**Annals of Operations Research manuscript No.**
(will be inserted by the editor)

# Google search volumes for portfolio management: performances and asset concentration

**Mario Maggi · Pierpaolo Uberti**

Received: date / Accepted: date

**Abstract** Google search volumes have proven to be useful in portfolio management. The basic idea is that high search volumes are related to bad news and risk increase. This paper shows additional evidence about the use of Google search volumes in risk management, for the Dow Jones Industrial Average index components from 2004 to 2017. To overcome the (time-series and cross-section) limitations Google imposes on data download, a renormalization procedure is presented, to obtain a multivariate sample of volumes, which preserve their relative magnitude. The results indicate that the volume normalization is relevant for portfolio performances. Renormalized Google search volumes yield poor results when they penalize the portfolio diversification. Instead, if the portfolio diversification can be kept to an acceptable level, the renormalized Google search volumes contribute to improving risk-adjusted performances.

**Keywords** Web searches · Google Trends · Portfolio management

**JEL classification** C81 · G11

## 1 Introduction

In the recent years, the increasing availability of web data fueled a wave of studies that analyzed the relationships between web searches and many aspects of social sciences. Since Ginsberg et al. (2009), Polgreen et al. (2008), who first used Google search volumes to forecast the diffusion of influenza, the information contents of Google queries have also been analyzed for other phenomena. In particular, some studies focused on the relations between web searches and economic variables (for

M. Maggi
Department of Economics and Business, University of Pavia
Via S. Felice, 5, 27100, Pavia, Italy
Tel.: +39-0382986234
E-mail: mario.maggi@unipv.it ORCID: 0000-0002-7233-4443

P. Uberti
DIEC Department of Economics, University of Genoa, Italy
E-mail: uberti@economia.unige.it ORCID: 0000-0001-8426-0368

example, see Askitas et al., 2009; Bijl et al., 2016; Dzielinski, 2012; Heiberger, 2015; Joseph et al., 2011; Kristoufek, 2015; Li et al., 2015; Liu et al., 2015; Mondria et al., 2010; Vlastakis and Markellos, 2012; Vozlyublennaia, 2014). Before the application in finance, Askitas et al. (2009) used search volumes to analyze economic conditions, such as unemployment rates. Alanyali et al. (2013) showed that there is evidence of the information flow from media to the financial market.

The information contents of Google search volumes has been documented in a financial framework by, among others, Da et al. (2011); Dzielinski (2012); Heiberger (2015); Kristoufek (2015); Li et al. (2015); Liu et al. (2015); Vlastakis and Markellos (2012); Vozlyublennaia (2014).

In Mondria et al. (2010) and Da et al. (2011), search engine volumes are proposed to approximate investor attention, which is known to influence stock market volatility. Also in Joseph et al. (2011), Google search volumes represent a proxy for investor attention, finding that search volumes are useful in predicting stock returns and trading volumes, especially abnormal movements. Moreover, high levels of Google search volumes are shown to be correlated to negative returns (see, for example, Bijl et al., 2016). In particular, Preis et al. (2014) prove that financial market downturns are preceded by rising investor concern, measured by Google Trends search volumes. Recently, following the previous evidences, different works have explored the possibility of exploiting the forecasting power of web data to set up asset allocation and trading strategies. For instance, Kristoufek (2015) and Preis et al. (2014) use Google search volumes with the aim of improving the return or the risk-return combination of a financial portfolio, while Bijl et al. (2016) propose implementing a trading strategy selling stocks with high Google search volumes and buying those with low search popularity.

This paper deepens the analysis by Kristoufek (2015) on the contribution of Google search volumes to the asset allocation performances of a portfolio containing the Dow Jones Industrial Average components. The basic idea is that web search volumes are related to the flow of news, principally bad news. The searched news can increase the trading activity and the volatility of the stocks, therefore the search volume may be considered as a risk indicator. We focus on the need to obtain a multivariate series of web search volumes whose sizes are proportional to the (undisclosed) real volumes. This goal is achieved by a renormalization procedure aimed at rebuilding the information not directly available online. The application of renormalized series yields different results with respect to Kristoufek (2015) and, although it is effective in reducing the risk, it seems to worsen the portfolio risk-adjusted performances. A deeper analysis on the portfolio concentration suggests that the use of Google volumes, after the renormalization procedure, leads to badly diversified portfolios with poor out-of-sample performances. The concentration issue depends on the huge differences between the relative search volumes of different items. It can be empirically shown that controlling for the portfolio diversification helps in improving the risk-adjusted performances and that our results are qualitatively not different from the ones obtained in Kristoufek (2015). Moreover, we provide evidences that the introduction of transaction costs does not significantly affect the results.

This paper proceeds as follows. Section 2 presents how to obtain Google search volumes through the Google Trends web page. Section 3 describes the procedure we set up to build a multivariate series of web search volumes, overcoming some limitations Google imposes on the disclosure of data. In Section 4 the web search

volumes are used to build a financial portfolio containing the stocks listed on the Dow Jones Industrial Average index (DJI in what follows). Section 5 analyzes the role of diversification in the performance of the portfolios built on the web search volumes. Therefore, we show the benefits of increasing the diversification of these portfolios. Section 6 concludes.

