Tourism Management 75 (2019) 1–12

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

Spurious patterns in Google Trends data - An analysis of the effects on tourism demand forecasting in Germany

Björn Bokelmann*, Stefan Lessmann

School of Business and Economics, Humboldt University of Berlin, Germany

ARTICLE INFO

Keywords:
Forecasting
Tourism demand
Big data analytics
Search query data
Google trends

ABSTRACT

Previous studies show that time series data about the frequency of hits for tourism-related search terms from Google (Google Trends data) is a valuable predictor for short-term tourism demand forecasting in many different tourism regions worldwide. The paper contributes to this literature in three ways. First, it shows that Google Trends data is useful for short-term predictions of monthly tourist arrivals in several German holiday regions. Second, the paper also demonstrates that the Google Trends time series we employ share certain patterns with Google Trends time series used in previous studies, including several studies totally unrelated to the tourism industry. We refer to these artefacts as “spurious patterns” and perform a detailed analysis of their negative impact on forecasting. Last, the paper proposes a method to sanitize Google Trends data and reduce the adverse impact of spurious patterns, thereby paving the way to develop statistically sound tourism demand forecasts.

1. Introduction

Along with the unprecedented growth of the global tourism industry in the last decades, tourism demand forecasting became an important research topic, reflected in the increasing number of articles published as Song and Li (2008) point out. Song and Li (2008) further note that the most commonly used measure of tourism demand is the number of arrivals at tourist accommodations, which is therefore used in this paper. Gunter and Önder (2015) state that the big importance of tourism demand forecasting comes from the perishable nature of tourism-related offerings: unsold hotel rooms or tourism services are lost revenue.

Tourism managers in business companies (Hotels and Travel Agencies), tourism associations and local authority organisations are confronted with the task of adapting the capacities necessary for their services to changing and in some cases fluctuating demand. Long-term decisions should be based on trend forecasts that take into account typical determinants of consumer decisions (demography, attitudes, income trends, etc.). In the short term, however, it is also possible and necessary to adapt some of the capacities (e.g. seasonal labour) flexibly to demand. In addition, demand can also be influenced in the short term by marketing measures, for example pricing policy (special offers). In order to be able to support such decisions, however, reliable forecasts about the bookings to be expected in the coming weeks and months are required. In this field, reliable predictions can improve tourism manager's decision making processes.

In their research review, Song and Li (2008) note that tourism demand forecasts are typically made by using either time series methods (forecasts based on the historical tourism demand data) or causal methods (multivariate forecasts using historical tourism demand data as well as certain economic variables). The authors conclude that tourism demand forecasting remains an open research field since there is no method which is generally considered most accurate.

In this paper, forecasts are made for monthly tourist arrivals (national as well as international) for several German tourism regions. Time series data about the frequency of hits for tourism-related search terms from Googles' service “Google Trends” is used as a predictor. In the following the abbreviation **GT** refers to Google Trends. Previous research shows that GT data can be used to forecast many economic variables like unemployment (Askitas & Zimmermann, 2009) and exchange rates (Bulut, 2018). Especially in tourism research, there has been a significant focus on GT data based predictions in the last years (Bangwayo-Skeete & Skeete, 2015; Park, Lee, & Song, 2017; Önder, 2017). However, little attention has been paid to a certain difficulty in the use of GT data for forecasting. There are a number of factors that might influence search behavior, which are unrelated to the target variable of the respective forecasting settings. The examples of section 2.2 demonstrate that GT time series share characteristic patterns even if the corresponding search queries are not thematically related. Lacking a topical relationship, it is implausible that search volume data would

* Corresponding author.
E-mail address: bjoern.bokelmann@gmx.de (B. Bokelmann).

https://doi.org/10.1016/j.tourman.2019.04.015
Received 28 September 2018; Received in revised form 11 February 2019; Accepted 23 April 2019
Available online 09 May 2019
0261-5177/

display such behavior. We refer to corresponding artefacts in GT time series as "spurious patterns" and hypothesize that they originate from a common factor that influences GT data. The existence of such a common factor – a determinant of GT data that is unrelated to the search query – would bias forecasts based on GT data and thus have far-reaching implications for forecasting based on this data in tourism management and beyond. The paper contributes to the literature through clarifying the origin of these spurious patterns, demonstrating how such artefacts impede forecast accuracy, and developing an approach to sanitize GT data. Another contribution comes from our empirical analysis that provides original insights on how GT data improves the accuracy of short-term predictions of monthly tourist arrivals in several holiday regions in Germany, and confirms the proposed data modification strategy to raise the predictive value of GT data.

2. Related work

2.1. Search engine data based tourism demand forecasting

There is a large number of studies about search engine data based tourism demand prediction. While GT data is used for tourism demand forecasting in most countries, studies on tourism demand forecasting in China are mainly based on data from the Baidu search engine. Due to the similarity of both data sources, statements about the prediction using one data source can usually be transferred to predictions using the other. For this reason, this section refers to "search engine data" in general, instead of distinguishing between GT data and Baidu data unless statements are made specifically for GT data only.

Yang, Pan, Evans, and Lv (2015b) suggest an algorithm to collect search terms related to a tourism region. Because there are often multiple possible search terms for one region, Li, Pan, Law, and Huang (2017) present a method for dimension reduction of the search engine data. Furthermore Zeynalov (2017) and Bangwayo-Skeete and Skeete (2015) evaluate the applicability of mixed data sampling in case search engine data and tourism demand data are given at different frequencies.

Nearly all studies use dynamic linear models (DLM's) or vector autoregressive models to make search engine data based predictions for tourism demand. The only exception is de Kort et al. (2017) who use machine learning methods. The value of search engine data for forecasting is mostly measured in comparison with the results of ARIMA or seasonal ARIMA (SARIMA) benchmark models.

Most of the search engine data based tourism demand forecasting studies make predictions for international tourism where they specify a target region and restrict their analysis to specific source regions. For example Park et al. (2017) consider tourist inflow from Japan to South Korea. In such a setting, search engine data from the source region is likely to represent the travel planning process, and lags of search engine data are likely to be correlated with the according tourism demand data. If someone in Japan, types "Seoul sights" he or she is likely to plan a trip to South Korea. Other examples of such international forecasting settings include Artola and Martínez-Galán (2012) who forecast numbers of British tourists in Spain, Gawlik, Kabaria, and Kaur (2011) who predict the amount of western tourists in Hong Kong and García Rodríguez (2017) who forecasts tourism demand for Mallorca coming from Germany and Britain. In contrast, this paper is concerned with tourism demand forecasting in Germany where there is an extraordinary big share of domestic tourism - in 2017 79% of the tourists in Germany were German citizens according to Destatis (2017). Therefore, predictions are not made for individual source regions but for domestic and international tourists summed up. This implies that the target region is part of the region from which the GT data is collected. As a consequence, travel related queries could be made from people planning to travel to a German tourism region or from people that have already arrived. Someone in Germany who types "Berlin Sights" could either live in a different part of Germany and plan to visit Berlin or could already stay in Berlin and search for activities.

This leads to a nowcasting setting, where contemporaneous and past GT data is combined with past tourism demand data to make predictions about the present. GT based nowcasting for tourism demand was already done by Choi and Varian (2012). In their study, the authors make predictions for month *t* based on past tourism demand data as well as GT data until the second week of month *t*. In contrast, in this paper aggregated GT data of the whole month *t* as well as past tourism demand data is used to make predictions for month *t*. According to Destatis (2018) the data of the German Federal Office of Statistics appears with a time lag of 6–8 weeks. Hence, the method yields estimates 2–4 weeks before the official numbers arrive.

To give an example how such estimates could be beneficial for decision making, consider the case of a hotel owner in one of the German tourism regions. If this owner has a bad booking situation in month *t* he would need to wait 6–8 weeks for the official numbers to assess whether there is currently an overall low tourism demand in the region or whether only his hotel is affected. With the use of tourist arrival estimates 2–4 weeks before the official numbers, this hotel owner can better adapt to the situation (e.g. change pricing).

The main focus of this paper is on the impact of the spurious patterns on the relationship between GT data and the target variable. It will be argued that these patterns are due to a certain "disturbance" of this relationship. For this research it is not important whether the target variable is related to lagged GT data (forecasting) or contemporaneous GT data (nowcasting). Therefore the term forecasting will be used in both cases unless a distinction is important for the understanding.

