Social Indicators Research (2021) 156:587–599
https://doi.org/10.1007/s11205-019-02250-5

ORIGINAL RESEARCH

Assessing Social Interest in Burnout Using Google Trends Data

Ana Maria Aguilera[1] · Francesca Fortuna[2] · Manuel Escabias[1] · Tonio Di Battista[2]

Accepted: 9 December 2019 / Published online: 14 December 2019

Abstract
Burnout is a serious problem in modern society and early detection methods are needed to successfully handled its multiple effects. The latter refer to working well-being, as well as to the affective, psychological, physiological, and behavioral well-being of workers. However, in many countries official statistics on this topic are not available. For this reason, we propose to use Google Trends data as proxies for the interest in burnout and to analyze them through the functional data analysis approach. The latter allows to address the so-called 'curse of dimensionality' of big data, enabling an effective statistical analysis when the number of variables exceeds the number of observations. Under this framework, the functional analysis of variance (FANOVA) model is used for testing a macro geographic area effect on search queries for the keyword "burnout" in Italy. The estimation of the FANOVA model is proposed in a finite dimensional space generated by a basis function representation. Thus, the functional model is reduced to a MANOVA model on the basis coefficients.

Keywords Burnout · Google trends data · FDA · FANOVA model

1 Introduction

The phenomenon of burnout has emerged as a major social issue in the United States in the mid-1970s, and its importance has grown significantly over the past thirty years in different countries. Burnout is typically defined as a three dimensional syndrome characterised by emotional exhaustion, depersonalization and lack of professional efficacy (Maslach and Jackson 1981). It has been originally identified as a syndrome which affects professions involving interpersonal interactions, such as health-care professionals, social workers and teachers (Maslach and Jackson 1981). However, empirical researches have shown that burnout may concern all employees, irrespective of their professional roles (Leiter and Schaufeli 1996; Maslach et al. 2008). Burnout has been mainly linked to work-related consequences, such as decreased productivity, job dissatisfaction, absenteeism and turnover

✉ Francesca Fortuna
francesca.fortuna@unich.it

[1] Department of Statistics and O.R., University of Granada, Granada, Spain

[2] DISFPEQ, "G. d' Annunzio" University, Pescara, Italy

(Shanafelt et al. 2009; Borritz et al. 2006). However, it has a strong impact not only on working well-being as it inevitably influences the private and social life of individuals. Indeed, burnout can affect health, giving rise to both physical and psychosomatic problems such as depression, anxiety, low self-esteem, guilt feelings, and low tolerance of frustration (Honkonen et al. 2006; Maslach et al. 2001). Moreover, worker reactions to job burnout can be manifested behaviorally and can include pathological dependencies and deviant attitudes such as increased smoking, drinking, aggressiveness or other mental disorders (Suner-Soler et al. 2014). As a consequence, burnout can results in family problems, work-home conflict, and a general reduction in the quality of life (Dyrbye et al. 2011; Lizano 2015). In this context, the role of social support in reducing the negative effects of burnout becomes fundamental, especially under the current situation of crisis in the world of work.

Although the importance of this phenomenon is now recognized, in many continents official statistics on the rates of burnout among workers are not available. The standard measuring instrument for burnout is the Maslach burnout inventory (MBI) (Maslach and Jackson 1986), which is a survey of 22 items with 6 degrees of response on a Likert scale, covering the three burnout dimensions: emotional exhaustion (EE), depersonalization (DP), and low sense of personal accomplishment (PA). Since the analysis of burnout is based on data surveys, its evaluation is expensive, time consuming, often focalized only on some professions, and the results are available with a large time delay.

For these reasons, we propose the use of Google Trends data as proxies for assessing burnout. The basic idea is that internet searches may be considered indicators of the public interest (Wilson et al. 2018). Indeed, in the modern society, big data seems to be a good information base to create reliable proxies of social indicators (Di Bella et al. 2018). Among the many big data sources, Internet traffic plays a fundamental role because it reflects several aspects of daily-life. Indeed, people reveal information about their needs, wants, interests, moods and phycological problems through their Internet search histories, which are stored in the form of Google Trends data (Zeynalov 2017). More specifically, we propose to analyze Google Trends data through the functional data analysis (FDA) approach (Ramsay and Silverman 2005; Ferraty and Vieu 2006) because data flowing from the web can be viewed as an infinite process, which continuously evolve over the time domain (Fortuna et al. 2018). Since functional data are infinite-dimensional objects, they provide a more suitable representation of Google Trends search queries than traditional multivariate vectors. Moreover, FDA allows to address the so-called 'curse of dimensionality' of big data, enabling an effective statistical analysis when the number of variables exceeds the number of observations.