## 2 Google Trends

It is well known that certain search engines, beyond providing services, base their business on the collection of data about the users' activity and profile. These data are useful to allow the web search providers to improve their search algorithms and to customize their services. Like other search engines, also Google acquires data about every query users type on its web page; Google decided to disclose a (small) part of these data through its service Google Trends.[1] These data can be useful for many applications. Goel et al. (2010) provided a survey of papers on this topic, specifically describing some of the advantages, together with the limitations, of web search derived data. They pointed out how these data are very easy to collect and to use, also underlining how often they may be regarded as helpful in nowcasting and forecasting frameworks, even providing a weak improvement in the predictive power of various models.

On the Google Trends page, data about search volumes are available starting from January 2004. The region and time window may be customized and multiple series can be downloaded in `csv` format. Google does not allow the downloading of raw volumes, but of only a normalized index named *Google Index* (GI in the following). GI takes integer values between 0 and 100. The maximum volume attained on the selected window is set to 100; all the other volumes are calculated accordingly and rounded to the nearest integer. In this way, the dynamic properties are retained, but the absolute size of the volume is lost.

Google imposes other limitations on the disclosed data. First, it is possible to download up to 5 multiple series. Second, there is a constraint relating to the length and the frequency of the series: the longer the time window, the lower the frequency of the data (monthly over 5 years, weekly from 3 months to 5 years, daily from 7 to 270 days, and so on). The actual information regarding search volumes is extremely precious and the Google database is enormous. Therefore, the real volumes are disclosed with reasonable prudence. In fact, GI is not derived from the real volumes, but from a sample of volumes drawn from the Google database. Consequently, the results can vary if the data of the same query are downloaded at different times. However, due to the random sampling, the differences are small enough not to be an issue.

For each query, a specification can be selected. For example, consider Coca-Cola, it is possible to choose between "Coca Cola" (Soft drink), "The Coca Cola Company" (Beverages company), "coca cola" (search term), therefore yielding different outputs. In the framework of asset allocation, we are interested in "The Coca Cola Company" (Beverages company). We adopt this approach for the considered stocks, i.e. we choose the specification which best indicates the company. On this point, there are various approaches in the literature. Some authors, for example

[1] See `https://trends.google.com/trends/`

Da et al. (2011) and Kristoufek (2015), use the company ticker as a search term. It is preferable to use the company name for different reasons. First of all, the investors collecting information on the web are principally non-professional and we do not think they use the ticker to identify a company. Secondly, professional managers using tickers have access to professional data providers, therefore they do not need to search tickers on the web. Moreover, in some cases, the ticker of a given stock may depend on the data provider. Finally, some tickers can have misleading meanings, like KO for Coca-Cola.

## 3 Renormalization of the GI series

This section discusses an issue that, as far as we are aware, has not already been fully taken into consideration. Here, we present how to obtain a sample of GI values with more than 5 series and with a length that does not necessarily depend on the frequency, nonetheless preserving the relative sizes of search volumes.

In this paper, we use a multivariate sample of weekly data from 2004 to 2017. To obtain this sample, we shall overcome the constraints indicated above regarding the length of time, the frequency, and the number of joint queries. To this end, we follow a renormalization procedure aimed at rebuilding the raw web search volumes, paying attention to the information loss due to rounding.

First of all, to obtain weekly series for the entire time window, we downloaded all the series for three periods, each period not longer than 5 years (for weekly data). The three series overlap at the boundary dates, so that it is possible to concatenate them, by matching the values for the overlapping weeks. We do not round the results to avoid the introduction of approximation errors and, thus, information loss. We illustrate the procedure with a simplified example. Consider the weekly GI values for the query "Goldman Sachs" downloaded on the two time intervals: from July 15, 2007 to July 15, 2012 and from July 15, 2012 to July 9, 2017. The gray line in Fig. 1 shows the downloaded data. Note that there are two maxima with GI = 100, one on each subperiod. In fact, the GI values are independently scaled with respect to the maximum value in each downloaded period. Therefore, on July 15, 2012 the two series display a visible jump in the boundary point, because their scales change. So, we modify the scale of the second period in order to obtain the same GI value for the week starting on July 15, 2012: the black line in Fig. 1 shows the result. We want to highlight that the rescaled values are not rounded to the closest integer to avoid the introduction of errors and the loss of information. For our dataset, the series are concatenated over the entire period July 2007-July 2017, using three downloaded subsamples.

To compose the multivariate sample of the GI values of all the DJI components, we downloaded the series 5 by 5 taking into account two main issues: *(i)* each set of 5 series has its own scale, so the sizes are not comparable across different sets; *(ii)* in case in the same sets there are series with highly different sizes, the smaller series suffers from a strong information loss, due to the rounding. To clarify the second point, let us consider, for example, the GI for "The Coca Cola Company" and "Cisco Systems". Fig. 2 reports the independent queries on the left pane, where both series have a maximum equal to 100, and this means that they are independently normalized. On the contrary, the right pane shows that, when the download is joint, i.e. the scale of the volumes is the same, only the larger series