With regard to the prediction setting, this paper distinguishes itself from prior literature by the high number of tourism time series used as well as by forecasting domestic and international tourism demand combined. In previous literature, only Yang et al. (2015b) and Li et al. (2017) forecast domestic tourism. But since domestic tourism in China could require flights of several hours to get to the destination, the consideration of domestic tourism in Germany is fundamentally different from both studies.

However, the key difference to any other GT based forecasting study is the analysis of the aforementioned spurious patterns of GT data and their impact on forecasts.

2.2. Spurious patterns in GT data

For each search term s, GT provides search index data in time series format. The size of the index indicates how many search queries were made for the terms s for each point in time. Regardless of which terms the GT data is collected for, it is striking that most time series show downward trends in the first years after 2004 and breaks in 2011 and 2016. We refer to these artefacts as spurious patterns. Some examples of visible spurious patterns in previous GT studies are given below.

Wu and Brynjolfsson (2015) among other sources, use GT data for the category "real estate" in the USA (Fig. 1). They assume that the initial downward trend and the following rise of the GT data are related to the peak of the real estate bubble in 2005 and the following recession until the recuperating of housing sales after 2011. Section 4.1.2 of this paper provides an alternative explanation.

Mccallum and Bury (2013) argue that the decline in the GT index for several environment-related terms indicates an eroding public interest in the environment. However, Ficetola (2013) responds that the declining index for these search terms is not necessarily a sign of decreasing interest and shows that there is a similar downward trend for several different unrelated search terms. Fig. 2 shows the GT index for one of the search terms used in both studies.

Önder (2017) uses GT data for the terms "Vienna" and "Barcelona" as well as for "Austria" and "Belgium" to predict tourism numbers in these regions. In the GT data for "Austria" in Fig. 3 there is an apparent downward trend in the first years which is not reflected in the tourism demand data in this study.

Vosen and Schmidt (2011) use several GT categories to forecast

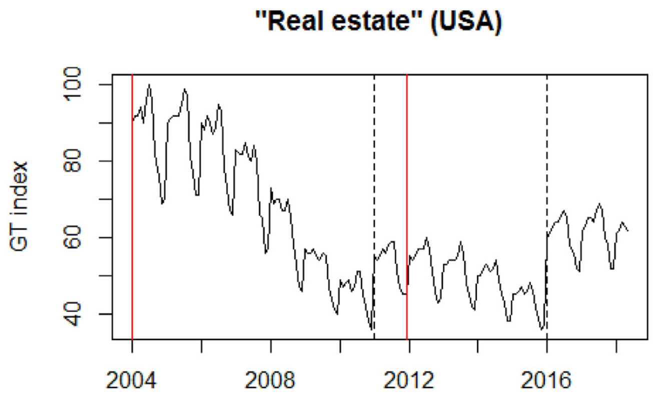

Fig. 1. GT index for the category "real estate" in the USA. The red lines represent the time span for which data was used by Wu and Brynjolfsson (2015). The dashed lines mark January 2011 and January 2016 when the data recording of GT changed. Note that Wu and Brynjolfsson (2015) used quarterly aggregates of this data. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

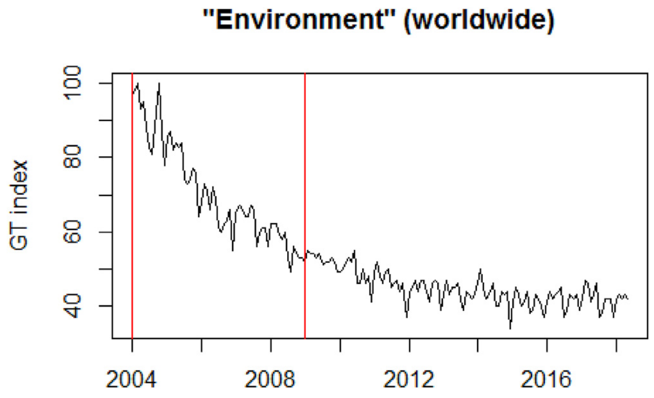

Fig. 2. GT index for the search term "environment" worldwide. The red lines represent the time span for which data was used by Ficetola (2013). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

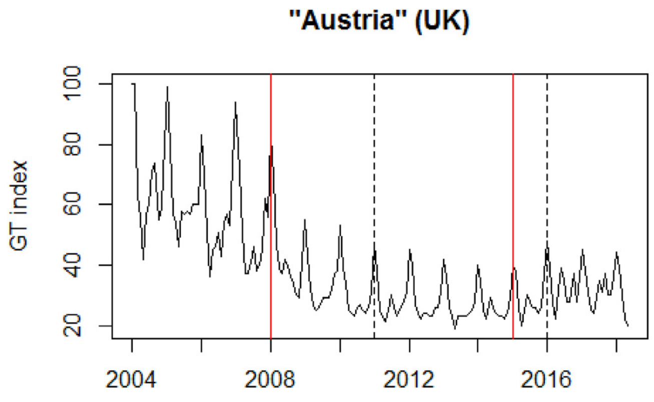

Fig. 3. GT index for the search term "Austria" in the UK. The red lines represent the time span for which data was used by Önder (2017). The dashed lines mark January 2011 and January 2016 when the data recording of GT changed. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

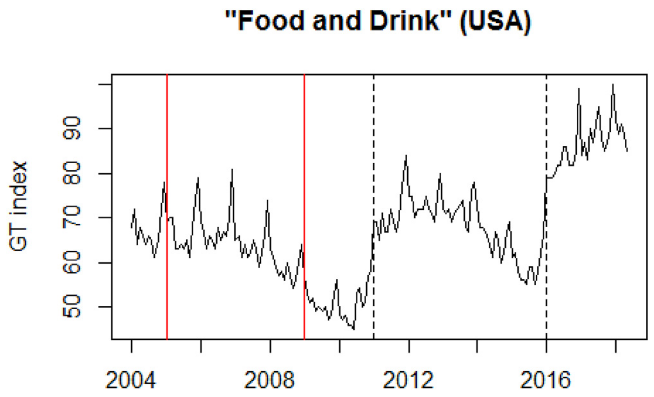

Fig. 4. GT index for the category "Food and Drink" in the USA. The red lines represent the time span for which data was used by Vosen and Schmidt (2011). The dashed lines mark January 2011 and January 2016 when the data recording of GT changed. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

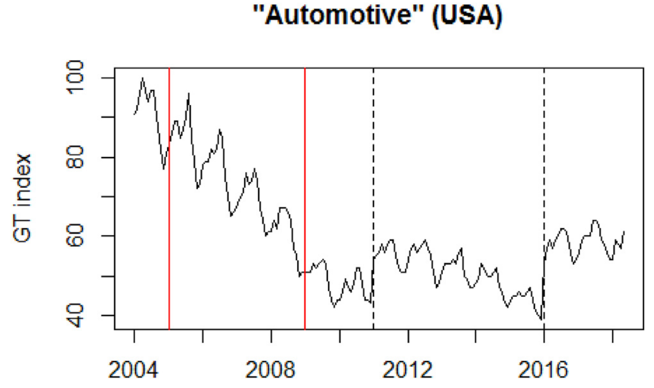

Fig. 5. GT index for the category "Automotive" in the USA. The red lines represent the time span for which data was used by Vosen and Schmidt (2011). The dashed lines mark January 2011 and January 2016 when the data recording of GT changed. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

private consumption in the USA. Figs. 4 and 5 show the GT index for two of these categories. In both graphs there are apparent breaks in 2011 and 2016. For the category "Automotive" a downward trend in the first years is discernible. For "Food and Drink" there is also a continuous downward trend in the first years which is only interrupted by the sudden shifts in 2011 and 2016.

There is no literature aiming at a detailed analysis of the spurious patterns in GT data. However, the possibility that long-term trends of GT time series for specific search terms could be due to a change in the search volume of the entirety of all search terms is already discussed by Stephens-Davidowitz and Varian (2014). This is possible since the GT data is an index representing the share of the search volume for a specific term compared to the total search volume of all search terms combined (see section 4.1).

After the apparent failure of the famous Google Flu Trends (a GT based influenza forecasting model), literature analyzing reasons for that failure appeared. Lazer, Kennedy, King, and Vespignani (2014) argue that spurious regression and a change in search behavior are among those reasons. Santillana, Zhang, Althouse, and Ayers (2014) and Yang, Santillana, and Kou (2015a) present continuously updating models for influenza forecasting in order to overcome the problem of the changing relationship between GT data and the target variable.

