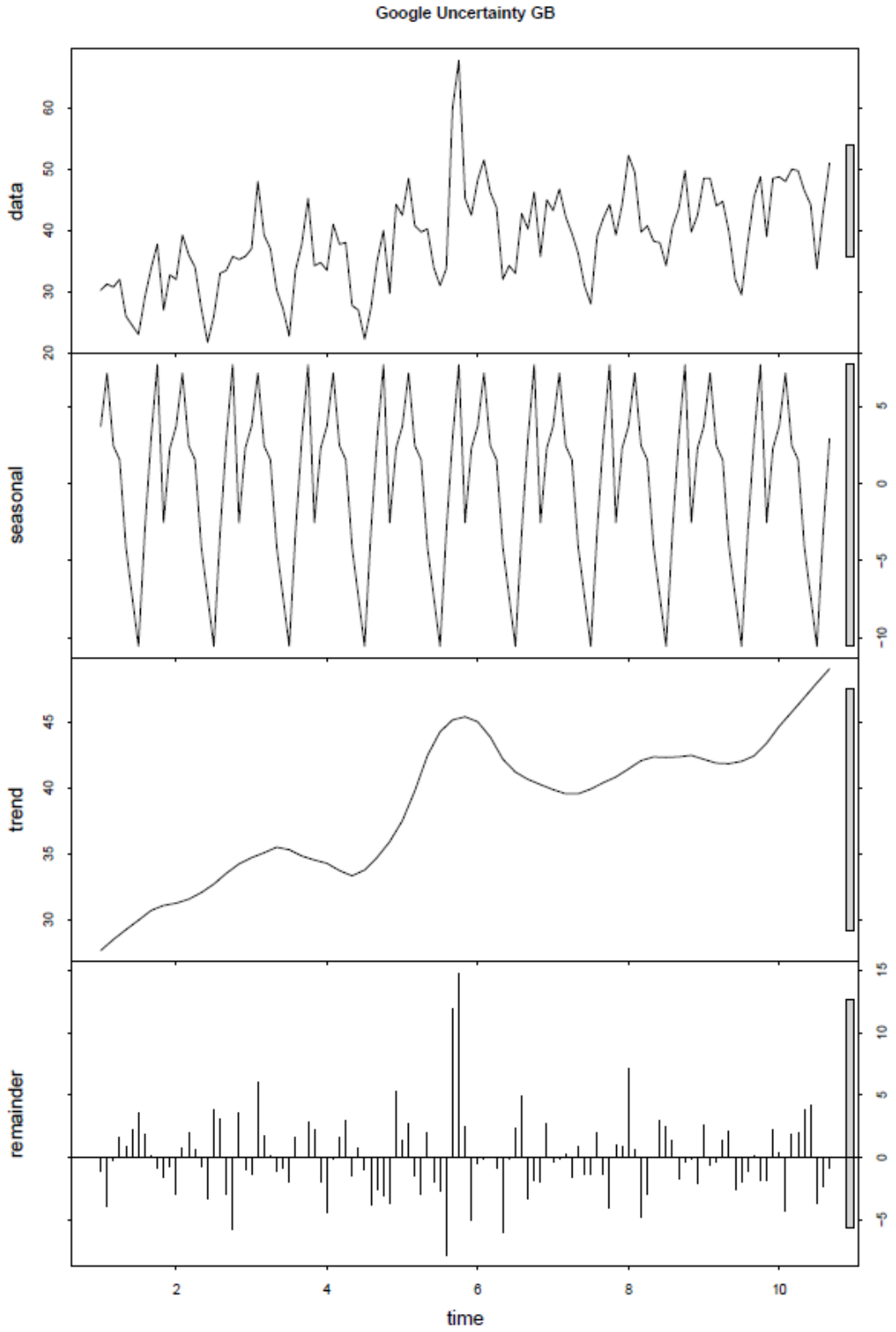


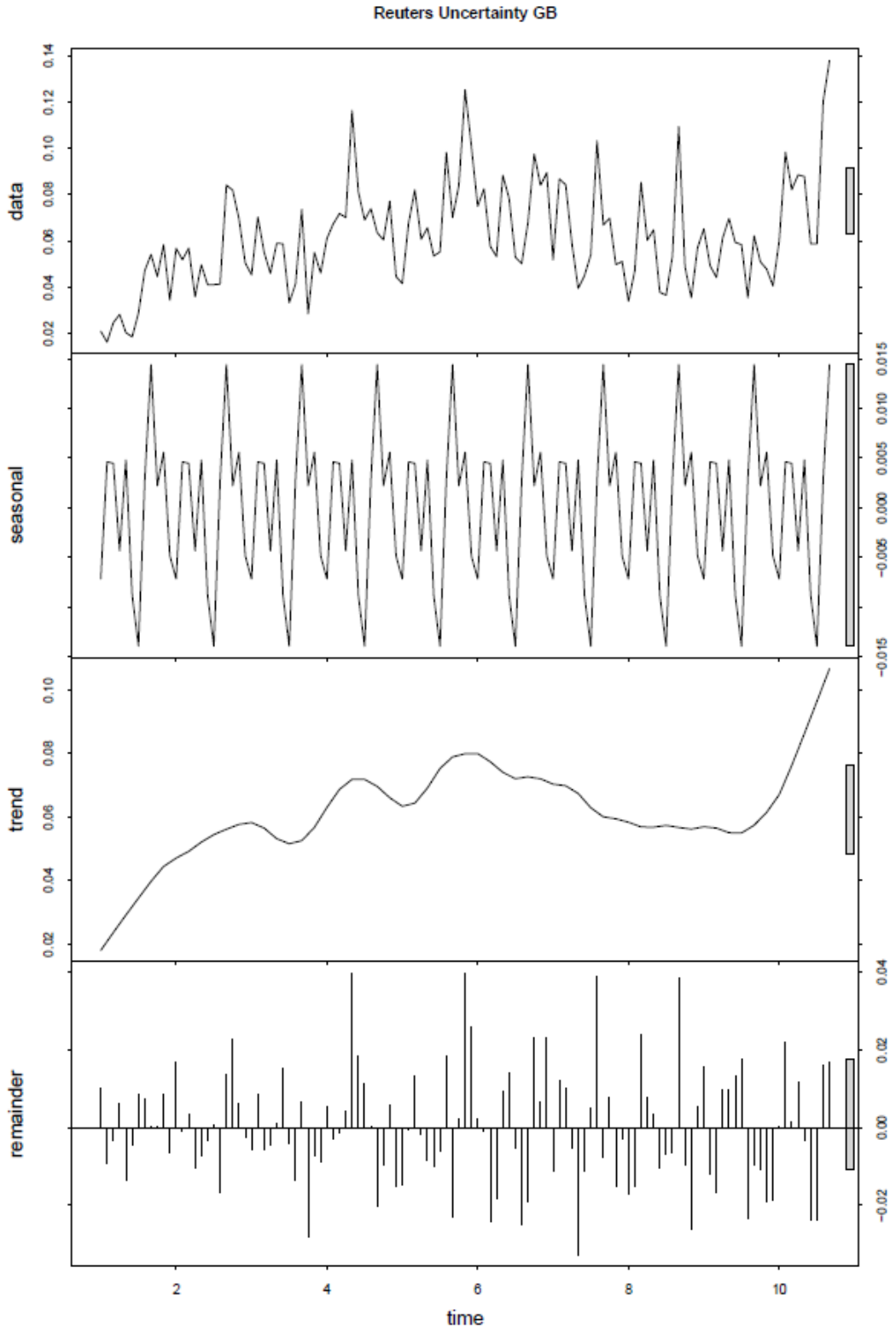


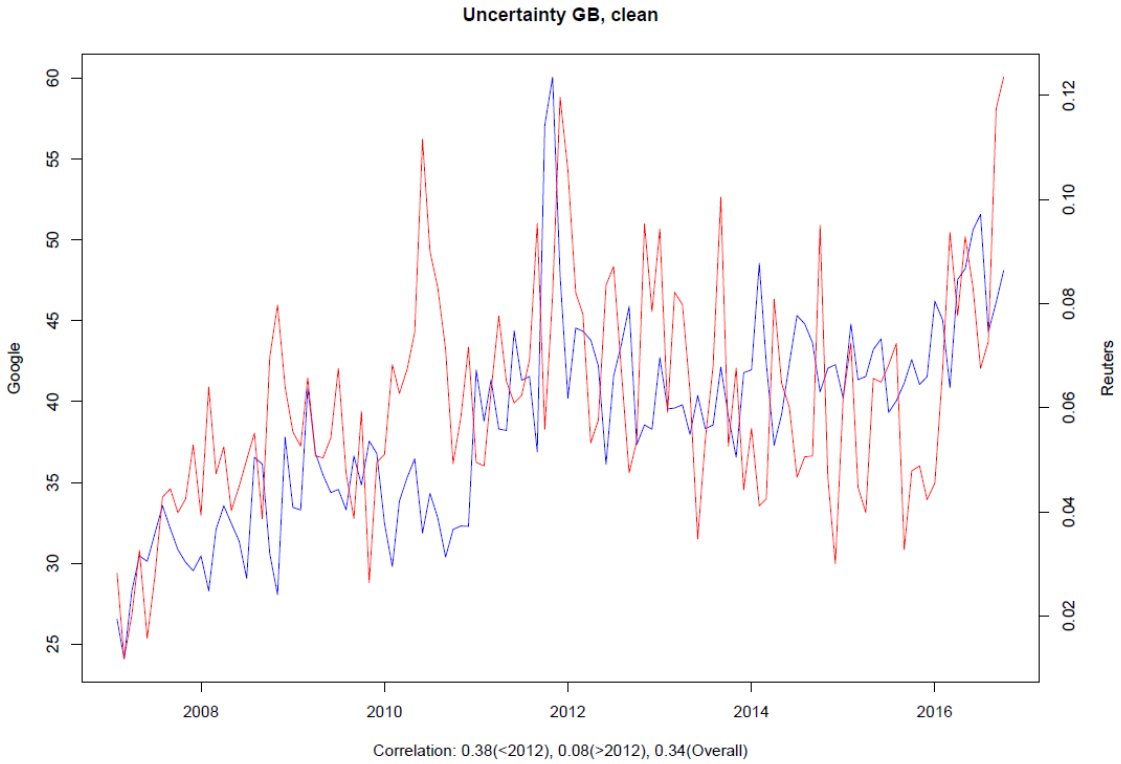



Google Uncertainty GB
20
30
40
50
60
2008
2010
2012
2014
2016


Reuters Uncertainty GB
0.05
0.10
0.15
0.20
0.25
2008
2010
2012
2014
2016


Google Uncertainty GB
data
seasonal
trend
remainder
time
2
4
6
8
10


Reuters Uncertainty GB
data
seasonal
trend
remainder
time
2
4
6
8
10


Uncertainty GB, clean
Google
Reuters
25
30
35
40
45
50
55
60
0.02
0.04
0.06
0.08
0.10
0.12
2008
2010
2012
2014
2016
Correlation: 0.38(<2012), 0.08(>2012), 0.34(Overall)

Getting in touch with the EU

In person
All over the European Union there are hundreds of Europe Direct Information Centres. You can find the address of the centre nearest you at: http://europa.eu/contact

On the phone or by e-mail
Europe Direct is a service that answers your questions about the European Union. You can contact this service
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by electronic mail via: http://europa.eu/contact

Finding information about the EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU Publications
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

EU law and related documents
For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU
The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

KS-TC-17-007-EN-N

Filtering techniques for big data and big data based uncertainty indexes

This work is concerned with the analysis of outliers detection, signal extraction and decomposition techniques related to big data. In the first part, also with the use of a numerical example, we investigate how the presence of outliers in the big unstructured data might affect the aggregated time series. Any outliers must be removed prior to the aggregation and the resulting time series should be checked further for outliers in the lower frequency. In the second part, we explore the issue of seasonality, also continuing the numerical example. Seasonal patterns are not easily identified in the high frequency series but are evident in the aggregated time series. Finally, we construct uncertainty indexes based on Google Trends and compare them to the corresponding Reuters-based indexes, also checking for outliers and seasonal components.

For more information
http://ec.europa.eu/eurostat/

Publications Office

ISBN 978-92-79-74460-0