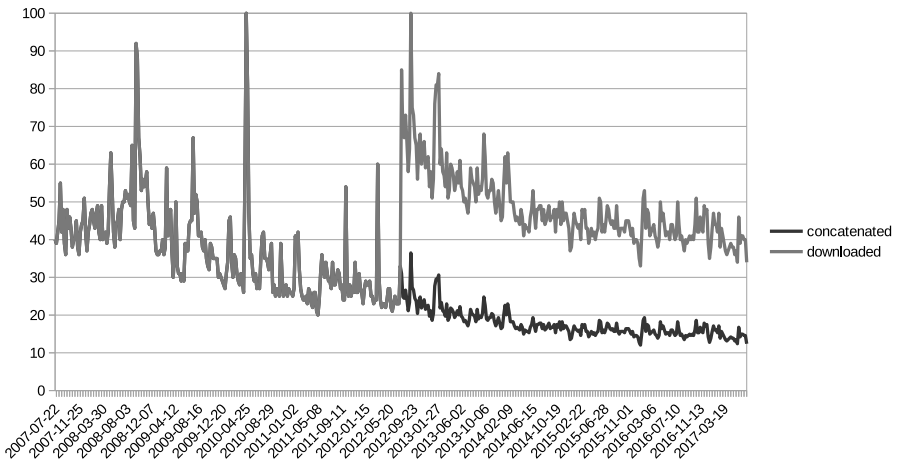


**Fig. 1** Weekly GI values from July 2007 to July 2017 for the query "Goldman Sachs"; data have been downloaded for two periods: July 2017-July 2012 and July 2012-July 2017. Downloaded (gray) and concatenated (black) search volumes.

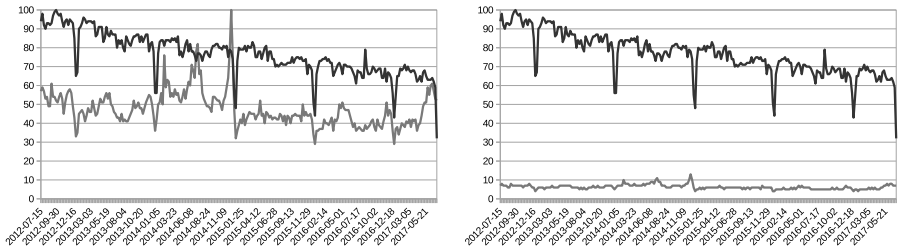

**Fig. 2** Weekly GI values from July 2012 to July 2017 for the queries "The Coca Cola Company" (gray) and "Cisco Systems" (black). Independet queries (left pane), joint queries (right pane).

peaks at 100, while the other is rescaled accordingly. Therefore, the dynamics of the smaller series, in this case "The Coca Cola Company", is compressed and the rounding produces a dramatic information loss. In fact, the GI for "The Coca Cola Company" ranges from 29 to 100 when downloaded alone, while it ranges from 4 to 13 (only integer values) when downloaded together with "Cisco Systems".[2]

We show with a simplified example how the procedure works. Assume it is possible to download up to 2 joint series, and the objective is to build a sample with 3 series. Consider the queries "The Coca Cola Company", "Goldman Sachs", and "Cisco Systems" from July 15, 2012 to July 9, 2017. To weaken the issue shown in Fig. 2, we chose to download the two pairs ("The Coca Cola Company","Goldman Sachs") and ("Goldman Sachs","Cisco Systems"). The downloaded search volumes are displayed in Fig. 3. Remark that in each plot there is a maximum value equal to 100 and all the remaining volumes are scaled accordingly. Like for time con-

[2] Potentially, the situation can be even more severe. Consider, for example, the joint download of "Pfizer" and "The Home Depot", where the GI of "Pfizer" is flattened to 1, losing all relative and dynamical information.

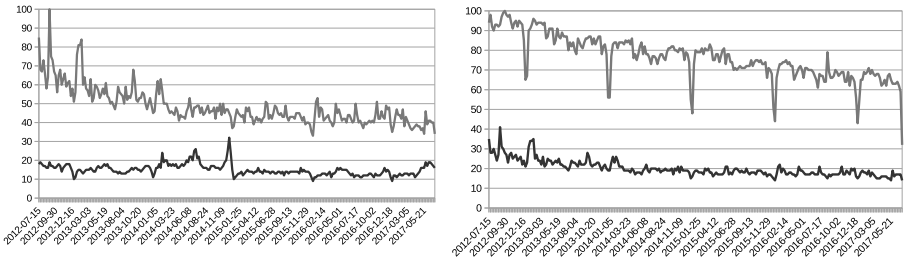

**Fig. 3** Weekly GI values from July 2012 to July 2017. Left pane "The Coca Cola Company" (black) and "Goldman Sachs" (gray); right pane "Goldman Sachs" (black) and "Cisco Systems" (gray).