In this paper, a method to modify GT data in order to reduce the (negative) impacts of the spurious patterns on forecasting is suggested. Each individual time series is divided by the average trend of GT time series for multiple queries from the same region. This is done to reduce the influence of the total search volume in the index described above. This method is related to the idea to analyze the trend of query data not individually but in comparison with several independent queries as suggested by Ficetola (2013). It is also related to the discussion by

Siliverstovs and Wochner (2018), that division of GT time series from the same region cancels out the denominator in the GT index representation and thus potentially unwanted factors on the forecasts.

3. Objectives of this study

The objectives of this paper are twofold: On the one hand, the aim is to evaluate whether GT data is a useful predictor for tourism within Germany. In this context, it shall be found out whether the spurious patterns have a (negative) impact on forecasting accuracy and whether it is possible to reduce this impact to generate more accurate forecasts.

On the other hand, a detailed analysis of the spurious patterns in GT data is made with the goal of a better understanding of their origins. This is potentially more far-reaching than the first objective since these spurious patterns seem to be an issue of GT data regardless of the actual application.

With regard to the aforementioned objectives, this paper answers two research questions:

(1) Does the use of GT data lead to improved forecasts compared to predictions made solely on historical tourist data?
(2) Do the spurious patterns have a negative impact on GT based forecasts and does GT data modification help to reduce this impact?

Finally, the paper aims to quantify the effect of factors causing the spurious patterns in the GT data. Therefore, in Appendix A an idea for a quantitative measure is given and the effects of data modification are evaluated.

4. Data

4.1. GT data

The goal was to enable GT data based forecasts for as many German tourism regions as possible. Therefore, a large number of search terms needed to be collected. Research about peoples use of search engines for holiday planning reveals that location-specific queries are most important (Xiang, Wang, O'Leary, & Fesenmaier, 2015) and that many tourism-related search terms are combinations of a location name and an expression of the specific travelers needs (e.g. “Berlin Hotel”) (Xiang & Pan, 2011).

To find search terms related to the planning of a holiday in Germany, the names of the top listed tourist destinations from two web pages ((GNTB, 2018) and (Tripadvisor, 2018)) were collected. In addition, the names of German rivers, lakes and mountains that were recommended by Google when entering the search terms “Deutschland Flüsse” (Germany rivers), “Deutschland Seen” (Germany lakes), “Deutschland Berge” (Germany mountains) were added. Finally, the names of all tourism regions specified by the German Federal Statistics Office complete the list of potential search terms. This made a total number of 269 search terms.

For some of the sights and locations chosen as search terms, there are words in other languages. But since most of the tourists in Germany have German mother tongue, it was chosen to collect the search terms in German.

For each search term, GT data between January 2008 and August 2017 was collected. GT data was taken from within Germany as well as from the whole world. This made two different data sets. For this paper, it is instructive to compare both data sets because they show different spurious patterns (see Figs. 6 and 7).

Choi and Varian (2012) state that because of Google Trend's “broad matching”, the resulting GT index for a search term measures the number of hits for that search term alone as well as in combination with other terms. E.g. entering “Berlin” leads to an index for “Berlin”, “Berlin Hotel”, “Berlin Airport”, “Berlin History” and so on. For more dignified results, GT offers the option to search within categories. According to Choi and Varian (2012) GT then calculates a probability for each search term to be related to the specified category and counts the hits according to that probability. E.g. “Berlin Hotel” and “Berlin Airport” are with high probability related to traveling, while “Berlin History” could also be searched for reasons unrelated to traveling. Therefore, only parts of the hits for “Berlin History” would be counted by the index. For the GT data that is used for this paper, the category “travel” was chosen. The data was downloaded using the *gtrends* function provided by Massicotte and Eddelbuettel (2017).

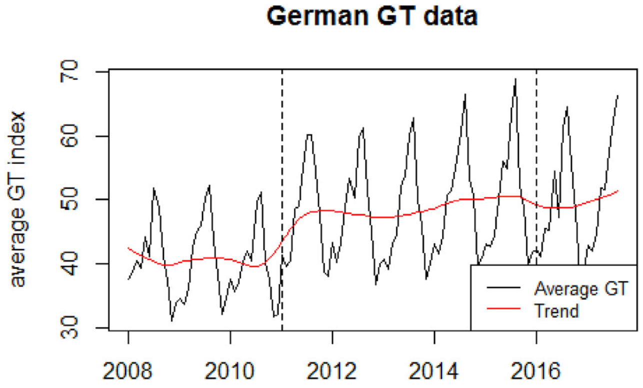

Fig. 6. Average of the 269 GT time series from Germany. The first dashed line correspond to Google's improvement of the localization of search volume. The second dashed line corresponds to the change of Google's data recording process.

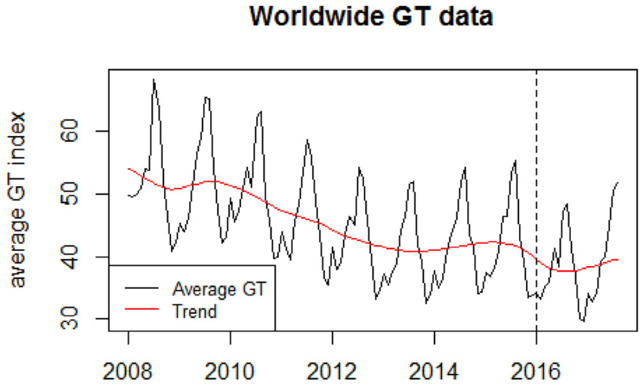

Fig. 7. Average of the 269 worldwide GT time series. The dashed line corresponds to the update in Google's data recording process.

4.1.1. Spurious patterns

Many of the GT time series from Germany showed a sudden break in January 2011 and January 2016. According to Google (2018a), at both times there were updates of Google's data recording process. Fig. 6 shows that both breaks are clearly visible in the average of all time series.

The worldwide GT data did not show any break in 2011, but many of these series had a significant downward trend until 2013 and also a break in January 2016. According to Google (2018a) the change in data recording in 2011 was only due to an improvement of the localization of search volume. Therefore, it does not affect the worldwide search volume recording since the collection of data is not restricted to a specific region. Fig. 7 shows that the downward trend and the shift are clearly visible in the average of all time series.

The difference between Figs. 6 and 7 is remarkable for the following reason: Both graphs represent search volume for the same queries and because the biggest share of tourism in Germany is domestic (79% in 2017 according to Destatis (2017)), it is very unlikely that the volume of tourism-related queries originating from outside Germany can explain such big differences. More plausible is that the difference between both graphs is due to factors unrelated to the actual search volume.

In the context of forecasting, a question about these spurious patterns is whether they are related to the respective target variable and therefore informative for forecasting or whether there is another explanation. The fact that several different GT time series share common patterns suggests that there is a hidden relationship between these time series. A plausible explanation for such hidden relationships can be found in the GT index representation.

According to Google (2018b) GT data is generated in the following way: Given a query i, a geographical region r, a frequency[1] and a time period t, an index is calculated. For each time point, the number $q_{i,r,t}$ of hits matching the query is divided by the total number of searches $Q_{r,t}$ within the region. The resulting time series is multiplied by a constant c_i which forces the maximum of the series to be 100.

$$g_{i,r,t} = c_i \cdot \frac{q_{i,r,t}}{Q_{r,t}} \tag{1}$$

However, as stated in Google (2018c) the results are only estimates from Google using a sampling method, which is why they depend on the day the data is collected. To account for this fact, each GT time series for the empirical study of this paper was collected at five different times, and the average was taken as suggested by Stephens-Davidowitz and Varian (2014).

As stated by Stephens-Davidowitz and Varian (2014) the index for a query is set to zero, if the number of hits matching that query is below a certain threshold. To account for this fact, equation (1) would need to be adjusted. But because (1) is only used to illustrate that GT time series from the same region share the same denominator this simplified representation is used.

4.1.2. Common disturbance factors

Similar downward trends of different GT time series from the same region can be explained by an increase of the denominator in equation (1) while there is no similar increase in the numerator. This could either be due to a change of the user population or a change of the populations search behavior. Ficetola (2013) argues that in the past mostly academics and computer scientists used the internet, while now it is adopted in the whole population of most countries.

With regard to the data collected, a plausible explanation for the downward trend in the worldwide GT data (see Fig. 7) would be that Google's user population outside Germany has grown at a higher rate than inside Germany. Since it seems likely that the majority of search volume for the chosen queries originates from within Germany, such difference in growth of the user population would lead to a faster increase of the denominator in equation (1) than in the numerator and hence cause a downward trend. This would also explain why there is no similar downward trend for the German GT data (see Fig. 6).

According to Google (2018a) in January 2011 Google's collection of regional search volume data changed and in January 2016 Google changed the general data recording process. This is a plausible explanation for the breaks in 2011 and 2016. Considering representation (1) it could be argued that these changes in data recording affect the numerator as well as the denominator for each GT time series.

For GT data $g_{i,r,t}$ which is used to forecast a target variable y_t, factors that are unrelated to y_t but affect the denominator of representation (1) will be denoted **common disturbance factors** in this paper. It will be argued that the spurious patterns (downward trend and breaks) are, to a large extent, caused by common disturbance factors.

In the context of predictive modeling the phenomenon of these common patterns indicates the possibility of concept drift: i.e. the relationship between input and target variables changes with time. A temporary downward trend or shift in the input variable, which is unrelated to the target variable, could be a cause for such concept drift. This has two implications:

On the one hand, this raises doubt about the reliability of predictive models based on the GT data. Even if the models prove to be accurate on a short test sample, the possibility of concept drift in the long run should be taken into consideration when assessing their usefulness.

On the other hand, the fact that several different GT time series share similar patterns and the possibility that representation (1) yields a mathematical explanation for these patterns could lead to a better understanding of them. Furthermore, it might help to reduce the impact of such factors and, as a consequence, reduce the problem of concept drift.

The main innovation of the paper lies in a firm analysis of the spurious patterns (downward trends and breaks in 2011 and 2016) and the collection of empirical evidence that their negative impact on forecasts can be reduced by GT data modification.

It could be argued that the problems resulting from the spurious patterns are not relevant for the future use of GT since the spurious patterns are only visible until 2016. But in regards of the technological development it seems unlikely not to expect dramatic changes in populations search behavior in the following years and as the example of the downward trend shows, such changes influence any individual GT time series. Therefore, the greatest value of the paper should perhaps be seen in the contribution to a better understanding of the mechanisms that influence GT data. Such understanding could be helpful for the future use of GT. Since previous research has already shown that GT data can be a valuable predictor of tourism demand, a better understanding of the potential problems and how to deal with them can lead to more reliable and accurate predictions.

4.2. Tourism demand data

As noted by Song and Li (2008) tourist arrivals are the most widely used measure for tourism demand in the forecasting literature and were therefore chosen as the target variable in this study. The monthly tourist arrival data[2] in the period between January 2009 and August 2017 was collected from the Federal Statistical Office of Germany.

For a meaningful comparison of methods, forecasts should be made on as many arrival time series as possible. The German tourist arrival data is disaggregated into more than 100 distinct tourism regions. Each region for which appropriate query data was found (see 5.2.1) was included into the forecasting competition.

Data until August 2015 was used as a training sample, the remaining two years' data constituted the test sample.

5. Empirical methodology

5.1. GT data modification

Given GT time series $g_{1,r,t}, \ldots, g_{N,r,t}$ for the same region, it is possible to remove the influence of the general search behavior in that region, which is contained in the denominator of the index representation, via division by the average of all GT time series:

$$\bar{g}_{i,r,t} = \frac{g_{i,r,t}}{\frac{1}{N}\sum_j g_{j,r,t}} \tag{2}$$

[1] The user has to specify a time period from which the GT data should be collected. If the period is longer than 5 years, monthly data is provided. For shorter periods weekly data is supplied.

[2] The data was not available in time series form. Instead, the respective version of the monthly survey with the arrival numbers of each tourism region had to be downloaded for each month between 2009 and 2017. These numbers were joined afterwards according to the name of each tourism region. Only regions with complete data between 2009 and 2017 were included in the study.

$$= \frac{c_i \frac{q_{i,r,t}}{Q_{r,t}}}{\frac{1}{N}\sum_j c_j \frac{q_{j,r,t}}{Q_{r,t}}} \tag{3}$$

$$= \frac{c_i q_{i,r,t}}{\frac{1}{N}\sum_j c_j q_{j,r,t}} \tag{4}$$

The possibility to eliminate the denominator $Q_{r,t}$ in such a way is already discussed by Siliverstovs and Wochner (2018). The resulting modified data is called **average divided** in the following.

Average division replaces the denominator $Q_{r,t}$ with $\frac{1}{N}\sum_j c_j q_{j,r,t}$. Even though this removes undesired influences from $Q_{r,t}$ on the index, it is doubtful whether the new denominator has better properties than the original one. Considering the case where $g_{1,r,t}$, ...,$g_{N,r,t}$ are tourism related, three potential problems arise in representation (4).

(1) The time series $\frac{1}{N}\sum_j c_j q_{j,r,t}$ has a strong seasonal variation as a result of the idiosyncratic seasonal variations of each individual time series. This common seasonality would be included in each individual time series $\bar{g}_{i,r,t}$.
(2) Due to the reduced number of query time series in the new denominator, there is an increased impact of outliers from the individual time series $q_{j,r,t}$.
(3) There could be an overall increase/decrease in $g_{1,r,t}$, …, $g_{N,r,t}$ due to overall tourism developments. Dividing by $\frac{1}{N}\sum_j g_{j,r,t}$ would remove this information.

To overcome issues (1) and (2) the individual GT time series $g_{1,r,t}$, …, $g_{N,r,t}$ could be divided by the trend of $\frac{1}{N}\sum_j g_{j,r,t}$ instead of average division. The trend was calculated by seasonal and trend decomposition using Loess as described by Cleveland, Cleveland, and Terpenning (1990).

GT time series modified in this way can be represented by

$$\tilde{g}_{i,r,t} = \frac{\frac{c_i q_{i,r,t}}{Q_{r,t}}}{F\left(\frac{1}{269}\sum_j c_j \frac{q_{j,r,t}}{Q_{r,t}}\right)_t} \tag{5}$$

where F is an operator that extracts the trend of a time series. GT time series modified by the division of the average trend are called **trend divided** in the following.

The main goal of this paper is to compare the original GT data with modified GT data regarding the informational value for predictions. Therefore, we applied the above described modification methods to the German GT data as well as to the worldwide GT data. This made three versions of German GT data and three versions of worldwide GT data for each keyword: original, average divided and trend divided.

5.2. Tourism demand forecasting competition

In section 4 it was shown that the GT data collected shows spurious patterns and it was argued that these patterns originate from common disturbance factors. As a possible consequence, the occurrence of concept drift for forecasting was mentioned.

In this chapter, the idea that common disturbance factors might cause concept drift and that GT data modification as described in 5.1 could help to reduce this effect is taken up. The comparison of the out-of-sample fit between models using the original GT data and models using modified GT data is a way to evaluate the effect of the data modification. If forecasts using the modified GT data are superior to the ones made on the original GT data, this provides evidence for the assumption of common disturbance factors and concept drift.

As stated in section 2.1 forecasts are made for tourist arrivals in month t by using past arrival data until month $t-1$ combined with GT data up to month t. Forecasts for a longer horizon are not made.

To generate meaningful results, forecasts needed to be made on a large number of datasets. Since comparisons involving GT based forecasts is only possible on regions for which appropriate GT data was found, the effort was taken to collect as many tourism-related search terms as possible. Due to the large number of data sets finally included in the empirical study, the comparison of forecasts should be classified as a forecasting competition according to Chatfield (2000, p. 157–158).

5.2.1. GT data selection

In section 4.1 it was described how 269 search terms were selected. But it was not discussed whether the respective GT data is suitable to make predictions. The value of GT data for forecasting can only be evaluated for appropriately chosen queries. Otherwise, the "garbage in, garbage out" principle would make comparisons with pure time series methods pointless.

Whether a GT time series is suitable as a predictor can best be evaluated in the context of building dynamic linear models. In Appendix B pre-whitened cross-correlation is described as a method to identify the lag order of the input variable in a dynamic linear model. The key idea of pre-whitened cross-correlation is to measure the linear relationship between (lags of) the input variable and the output variable, conditional on the historical data of the output variable. According to Hamilton (1994, p. 557–562) this helps to overcome the problem of spurious regression.