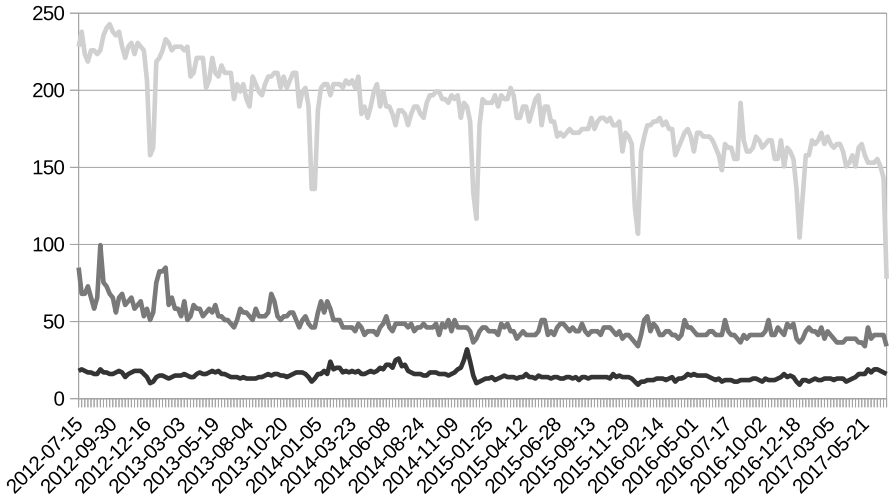


**Fig. 4** Weekly GI values from July 2012 to July 2017 for the rescaled series "The Coca Cola Company" (black), "Goldman Sachs" (gray), and "Cisco Systems" (light gray).

catenation, there are some overlapping data; in this case, the "Goldman Sachs" time series. It is clear that the "Goldman Sachs" series has the same behavior in the two graphics, except for its scale. Therefore, we change the scale of the entire second pair ("Goldman Sachs","Cisco Systems"), to match the values of "Goldman Sachs" in the first pair. The result is shown in Fig. 4, where the three series are scaled with respect to "Goldman Sachs" and so the values for "Cisco Systems" are larger than 100. We underline also in this case that the rescaled values are not rounded to the closest integer to avoid the introduction of rounding errors and the loss of information. To obtain a multivariate sample that preserves the relative size of the search volumes, we apply the described concatenation and renormaliztion procedures.

## 4 Asset allocation based on Google search volumes

Following the approach proposed by Kristoufek (2015), we use the GI values data to find the weights of a (long only) portfolio composed by the DJI stocks.[3] The basic idea is that web search volumes are related to bad news. Therefore, the web search volume is used as a risk indicator: a rise in the interest on a given stock can indicate that many people collect information for trading purposes. In fact, Heiberger (2015), Kristoufek (2015) and Vozlyublennaia (2014) provide evidence of the relation between web searches and trading volumes. An increase in trading produces a possible increase in the price volatility (see also Dzielinski, 2012; Vlastakis and Markellos, 2012).

Considering the web search volume as a risk measure, it is possible to extend to the GI the *equal risk contribution* (ERC) rule proposed by Maillard et al. (2010) to manage risk. Consider, for simplicity, the case of non correlated returns; following the ERC rule, the portfolio weights are set to

$$w_{i,t+1} = \frac{V_{i,t}^{-1}}{\sum_{j=1}^{N} V_{j,t}^{-1}}, \tag{1}$$

where $V_{i,t}$ is the volatility of asset $i$ at time $t$. This way, each asset provides the same contribution $w_{i,t+1}V_{i,t}$ to the portfolio risk. The resulting weights are proportional to the reciprocal of the volatilities, yielding large weights for low volatility stocks and underweighting the most volatile assets.

The rule in Eq. (1) can be generalized in two directions: *(i)* allow $V_{i,t}$ to be any risk measure; *(ii)* introduce a parameter controlling for the relevance of $V_{i,t}$ in fixing the weights. We continue to call the generalized rule as ERC rule, even though it is evident that the argument concerning the equal contribution to the portfolio risk cannot apply.

Now consider the web search volume as a risk indicator, so let $V_{i,t}$ be the GI for stock $i$, at time $t$. The ERC rule yields the portfolio weights

$$w_{i,t+1} = \frac{V_{i,t}^{-\alpha}}{\sum_{j=1}^{N} V_{j,t}^{-\alpha}}, \tag{2}$$

where $\alpha$ is the additional parameter that controls for the relevance of $V_{i,t}$. $\alpha$ also allows to consider more general cases. In fact, for $\alpha = 0$, the portfolio is uniform, $w_i = \frac{1}{N},\ i = 1, \ldots, N$. In other words, the uniform portfolio is a special case of the rule in Eq. (2). For $\alpha > 0$, $w_i$ decreases with $V_{i,t}$, underweighting stocks with large GI values. The weights behave like those of the ERC rule in Eq. (1), with more a pronounced departure from the uniform portfolio as $\alpha$ increases. For $\alpha < 0$, $w_i$ increases with $V_{i,t}$, overweighting stocks with large GI values. This case also allows to consider allocations that prefer assets with high risk.

We set up portfolios applying the ERC rule in Eq. (2), on the basis of weekly data (GI values and stock prices) of the stocks listed on the Dow Jones Industrial Average index (DJI), from July 2004 to July 2017. The stocks we consider are those listed on the DJI in 2017, net of stocks with incomplete series: we end up

[3] Kristoufek (2015) proposes a comparison between the use of web search volumes in-sample and out-of-sample. In this paper, we only consider the so called out-of-sample analysis, which is appropriate to build an asset allocation rule.

with a sample of 26 stocks. The weekly GI values refer to the search terms "Company name" as company and not as products (e.g. "The Coca Cola Company" (Beverages company) instead of "Coca Cola" (Soft drink)). The GI series are downloaded and renormalized according to the procedure described in Section 3. To analyze performances, we compare the results obtained in case of univariate data (independent GI downloads, series by series) with those obtained in case of renormalized data.