The strongest relationship between GT data and the arrival data was expected to be contemporaneous. Siliverstovs and Wochner (2018) argue that the travel planning process involves multiple stages and that different tourists execute the first stages at different times. The last stage involves the acquisition of location-specific information and is performed by the majority of tourists in a short period prior to departure. This is why the strongest information of location-specific GT queries on monthly arrivals is contemporaneous. Visual comparison of the GT data collected with the arrival data confirmed this.

Since it was argued that the linear relationship between the GT data and the arrival data could be evaluated by pre-whitened cross-correlation and that the strongest relationship should be contemporary, the pre-whitened cross-correlation at lag 0 was chosen as a measure for the suitability of GT data for forecasting.

As argued above, a selection was necessary to ensure a meaningful comparison with pure time series methods. However, the selection of GT data should be made without using the test set, because otherwise it could be argued that selection in favor of the GT based models took place. Therefore, the empirical pre-whitened cross-correlation was only calculated on the training set.

Finally, it had to be considered that for the German GT data as well as for the worldwide GT data, three different versions (original, average divided, trend divided) had to be compared. Queries were included in the competition if at least two of the three versions[3] had a positive and significant pre-whitened cross-correlation at 5% level (see Appendix B.2). According to this procedure, 58 German GT time series and 49 worldwide GT time series were selected.

5.2.2. Forecasting models

A large number of forecasting models had to be estimated: For each of the 49 queries for the German GT data and the 58 queries of the worldwide GT data three different GT based models had to be estimated (according to the three versions of GT data). In addition, seasonal ARIMA models (SARIMA) needed to be estimated for each selected tourism region (see 6.1). This made a total of 356 models, the estimation of which required an automatized model building algorithm.

[3] Since the different versions of GT data needed to be compared, for each query either none or all versions needed to be selected. The decision to include queries if two of the three versions had positive, significant pre-whitened cross-correlation was made to have a criterion which is neither too restrictive nor too weak.

Athanasopoulos, Hyndman, Song, and Wu (2011) perform a tourism demand forecasting competition with automatized time series model selection. Therefore, this paper was taken as a reference.

In the following, the models used in the competition are described. The notation of Box and Jenkins (1976) is used. B denotes the backward shift operator $By_t = y_{t-1}$. The polynomials $\phi(B)$, $\Phi(B^{12})$ and $\psi(B)$, $\Psi(B^{12})$ represent the autoregressive and moving average operators respectively. A detailed introduction of SARIMA and dynamic linear models can be found in Box and Jenkins (1976) and will not be given in this paper.

5.2.2.1. SARIMA models. Due to a change in data recording by the German Federal Office of Statistics in 2012, a level shift $1_{t<2012}$ was included as an intervention variable.[4] This lead to the specification

$$(1-B)^d(1-B^{12})y_t = c + \gamma(1-B)^d(1-B^{12})1_{t<2012} + \frac{\Psi(B^{12})}{\Phi(B^{12})}\frac{\psi(B)}{\phi(B)}\varepsilon_t$$

with $c = 0$ for $d = 1$ and $\varepsilon_t \sim WN$. The models were fitted using conditional sum of squares.

The procedure for the identification of the orders for $\phi(B)$, $\Phi(B^{12})$, $\psi(B)$, $\Psi(B^{12})$ and the order of differencing d is described in Appendix B. The algorithm is similar to the one used for automatized SARIMA model building in Athanasopoulos et al. (2011).

5.2.2.2. GT based DLM's. With the level shift $1_{t<2012}$ [4] and the coefficients of the (lagged) GT data, the DLM's were of the form

$$(1-B)^d(1-B^{12})y_t = c + \gamma(1-B)^d(1-B^{12})1_{t<2012} + \sum_k a_k(1-B)^d(1-B^{12})g_{t-k} + \frac{\Psi(B^{12})}{\Phi(B^{12})}\frac{\psi(B)}{\phi(B)}\varepsilon_t$$

with $c = 0$ if $d = 1$ and $\varepsilon_t \sim WN$. The models were fitted using conditional sum of squares.

The model identification consisted of three aspects. The lag order selection of the input GT data was made using the empirical pre-whitened cross-correlation (see Appendix B.2) between the GT data and the arrivals. It was assumed that tourists usually plan their holidays only a limited number of month ahead which is why little signal was expected in GT data with a lag order higher than 6. Therefore contemporaneous and lags of GT data (up to six months) were included if their empirical pre-whitened cross-correlation with the arrival data was positive and significant at a 5% level (see Appendix B.2).

The identification of the order of differencing d along with the order selection of $\phi(B)$, $\Phi(B^{12})$, $\psi(B)$ and $\Psi(B^{12})$ is described in Appendix B.

5.2.3. Forecast evaluation

There are several different error metrics that could be applied to compare different forecasting methods, and for tourism demand forecasting it is not clear which one is the most appropriate. However, in this paper forecasts across time series of different scale are compared. Therefore, scale-independence of the error metric is a natural requirement. E.g. a scale-dependent error metric like the mean square error would give relatively more weight to the most frequently visited tourism regions.

Since the different error metrics might yield different results regarding the relative performance of the methods, two different error metrics were used:

$$MAPE = \frac{1}{T-N}\sum_{t=N+1}^{T}\frac{|y_t - \hat{y}_t|}{y_t}$$

$$MASE = \frac{1}{T-N}\sum_{t=N+1}^{T}\frac{|y_t - \hat{y}_t|}{\frac{1}{N-12}\sum_{t=13}^{N}|y_t - y_{t-12}|}$$

Here T is the length of the time series, and N is the length of the in-sample period. $\hat{y}_t$ denotes the forecast at point t. The mean absolute percentage error (MAPE) is one of the most commonly used measures in tourism demand forecasting according to Song and Li (2008). In contrast, the mean absolute scaled error (MASE), suggested by Hyndman and Koehler (2006), is less widespread but overcomes several disadvantages of the MAPE (like a deficiency in the case $y_t = 0$) and has a natural interpretation as a comparison to the seasonal naive forecast.

To compare forecasts of two different methods on one time series Diebold and Mariano (2002) derive an asymptotic test statistic in the following way: For a loss function L (in this paper $L(y_t, \hat{y}_t) = |y_t - \hat{y}_t|$) a time series is defined by $d_t = L(y_{1,t}, \hat{y}_{1,t}) - L(y_{2,t}, \hat{y}_{2,t})$. The null hypothesis of equal performance $E[d_t] = 0$ can be tested using the statistic $\frac{\bar{d}_t}{\hat{V}(\bar{d}_t)^{1/2}}$. Several of the GT based tourism demand forecasting studies in section 2.1 use the Diebold-Mariano test for comparison.

In the forecasting competition, the number of time series where one method performed superior over another was calculated. Section 6.1 presents comparisons of each GT based model against the SARIMA models. In section 6.2 the DLM's using the original data are compared with the ones using trend divided GT data.

6. Results

6.1. The informational value of GT data

The first question regarding the performance of the forecasting methods is whether the use of GT data leads to improved forecasts in comparison with methods based solely on historical data.

For 33 tourism regions queries were found, such that the German GT data met the criterion described in section 5.2.1. For the worldwide GT data, 35 regions were found such that the criterion was passed.

Because for some regions there were multiple possible queries, for each of these queries a model was fitted. Only the model with the lowest in-sample MAPE was used to make forecasts for the region.

The results of Tables 1 and 2 show that the seasonal naive forecast is significantly outperformed by all other methods. Furthermore all GT based methods are superior to the SARIMA forecasts which can be seen from the Diebold-Mariano test results as well as from the MAPE and MASE. For the models based on German trend divided GT data MAPE and MASE are in average 14.8% and 10.8% lower than for the SARIMA forecasts. For the international data, the improvement is 9.1% and 7.7% respectively.[5]

In summary, the results clearly indicate a positive informational value of the GT data in the forecasting competition. The best results were achieved by using trend divided GT data.

6.2. The effect of GT data modification

The second and most significant research question is whether the data modification described in 5.1 leads to improved GT based forecasts compared with the original GT data.

The results of section 6.1 indicate a positive effect of trend division but the forecasts using the average divided GT data did in general not perform better than the forecasts using the original data. Therefore, the following analysis is restricted to the comparison of the trend divided GT data with the original GT data.

[4] Until January 2012 each German accommodation provider with the capacity to host at least 9 people (3 for campsites) had to report the monthly arrivals of tourists. Due to EU law, reporting was only required in case of capacity for at least 10 people (also for camp sites) afterwards. However, the effect of this change on the forecasting results was considered limited since no apparent breaks were visible in the arrival data.