As a first explorative result, we evaluate the contribution of using GI values, setting $\alpha$ equal to 0, 0.1, 0.5, and 1 in Eq. (2), to progressively introduce the GI information. In Table 1, the rows "renorm" show that, with respect to the index, the use of GI values increases the average and the standard deviation of returns, whereas the $\text{VaR}_{5\%}$ is lower for the smaller values of $\alpha$. Table 1 also reports the results obtained using GI values independently downloaded, without renormalization (rows "non norm"). These data lead to better results than the renormalized ones in terms of average return, whereas the standard deviation and the $\text{VaR}_{5\%}$ are lower only for $\alpha = 1$. For the sake of completeness, also the uniform portfolio is considered (i.e. $\alpha = 0$). These results show that the renormalization procedure is not neutral, but it substantially impacts the performances of the strategies. Moreover, Table 1 shows the performance indicators when transaction costs are taken into account. Transaction costs are set to 5 basis points, both on buy and sell. The results shown in Table 1 raise some points in contrast to the assumption that the web search volume is a risk indicator. In fact, *(i)* the uniform portfolio has good performances; *(ii)* non normalized series, which totally neglect the relative sizes of the various searches, perform better than renormalized series. To better understand whether the GI values are really a valuable and consistent tool for asset management, we deepen the analysis.

We also remark that the introduction of transaction costs obviously worsens performances (see Table 1). However, the weekly turnover caused by the application of the active strategy in Eq. (2) produces transaction costs that do not completely erode performances and, in particular, do not qualitatively alter the comparison between the various cases we consider. For this reason, and because the effective specification and size of transaction costs may vary upon the context, in what follows we assume no transaction costs.

First of all, we compute the average, the standard deviation, the $\text{VaR}_{5\%}$, and the Sharpe ratio of the portfolio returns, for $\alpha \in [0, 2]$, and we investigate the contribution of GI data with respect to the case $\alpha = 0$. Fig. 5 presents the results. The use of renormalized GI values produces a reduction in the risk indicators for small $\alpha$. This reduction is balanced by a reduction in the expected return such that the Sharpe ratio reduces as well. However, the performances worsen for higher $\alpha$. It is worth noting that, without the renormalization, the portfolio performances improve in terms of all the reported indicators, when $\alpha$ increases, but is less than around 1.2. For larger $\alpha$, the risk indicators slightly increase, but the increase in the average more than compensates the riskiness, producing a decreasing Sharpe ratio. This last result is in line with Kristoufek (2015). Although the renormalization of the GI values makes the asset allocation consistent with the assumption that web search volume is a risk indicator, renormalized GI values produce quite disappointing performances.

We highlight that the two cases significantly differ in terms of portfolio diversification. This is the consequence of the scale of the GI values. In fact, each non

| no transaction costs | | mean | stdev | $\text{VaR}_{5\%}$ | Sharpe |
|---|---|---|---|---|---|
| DJI | | 0.00113 | 0.02245 | 0.03848 | 0.00503 |
| $\alpha = 0$ | | 0.00145 | 0.02361 | 0.03858 | 0.06126 |
| $\alpha = 0.1$ | renorm | 0.00138 | 0.02353 | 0.03752 | 0.05867 |
| | non norm | 0.00147 | 0.02355 | 0.03857 | 0.06234 |
| $\alpha = 0.5$ | renorm | 0.00129 | 0.02319 | 0.03560 | 0.05551 |
| | non norm | 0.00156 | 0.02332 | 0.03655 | 0.06686 |
| $\alpha = 1$ | renorm | 0.00132 | 0.02484 | 0.03889 | 0.05320 |
| | non norm | 0.00169 | 0.02312 | 0.03472 | 0.07293 |
| with transaction costs | | mean | stdev | $\text{VaR}_{5\%}$ | Sharpe |
| $\alpha = 0$ | | 0.00135 | 0.02371 | 0.03909 | 0.05692 |
| $\alpha = 0.1$ | renorm | 0.00134 | 0.02351 | 0.03753 | 0.05691 |
| | non norm | 0.00143 | 0.02355 | 0.03858 | 0.06063 |
| $\alpha = 0.5$ | renorm | 0.00123 | 0.02317 | 0.03563 | 0.05288 |
| | non norm | 0.00149 | 0.02331 | 0.03657 | 0.06413 |
| $\alpha = 1$ | renorm | 0.00110 | 0.02505 | 0.03893 | 0.04383 |
| | non norm | 0.00164 | 0.02311 | 0.03523 | 0.07083 |

**Table 1** Portfolio returns based on GI values, compared to the DJI index and the $\alpha = 0$ uniform portfolio; without transaction costs (top part), with transaction costs (bottom part). Normalized (renorm) search volumes are obtained through the presented renormalization procedure, non normalized (non norm) search volumes are simply the GI values downloaded one by one for each stock.