[5] The percentages were calculated according to $\frac{error_{orig} - error_{trend}}{error_{orig}}$.

Table 1
Comparison of forecasting results for DLM's using German GT data against SARIMA and seasonal naive forecasts.

	naive12	sarima	gt_orig	gt_mean	gt_trend
mean MAPE	7.04%	6.41%	5.92%	5.99%	**5.46%**
median MAPE	6.94%	5.85%	5.36%	5.68%	**5.11%**
mean MASE	1.191	1.023	0.973	0.979	**0.912**
median MASE	1.193	0.994	0.903	0.945	**0.901**
DM against SARIMA	/	/	11/5	10/4	14/1

The results are based on forecasts for 33 tourism regions. gt_orig stands for the DLM using unmodified GT data. gt_mean and gt_trend denote the DLM's using mean and trend divided GT data respectively. DM against SARIMA counts how often the method in the respective column outperformed/was outperformed by the SARIMA forecasts according to the Diebold-Mariano test at 5% significance level.

Table 2
Comparison of forecasting results for DLM's using worldwide GT data against SARIMA and seasonal naive forecasts.

	naive12	sarima	gt_orig	gt_mean	gt_trend
mean MAPE	6.77%	6.28%	5.86%	5.81%	**5.71%**
median MAPE	6.51%	5.85%	5.43%	**5.27%**	5.34%
mean MASE	1.113	0.986	0.936	0.931	**0.910**
median MASE	1.126	0.974	0.911	0.889	**0.860**
DM against SARIMA	/	/	8/1	8/0	9/1

The results are based on forecasts for 35 tourism regions. gt_orig stands for the DLM using unmodified GT data. gt_mean and gt_trend denote the DLM's using mean and trend divided GT data respectively. DM against SARIMA counts how often the method in the respective column outperformed/was outperformed by the SARIMA forecasts according to the Diebold-Mariano test at 5% significance level.

In contrast to the situation in 6.1, tourism regions were included in the competition multiple times when there were multiple possible queries. This was done because each participating method generates different forecasts in one region according to the query data used (in contrast to methods solely based on tourist data). Therefore, for the comparison using the German GT data 58 time series were used and for the comparison of worldwide GT data, 49 time series were included.

Tables 3 and 4 show the results of a forecasting competition between the DLM's using unmodified GT data and the DLM's that use trend divided GT data. For the German GT data as well as for the worldwide GT data DLM's using trend divided GT data outperformed their counterparts using unmodified GT data. For the German GT data, the MAPE and MASE are in average 6.1% and 5.8% lower for the models based on trend divided GT data respectively. For the worldwide GT data, the improvement is 2.5% and 3.2% respectively.[5]

Table 3
Comparison of forecasting results for DLM's using trend divided German GT data against DLM's using original GT data.

	gt_orig	gt_trend
mean MAPE	6.06%	**5.69%**
median MAPE	5.93%	**5.37%**
mean MASE	0.955	**0.900**
median MASE	0.900	**0.877**
DM against gt_orig	/	**16/7**

The results are based on forecasts for 58 pairs (g_t, y_t) of GT time series and the corresponding arrival time series. gt_orig stands for the DLM using unmodified GT data. gt_trend denote the DLM's using trend divided GT data. DM against gt_orig counts how often the model in the respective column outperformed/was outperformed by the DLM using unmodified GT data according to the Diebold-Mariano test at 5% significance level.

Table 4
Comparison of forecasting results for DLM's using trend divided worldwide GT data against DLM's using original GT data.

	gt_orig	gt_trend
mean MAPE	5.89%	**5.74%**
median MAPE	5.61%	**5.34%**
mean MASE	0.914	**0.885**
median MASE	0.910	**0.837**
DM against gt_orig	/	**18/6**

The results are based on forecasts for 49 pairs (g_t, y_t) of GT time series and the corresponding arrival time series. gt_orig stands for the DLM using unmodified GT data. gt_trend denote the DLM's using trend divided GT data. DM against gt_orig counts how often the model in the respective column outperformed/was outperformed by the DLM using unmodified GT data according to the Diebold-Mariano test at 5% significance level.

7. Conclusion

First, for a large number of German holiday regions, it was examined whether the use of GT data is helpful for predicting tourist arrivals. The benefit of GT data was clearly proven. However, contemporaneous GT data was used in this study, therefore decision-makers could not use these forecasts to prepare or to anticipate the arriving number of tourists. The benefit of such forecasts is to provide quick estimates of tourist numbers 2–4 weeks before the official numbers arrive. As a possible extension of this approach, it could be evaluated whether accurate predictions could be made using the first two weeks instead of the aggregated GT data for the whole month. This was already successfully done by Choi and Varian (2012) for target region Hong Kong. Such an approach would generate estimates 4–6 weeks before official numbers arrive.

Second and most importantly, this paper aims at a firm analysis of the spurious patterns in GT data. In section 2 it was shown that GT time series for several unrelated search terms exhibit these patterns (downward trend and breaks in 2011 and 2016). The concept of “common disturbance factors” was described in section 4. It was argued that common disturbance factors might be the reason for the spurious patterns and that they could cause concept drift in forecasting models using GT data. To test this hypothesis and to evaluate whether GT data modification by mean division or trend division could help to reduce a possible negative impact on forecasts, predictions based on the original GT data were compared with predictions based on the two types of modified GT data. The forecasts using the trend divided GT data were slightly superior to the ones using the original data. This is in line with the assumption of common disturbance factors and the implication that trend division could reduce the negative impact on forecasting.

From the analysis of the spurious patterns, consequences for the use of GT for the prediction of tourism in particular, as well as for the use of GT data in general can be derived. Regarding tourism demand forecasting there is a number of studies were GT data from Germany is used as a predictor (Camacho & Pacce, 2017; García Rodríguez, 2017; Zeynalov, 2017; Önder, 2017). Accordingly forecasts of these studies are likely affected by spurious patterns and trend division might lead to improved forecasts. To give a more general impression of the effects of spurious patterns, it has been shown that authors of some previous studies may have been misled by these patterns: The downward trend and the following rise in 2011 of the GT index for category “real estate” (see Fig. 1) could be explained by common disturbance factors rather then by its relationship to the real estate market as assumed by Wu and Brynjolfsson (2015). Arguing that the downward trend for the search term “environment” in Fig. 2 is due to decreasing interest towards environmental topics as claimed by Mccallum and Bury (2013) is also questionable regarding the analysis performed in this paper.

Once clear evidence has been provided of the potential problems with the use of GT data, the question naturally arises as to what extend trend division offers a solution. Although trend division is likely to reduce the impact of spurious patterns in many cases, the method is not sufficient to solve the problem completely. On the one hand, it requires the selection of a large number of appropriate queries for the calculation of the trend. This might in general not be so straightforward as in the forecasting setting of this paper. On the other hand, trend division might include its own disturbance into the GT data as discussed in 5.1 and besides that most likely does not remove the spurious patterns completely. For the latter statement the reader is referred to Appendix A. Therefore, rather than in providing a guide on how to fix problems in the GT data, the main contribution of this paper should be seen in broadening the understanding of challenges regarding the use of GT data. To use the full potential of the powerful data source Google Trends and to avoid being trapped by the inherent challenges regarding the use of GT data, a better understanding of the data source is necessary.

Appendix A. Quantitative analysis of common disturbance factors

In the previous parts of the paper, spurious patterns were identified visually and the effect of data modification was evaluated by the analysis of the forecasting results. For a more comprehensive approach towards handling the problem of spurious patterns and common disturbance factors, it would be helpful to have a quantitative measure to identify common disturbance factors and to evaluate the effect of data modification.

The key idea for the definition of such a measure is that GT time series $g_{1,r,t}$, …, $g_{N,r,t}$ from the same region behave similarly whenever there is a big change in the common denominator $Q_{r,t}$. E.g. if $Q_{r,t}$ strongly increases most of the time series $g_{1,r,t}$, …, $g_{N,r,t}$ should decrease at a similar rate. Therefore, common disturbance factors are related to a “high similarity” between the time series.

To quantitatively evaluate where time series behave distinctively similar, a dissimilarity measure has to be defined.

Appendix A.1Definition of the dissimilarity D_t^l

In the following (y_t) denotes a time series while y_t stands for the value of (y_t) at time t. This distinction in the notation is necessary for a better understanding of the following definitions.

First, it should be noted that GT time series are in general non-stationary. Therefore, the means and variances of the individual time series can change with time, potentially leading them to drift apart or together in the long run. This behavior should preferably not influence the dissimilarity defined. Considering the time series within a restricted interval I and standardizing each time series within this interval reduces the effect of non-stationarity.

Let S^I denote the operator for standardizing within an interval. $S^I(y_{i,t})$ stands for the time series $(y_{i,t})$ in the interval I after standardizing in I. The value of $S^I(y_{i,t})$ at point t is given by

$$S^I(y_{i,t})_t = \frac{y_{i,t} - \bar{y}_{i,t}}{\hat{\sigma}(y_{i,t})} \quad \text{for all } t \in I \tag{A.1}$$

with $\bar{y}_{i,t} = \frac{1}{|I|}\sum_{t\in I} y_{i,t}$ and $\hat{\sigma}(y_{i,t}) = \sqrt{\frac{\sum_{t\in I}(y_{i,t} - \bar{y}_{i,t})^2}{|I\,| - 1}}$ and $|I|$ the number of observations for $(y_{i,t})$ in I.

For a group $(y_{i,t})_{i=1,\dots,N}$ of time series and an interval I, the group average (μ_t^I) of the standardized series in the interval is given by

$$\mu_t^I := \frac{1}{N}\sum_{i=1}^{N} S^I(y_{i,t})_t \quad \text{for all } t \in I \tag{A.2}$$

The variation between time series in the interval I is measured by

$$D^I := \frac{1}{N}\sum_{i=1}^{N} d(S^I(y_{i,t}), (\mu_t^I)) \tag{A.3}$$

where d denotes the Euclidean distance.

After specifying a length l, for each point t (with at least l neighbors to the left and right) the dissimilarity D_t^l at point t is defined by

$$D_t^l := D^{[t-l,t+l]} \tag{A.4}$$

The dissimilarity measure depends on the length l. The lower l is chosen, the less the impact of non-stationarity on the result. However, for a very small l long-lasting common disturbance factors might not be detected.

Appendix A.2Application of the dissimilarity measure on the GT data

The dissimilarity measure was applied to the GT data collected. Because the GT time series share a similar seasonality, seasonal adjustment needed to be performed first. Otherwise, there would have been low dissimilarity at certain months solely due to the seasonality. Seasonal adjustment was performed by Seasonal and Trend Decomposition using Loess. Seasonality was specified as additive for all GT time series.

Figures A.8 and A.9 show the dissimilarity for all versions of GT data from Germany and worldwide. For the original German GT data, there is a marked slump around January 2011 and a less significant but visible slump around January 2016. The worldwide original GT data shows a slump between 2009 and 2013 and a more significant slump around January 2016. The spurious patterns visible in the averaged GT data (see Figs. 6 and 7) seem to be the most likely explanation for the slumps in dissimilarity. The fact that the slumps start and end approximately 12 months before/after 2011 and 2016 can be explained by choice of $l = 12$.

For the trend-divided time series, there are similar slumps as in the original GT data. But these are far less significant. The average divided GT data has a constantly lower dissimilarity than the two other versions and does not show any significant slumps.

To give an explanation for the results, the representations of the three GT versions from section 5.1 are considered again: The original GT data $g_{1,r,t}$, $g_{2,r,t}$, …,$g_{269,r,t}$ can be represented by

$$g_{i,r,t} = c_i \frac{q_{i,r,t}}{Q_{r,t}} \tag{A.5}$$

The average divided GT data $\bar{g}_{1,r,t}$, $\bar{g}_{2,r,t}$, ..., $\bar{g}_{269,r,t}$ has the representation

$$\bar{g}_{i,r,t} = \frac{c_i q_{i,r,t}}{\frac{1}{269}\sum_j c_j q_{j,r,t}} \tag{A.6}$$

And the trend divided GT data $\tilde{g}_{1,r,t}$, $\tilde{g}_{2,r,t}$, ..., $\tilde{g}_{269,r,t}$ has the representation

$$\tilde{g}_{i,r,t} = \frac{\frac{c_i q_{i,r,t}}{Q_{r,t}}}{F\left(\frac{1}{269}\sum_j c_j \frac{q_{j,r,t}}{Q_{r,t}}\right)_t} \tag{A.7}$$

The fact that the average divided GT data does not show slumps at the times of the spurious patterns could be explained by the fact that there is no $Q_{r,t}$ in representation A.6. This would also explain why the slumps did not completely vanish for the trend divided GT data: The operator F might remove some variation from $\frac{1}{269}\sum_j c_j \frac{q_{j,r,t}}{Q_{r,t}}$ which is due to the effect of common disturbance factors on $Q_{r,t}$. This is why a part of the effect of the common disturbance factors on the numerator $\frac{c_i q_{i,r,t}}{Q_{r,t}}$ is not removed by the trend division and leads to the remaining distinct similarity.

Appendix A.3The utility of the dissimilarity measure

According to the above discussion the potential utility of the dissimilarity is twofold: On the one hand, it provides a method to detect common disturbance factors.

On the other hand, the effect of GT data modification methods can be evaluated. For example, Figs. A.8 and A.9 indicate that trend division leaves parts of the effect which is attributed to common disturbance factors. However, this does not mean that the goal of GT data modification should be to achieve the highest dissimilarity. As discussed in section 5.1 the GT data could, for example, have a common upward trend due to an overall increase of tourism in Germany. Removal of this trend would likely lead to higher dissimilarity but probably not improve the value of the GT data as a predictor. The results of the tourism demand forecasting competition show that, despite the higher dissimilarity, forecasts based on the average divided GT data are inferior to the ones based on trend divided GT data.

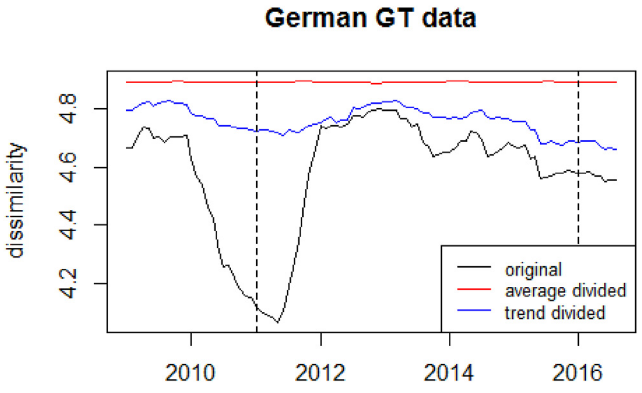

Fig. A.8. Representation of D_t^{12} for the German GT data.

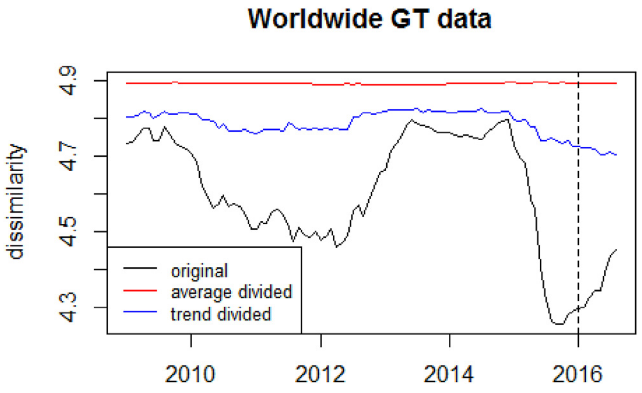

Fig. A.9. Representation of D_t^{12} for the worldwide GT data.

Appendix B. Model identification

The SARIMA models, as well as the GT based dynamic linear models, can both be written as

$$(1 - B^{12})y_t = c + \gamma(1 - B^{12})1_{t<2012} + \sum_k a_k(1 - B^{12})g_{t-k} + N_t \tag{B.1}$$

with a seasonal ARIMA process N_t, such that $\frac{\Phi(B^{12})}{\Psi(B^{12})}\frac{\phi(B)}{\psi(B)}(1-B)^d N_t = \varepsilon_t \sim WN$.

In case of the SARIMA models, the coefficients a_k are set to 0. Thus only N_t needs to be specified, which corresponds to the identification of the orders p, q, P and Q of the polynomials $\phi(B)$, $\psi(B)$, $\Phi(B^{12})$ and $\Psi(B^{12})$ along with the order of integration d. For the GT based DLM's the input lag order needs to be specified in addition to the specification of N_t. In the following the order selection procedure for N_t is described. For the GT based DLM's the procedure requires that the input lag order is already selected. The according procedure for the input lag order selection is described in Appendix B.2.