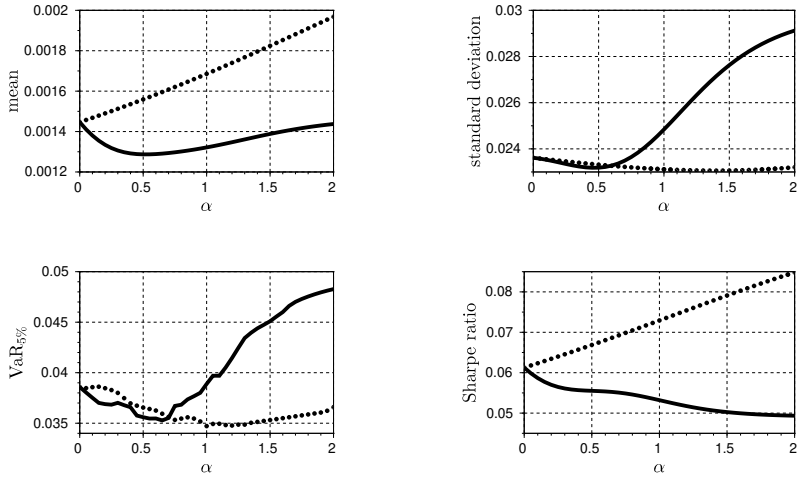


**Fig. 5** Returns average, standard deviation, $\text{VaR}_{5\%}$, and Sharpe ratio of the portfolios composed through Eq. (2). Renormalized GI values (continuous), non normalized GI values (dotted lines).

normalized series have GI values of comparable sizes, having a maximum set to 100. For example, considering again the GI values for "The Coca Cola Company" and "Cisco Systems" discussed in Section 3, the average values of non normalized series are 47.42 and 76.65, respectively; while, after the renormalization, the average values become 15.00 and 186.14, respectively. Obviously, the more the relative sizes of GI are different, the more extreme the weights are, resulting in concentrated portfolios. By consequence, we obtain that the non normalized GI

values produce diversified portfolios, whereas the renormalized GI values yield unbalanced portfolios, because the relative sizes of the GI values have a very wide range. In addition, the portfolio concentration increases with $\alpha$. To measure the portfolio diversification, we compute the Gini coefficient of the weights for $\alpha = 1$, obtaining 0.8078 in case of renormalization, and 0.2389 in case of non normalization.[4] It is clear that the former suffers from low diversification. This fact induces us to study the portfolio diversification and concentration, because we think it may have a central role for the performances.

## 5 GI scaling and diversification

The difference between the non normalized and renormalized portfolios presented in Section 4 is the scaling of the individual GI series. It appears that this scaling has strong consequences on performances. To explore the effects of scaling, we perform a Monte Carlo exercise. We randomly scale the GI values, drawing the scaling factors from a Weibull distribution, which we identify with $\mathsf{Wei}(1, u)$, where the scale parameter is equal to 1 and the shape parameter $u$ is drawn from a uniform distribution $U(0.1, 5)$. This distribution allows us to obtain portfolios with a wide range of Gini coefficients, which vary from 0.085 to 0.999. Therefore, the distribution is suitable to study the diversification impact on performances. Fig. 6 shows the average and the $5^{\text{th}}$ and $95^{\text{th}}$ percentiles obtained from the simulated distributions of the portfolio returns. We remark that, on average, the GI contribution is valuable. In fact, on average, as $\alpha$ increases, the expected return increases and, while the standard deviation increases too, the Sharpe ratio increases, so that the risk-adjusted performances improve with $\alpha$. In terms of $\mathrm{VaR}_{5\%}$, the average behavior is almost flat. However, the variability of the results is quite large and no monotone behavior of the performance indicator is guaranteed.

To focus on the portfolio diversification, we study the conditional distribution of returns with respect to the Gini coefficient of the portfolio weights. We condition the distribution to the value of the Gini coefficient for $\alpha = 1$. In other words, we generate a random scaling, then we compute the portfolio returns for $\alpha \in [0, 2]$, and we associate to each random generation the Gini coefficient of the portfolio obtained with $\alpha = 1$. Fig. 7 shows that, when the Gini coefficient is less than or equal to 0.5, the Sharpe ratio increases with $\alpha$, suggesting that, in case of diversified portfolios, the stronger GI is considered, the better the risk-adjusted performances.[5] On the contrary, for values of the Gini coefficient larger than 0.5, the best performances are obtained for small values of $\alpha$, that is when the GI have a very weak effect and the portfolio is close to the uniform one. The effect

[4] Recall that the Gini coefficient is a concentration measure widely used in economics and statistics, principally to analyze income inequalities (see Gini, 2005). The Gini coefficient ranges from 0 (perfect equality: all the individuals have the same income) to 1 (maximum concentration: 1 individual earns all the income in the economy). In this paper, we use the Gini coefficient as a portfolio concentration measure. This way, we have an indicator of the diversification of the considered portfolios. Here and in the following, we report the Gini coefficient of the average portfolio weights on the considered period: $\mathrm{Gini}\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right)$, where $w_t$ is the vector of the portfolio weights at time $t$.

[5] Remark that the Gini coefficient of the weights of common stock indexes is around 0.5 or below.

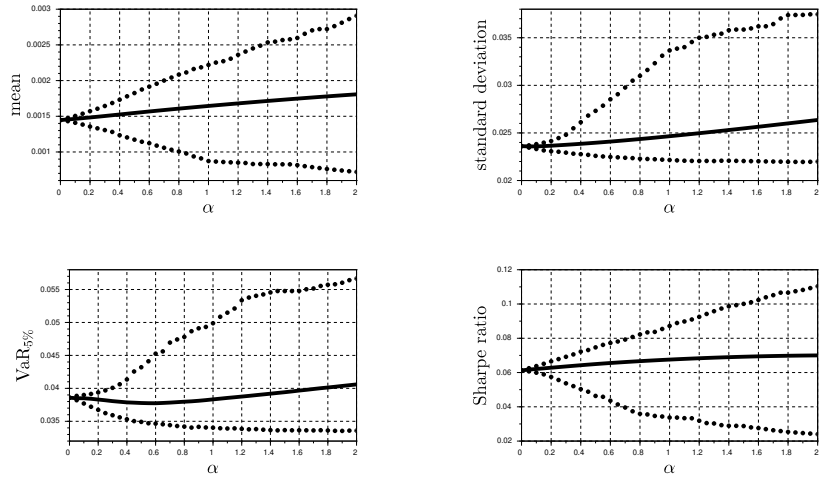


**Fig. 6** Average, standard deviation, $\mathrm{VaR}_{5\%}$, and Sharpe ratio of the returns of the portfolios composed through Eq. (2). Monte Carlo: 1000 random GI normalizations drawn from a Weibull distribution $\mathsf{Wei}(1, u)$, with scale parameter equal to 1 and shape parameter $u$ such that $u \sim U(0.1, 5)$; $\alpha \in [0, 2]$. MC average (continuous), $5^{\text{th}}$ and $95^{\text{th}}$ percentiles (dotted).

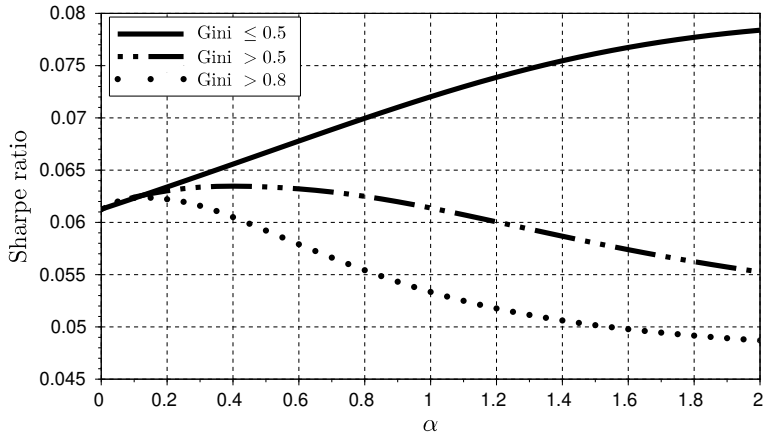


**Fig. 7** Portfolio return Sharpe ratio. The lines describe the average Sharpe ratio, conditional on the portfolio concentration. Monte Carlo: 1000 random GI normalizations drawn from a Weibull distribution $\mathsf{Wei}(1, u)$, with scale parameter equal to 1 and shape parameter $u$ such that $u \sim U(0.1, 5)$; $\alpha \in [0, 2]$. The Gini coefficient is computed for $\alpha = 1$.

is even more pronounced for highly concentrated portfolios (see the case Gini $> 0.8$ in Fig. 7). The slight increase of performances for small values of $\alpha$ obtained in all cases can be explained by the fact that, when $\alpha$ is low, portfolios do not substantially deviate from the uniform case, therefore the diversification is good.

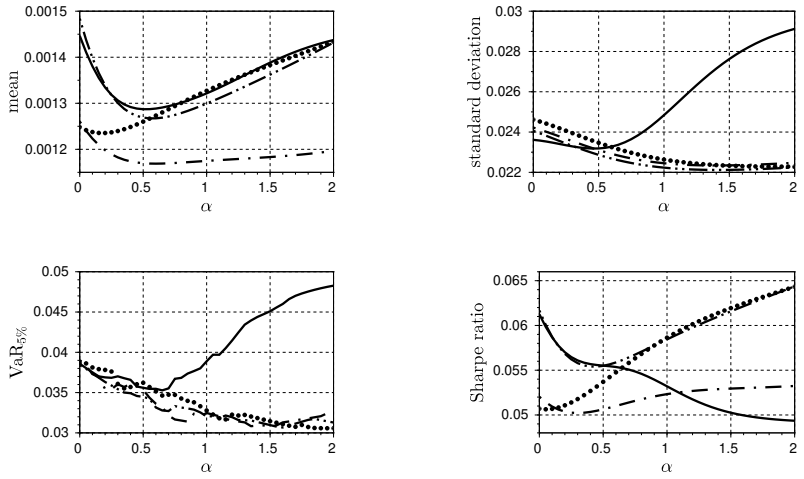


**Fig. 8** Return means, standard deviations, $\mathrm{VaR}_{5\%}$, and Sharpe ratios of the portfolio composed by eliminating the stocks with the $k$ largest and the $k$ smallest GI values. Renormalized GI values. $k = 0$ (—), $k = 1$ (– ·· –), $k = 2$ (– · –), $k = 3$ (· · ·).