Appendix B.1SARIMA order selection

The SARIMA order $(p, d, q)(P, D, Q)$ to represent N_t is selected in the following way:

(1) $D = 0$ per default since equation (B.1) is already in seasonal differences.
(2) OLS regression is performed for equation (B.1) with $c = 0$ to calculate the residuals $(\hat{N}_t)$
(3) A Kwiatkowski–Phillips–Schmidt–Shin unit root test is applied to $(\hat{N}_t)$. If the null hypothesis of trend stationarity is rejected at a 5% level d is set to 1.
(4) From this point on, different models for N_t are specified and estimated. To start, the following SARIMA orders are chosen: (2,d, 2) (1,0,1), (0,d, 0) (0,0,0), (1,d, 0) (1,0,0) and (0,d, 1) (0,0,1). The model with the smallest AIC value becomes the tentative model.
(5) Up to thirteen models with the orders p,q,P,Q close (± 1) to the ones from the tentative model are fitted. If a model with a smaller AIC than the tentative model is found, it becomes the new tentative model and step 5 starts again. If this is not the case, the tentative model becomes the final model.

This method is described in Athanasopoulos et al. (2011). The algorithm for the SARIMA order selection is implemented in the *auto.arima* function from Hyndman and Khandakar (2008).

Appendix B.2Input lag order selection

Pre-whitened cross-correlation can best be described in the context of dynamic linear modeling where one needs to decide whether certain lags of g_t should be included in the model because they have an influence on y_t. Due to possibly similar auto-correlation structure of g_t and y_t, the classical Pearson correlation coefficient can be misleading when a measure of the linear relationship is required. This phenomenon is known as "spurious regression" and described in Hamilton (1994, p. 557–562). Pre-whitened cross-correlation overcomes this issue by removing the shared auto-correlation structure.

Under the assumption that g_t is a SARIMA process, there exists an operator F fulfilling $Fg_t = \varepsilon_t$ where ε_t is white noise. The pre-whitened cross-correlation between g_t and y_t at lag k is given as

$$cor(Fg_{t-k}, Fy_t) \tag{B.2}$$

As proven by Box and Jenkins (1976, p. 379–382) the pre-whitened cross-correlation at lag k corresponds to a_k in representation B.1 Therefore the empirical pre-whitened cross-correlation can be used to identify the input lag order in a dynamic linear model.

It can further be shown that if the pre-whitened cross-correlation of g_t and y_t at lag k is zero, the empirical pre-whitened cross-correlation is asymptotically normally distributed with standard deviation $(T-k)^{-\frac{1}{2}}$ (Box & Jenkins, 1976, pp. 379–382).

Note that the derived distribution only holds under the assumption that F is known in advance and really fulfills $Fg_t \sim WN$. In practice F is unknown, and an approximate SARIMA model for g_t needs to be specified from which only estimates of the innovations $\hat{\varepsilon}_t$ can be obtained. In this paper the algorithm from Appendix B.1 was applied to estimate a SARIMA model for each GT time series $g_{i,t}$. The residuals of the estimated model are used as estimates for Fg_t.

References

Artola, C., & Martínez-Galán, E. (2012). *Tracking the future on the web: Construction of leading indicators using internet searches. Banco de Espana occasional paper No. 1203*. Online available at: https://doi.org/10.2139/ssrn.2043056.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly, 55*, 107–120.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*, 822–844.

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454–464.

Box, G. E., & Jenkins, G. M. (1976). Time series analysis. Forecasting and control. *Holden-day series in time series analysis*(Revised ed.). San Francisco: Holden-Day 1976.

Bulut, L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting, 37*, 303–315.

Camacho, M., & Pacce, M. J. (2017). Forecasting travellers in Spain with google's search volume indices. *Tourism Economics, 24*, 434–448.

Chatfield, C. (2000). *Time-series forecasting.* Chapman and Hall/CRC.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *The Economic Record, 88*, 2–9.

Cleveland, R. B., Cleveland, W. S., & Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*, 3.

Destatis (2017). *Ergebnisse der Monatserhebung im Tourismus.* Data retrieved at https://www.destatis.de/DE/Publikationen/Thematisch/BinnenhandelGastgewerbeTourismus/AlteAusgaben/TourismusMAlt.html Visited on 2017-10-15.

Destatis. Monatserhebung im Tourismus- Qualitätsbericht. (2018). Online availabe at https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/Qualitaetsberichte.html Visited on 2018-05-29.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*, 134–144.

Ficetola, G. F. (2013). Is interest toward the environment really declining? The complexity of analysing trends using internet search data. *Biodiversity & Conservation, 22*, 2983–2988.

García Rodríguez, Ó. (2017). Forecasting tourism arrivals with an online search engine data: A study of the balearic islands. *PASOS : Revista de Turismo y Patrimonio Cultural, 15*, 943–958.

Gawlik, E., Kabaria, H., & Kaur, S. (2011). Predicting tourism trends with google insights. Online available at: Ωcs229.stanford.edu/proj2011/GawlikKaurKabaria-PredictingTourismTrendsWithGoogleInsights.pdf.

GNTB (2018). *The top 100 sights and attractions in Germany.* http://www.germany.travel/en/towns-cities-culture/top-100/germany-travel-attractions.html visited on 2018-02-23.

Google (2018a). *Google trends – search volume in Germany for "usedom".* https://trends.google.de/trends/explore?date=all&geo=DE&q=Usedom visted on 2018-05-18.

Google (2018b). *Google Trends help – how Trends data is adjusted.* https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 visited on 2018-04-23.

Google (2018c). *Google Trends help – trends graphs and forecasts.* https://support.google.com/trends/answer/4359597?hl=en visted on 2018-04-23.

Gunter, U., & Önder, I. (2015). Forecasting international city tourism demand for paris:

Accuracy of uni-and multivariate models employing monthly data. *Tourism Management, 46*, 123–135.

Hamilton, J. D. (1994). *Time series analysis*, *Vol. 2*. Princeton: Princeton university press.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 26*, 1–22.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679–688.

de Kort, R. E. (2017). *Forecasting tourism demand through search queries and machine learning.* Bank for International Settlements. IFC Bulletins chapters 44, Online available at: https://www.bis.org/ifc/publ/ifcb44f.pdf.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google Flu: Traps in big data analysis. *Science, 343*, 1203–1205.

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66.

Massicotte, P., & Eddelbuettel, D. (2017). gtrendsR: Perform and display google trends queries. https://github.com/PMassicotte/gtrendsR_r_package_version_1.4.1.

Mccallum, M. L., & Bury, G. W. (2013). Google search patterns suggest declining interest in the environment. *Biodiversity & Conservation, 22*, 1355–1367.

Önder, I. (2017). Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research, 19*, 648–660.

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using google trends data. *Journal of Travel & Tourism Marketing, 34*, 357–368.

Santillana, M., Zhang, D. W., Althouse, B. M., & Ayers, J. W. (2014). What can digital disease detection learn from (an external revision to) Google Flu Trends? *American Journal of Preventive Medicine, 47*, 341–347.

Siliverstovs, B., & Wochner, D. S. (2018). Google Trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization, 145*, 1–23.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting–A review of recent research. *Tourism Management, 29*, 203–220.

Stephens-Davidowitz, S., & Varian, H. (2014). *A hands-on guide to Google data.* Mountain View, CA: Google. Online available at: http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf.

Tripadvisor. Activities Germany. (2018). https://www.tripadvisor.de/Attractions-g187275-Activities-Germany.html visited on 2018-04-03.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting, 30*, 565–578.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. *Economic analysis of the digital economy* (pp. 89–118). University of Chicago Press.

Xiang, Z., & Pan, B. (2011). Travel queries on cities in the United States: Implications for search engine marketing for tourist destinations. *Tourism Management, 32*, 88–97.

Xiang, Z., Wang, D., O'Leary, J. T., & Fesenmaier, D. R. (2015). Adapting to the internet: Trends in travelers' use of the web for trip planning. *Journal of Travel Research, 54*, 511–527.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015b). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397.

Yang, S., Santillana, M., & Kou, S. C. (2015a). Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences, 112*, 14473–14478.

Zeynalov, A. (2017). *Forecasting tourist arrivals in prague: Google econometrics. Munich personal RePEc archive.* Munich: MPRA. Paper No. 83268. Online available at: https://mpra.ub.uni-muenchen.de/83268/.