It appears clear that, to improve the risk-adjusted performances, the portfolio diversification should be kept under control even if the weights are chosen on the basis of GI values. However, we cannot renormalize as we like, because we cannot arbitrarily choose the relative scale of web search volumes. For this reason, to control for the concentration, we rank the stocks according to the average magnitude of their web search volumes and we progressively delete the stocks with the $k$ largest and the $k$ smallest GI values, with $k = 0, \ldots, 3$. The Gini coefficients of the portfolio weights for $\alpha = 1$ and $k = 0, \ldots, 3$ are 0.8078, 0.6009, 0.5862, and 0.5571, respectively. Fig. 8 summarizes the results. First of all, eliminating the most extreme cases, the risk indicators become decreasing with $\alpha$, consistently with the assumption that the GI values can be used as risk indicators. Moreover, as the more extreme GI sizes are eliminated, the profiles of the Sharpe ratio become more regular and increasing.[6] The change occurs when the Gini coefficient approaches 0.5, confirming that the diversification is a relevant feature for asset allocation.

## 6 Conclusion

We investigated the usefulness of the Google search volume data in order to improve the performances of asset allocation. We used the Google Index (GI), i.e. the data freely available from the Google Trends web site, as an indicator of the attention about a given search query. The basic assumption is that the attention of the web concerning a company name is related to a general concern about the company; therefore the GI can be used as an indicator/predictor of the risk related

[6] The qualitative behavior of the indicators for $k = 4, 5, 6$ is similar to the one with $k = 3$. For the sake of interest and space, we do not report the results for such values of $k$, although they are available upon request.

to the stock. We found some evidence about the information contents of the GI when it is used to select the weights of a portfolio. Following the interpretation of the GI as a risk measure, we proposed a renormalization procedure that allowed to obtain a consistent multivariate sample, overcoming the limitations Google imposes on the data disclosure. Finally, a portfolio concentration analysis highlighted that the renormalized GI values may lead to poor performances when the resulting portfolios suffer from a lack of diversification. In fact, we showed that, when the portfolio concentration is kept under control, the use of renormalized GI values leads to an improvement of the risk-adjusted performance of the portfolio, confirming the information contents of web search volumes. Our results, together with the ones found by Kristoufek (2015), Kristoufek (2015), Mondria et al. (2010) and Preis et al. (2014), may suggest that asset managers could add GI volumes to their information set, in order to improve the performances of asset allocation. Consequently, for a practical and operational application, a deeper analysis of technical aspects (such as transaction costs, signal strength, investing horizon, rebalancing frequency, ...) should be performed. Obviously, this work is beyond the scope of this paper, but it is an interesting subject for a subsequent research.

**Acknowledgements** We would like to thank two anonymous Referees for their useful suggestions that helped us to improve the paper.

## References

Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the Relationship Between Financial News and the Stock Market. *Scientific Reports*, 3, 3578.

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107–120.

Bijl, L., Kringhaug, G., Molnr, P., & Sandvik, E. (2016). Google Searches and Stock Returns. *International Review of Financial Analysis*, 45, 150–156.

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *The Journal of Finance*, 66(5), 1461–1499.

Dzielinski, M. (2012). Measuring Economic Uncertainty and its Impact on the Stock Market. *Finance Research Letters*, 9(3), 167–175.

Gini, C. (2005). On the Measurement of Concentration and Variability of Characters. *METRON-International Journal of Statistics*, 63(1), 1–38.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, 457, 1012–1014.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting Consumer Behaviour with Web Search. *Proceedings of the National Academy of Sciences*, 107(41), 1–5.

Heiberger, R. H. (2015). Collective Attention and Stock Prices: Evidence from Google Trends Data on Standard and Poor's 100. *PLoS One*, 10(8), e0135311.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting Abnormal Stock Returns and Trading Volume Using Investor Sentiment: Evidence From Online Search. *International Journal of Forecasting*, 27(4), 1116–1127.

Kristoufek, L. (2013). Can Google Trends Search Queries Contribute to Risk Diversification?. *Scientific Reports*, 3, 2713.

Kristoufek, L. (2015). Power-Law Correlations in Finance-Related Google Searches, and Their Cross-Correlations with Volatility and Traded Volume: Evidence from the Dow Jones Industrial Components. *Phisica A*, 428, 194–205.

Li, X., Ma, J., Wang, S., & Zhang, X. (2015). How does Google Search Affect Trader Positions and Crude Oil Prices? *Economic Modeling*, 49, 162–171.

Liu, Y., Chen, Y., Wu, S., Peng, G., & Lv, B. (2015). Composite Leading Search Index: A Preprocessing Method of Internet Search Data for Stock Trends Prediction. *Annals of Operations Research*, 234(1), 77–94.

Maillard, S., Roncalli, T., & Teïletche, J. (2010). The Properties of Equally Weighted Risk Contribution Portfolios. *Journal of Portfolio Management*, 36(1), 60–70.

Mondria, J., Wu, T., & Zhang, Y. (2010). The Determinants of International Investment and Attention Allocation: Using Internet Search Query Data. *Journal of International Economics*, 82(1), 85–95.

Polgreen, P. M., Chen, Y., & Pennock, D. M. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47, 1443–1448.

Preis, T., Moat, H. S., & Stanley, E. (2014). Quantifying Trading Behavior in Financial Markets Using *Google Trends*. *Scientific Reports*, 3, 1684.

Vlastakis, N., & Markellos, R. N. (2012). Information Demand and Stock Market Volatility. *Journal of Banking & Finance*, 36(6), 1808–1821.

Vozlyublennaia, N. (2014). Investor Attention, Index Performance, and Return Predictability. *Journal of Banking & Finance*, 41, 17–35.