Under this framework, the functional analysis of variance (FANOVA) model has been applied for studying the relationship between the functional queries and an explanatory categorical variable. This issue is usually handled by decomposing the response variable into contributions of the overall mean and the main effects, which are both functional. In particular, the problem of testing the null hypothesis of equality of mean functions across different groups has been considered. In the literature, the FANOVA testing problem has been addressed by several authors according to different methods. A first approach has been proposed by Ramsay and Silverman (2005) by applying standard ANOVA techniques directly on the discretized values of the functions at specified points of the domain. Then, a series of corresponding pointwise F-tests can be performed. A crucial drawback of this approach is that it might cause a loss of knowledge, because it ignores the continuous nature of the parameters. Other approaches are based on dimension-reduction and smoothing techniques (Faraway 1997; Ferraty et al. 2007; Maturo et al. 2018; Di Battista et al. 2016). Different points of view have been

handled in Cuevas et al. (2004) and Shen and Faraway (2004), where, a functional F-statistic based on the analysis of the squared norms has been proposed. On the other hand, Hall and Van Keilegom (2007) have proposed a bootstrap test for detecting differences between two mean functions in a nonparametric regression setting. Finally, other procedures have considered FANOVA modelling from a Bayesian point of view (Behseta et al. 2007; Kaufman and Sainy 2010). In this paper, the estimation of the FANOVA model has been considered in a finite dimensional space generated by a basis. Then, the problem has been reduced to a finite multivariate ANOVA (MANOVA) model on the vector of basis coefficients.

The procedure has been used to evaluate the social interest in burnout over a five year period in Italy and to verify the existence of possible differences in burnout interest among macro geographic areas characterized by different socio-economic conditions. In this application, we have considered free Google Trends data associated with the number of search queries (that is the geographical factor) but additional information stored by big data sources could provide more evidence on social phenomena. Unfortunately, big data are private process-produced data whose full access is not always possible and not free. However, we point out that the main aim of this paper is to provide an original methodological approach for the analysis of social indicators based on big data through the FDA approach.

The remainder of the paper is organized as follows. Section 2 illustrates the materials and methods. In particular, it begins with the analysis of Google Trends data in a functional framework and continues introducing the basis function type approach for the FANOVA model. In addition, it briefly reviews well-known tests for the one-way MANOVA problem, contextualizing them to the functional case. Section 3 shows the main results obtained by applying the proposed approach to a real data-set concerning Google trends data for the search query "burnout" across Italian regions. The Section ends with the description of the R code. Finally, Sect. 4 presents the discussion and conclusions of this study.

2 Materials and Methods

2.1 Google Trends Data in a Functional Framework

Google Trends (http://www.google.com/trends) is a keyword research tool that provides real time trend data regarding interest as operationalised by Internet search volume. It shows how often search terms are entered in Google search engine relative to the total search volume over time since 2004 and across different geographical locations. The search query index does not represent the raw levels of search queries, but rather a relative search volume index, *RSV*, that is:

$$RSV(q, r, t) = \frac{s(q, r, t)}{\sum_{q \in Q(r,t)} s(q, r, t)}, \tag{1}$$

where $s(q, r, t)$ denotes the number of search queries for keyword q in a specific geographical area r at time t, and $Q(r, t)$ is the set of all search queries from location r at time t. *RSV* is then normalized by the highest query share of that term over the time series as follows (Choi and Varian 2012):

$$GRSV(q,r,t) = \frac{RSV(q,r,t)}{max_{t\in\mathcal{T}} RSV(q,r,t)} \times 100 \quad \in [0,100], \tag{2}$$

where $\mathcal{T}$ is the time interval under consideration. The large amount of Google Trends data, the high frequency at which they are produced, their easy and free availability, make them important data sources for official statistics (Glasson et al. 2013) and social indicators (Di Bella et al. 2018). Several researches have revealed the usefulness of Internet search behaviour from Google Trends for short-term economic prediction of unemployment rate, tourism demand, suicide death and health-related topics [see Goel et al. (2010) for a wide review on this topic].

Since Google Trends data continuously flow from the server of a web site, they can be seen as functions in a continuous domain, rather than scalar vectors (Fortuna et al. 2018). Despite the continuous nature of functional data, in real applications, sample curves are observed with error in a discrete set of sampling points, $t_1 < t_2 < \cdots < t_L$ of $\mathcal{T}$. Specifically, let $y_j(t) = \left\{ y_j(t_l) \right\}_{l=1}^{L}$, $j = 1, 2, \ldots, n$, be a functional variable observed in a discrete set of sampling points, $l = 1, 2, \ldots, L$, in the temporal domain $\mathcal{T}$. Let us also assume that $y(t) \in L^2(\mathcal{T})$, where $L^2(\mathcal{T})$ is the Hilbert space of square integrable functions with the usual inner product $<f, g> = \int_{\mathcal{T}} f(t)g(t)\,dt$, $\forall f, g \in L^2(\mathcal{T})$ and the L^2-norm $||f|| = <f,f>^{1/2} < \infty$. Thus, the observed data satisfy the following statistical model:

$$y_{jl} = y_j(t_l) + \epsilon_{jl} \quad l = 1, \ldots, L; j = 1, \ldots, n, \tag{3}$$

with y_{jl} being the observed value for the jth sample path at the sampling point t_l. One usual solution to reconstruct the functional form of the n samples starting from the discrete observations, is to assume that sample paths belong to a finite-dimension space spanned by a basis $\{\phi_1(t), \phi_2(t), \ldots, \phi_K(t)\}$, so that they can be expressed as follows:

$$y_j(t) = \sum_{k=1}^{K} a_{jk}\phi_k(t), \quad j = 1, \ldots, n, \tag{4}$$

or equivalently in matrix notation

$$\boldsymbol{y}(t) = \boldsymbol{A}\boldsymbol{\phi}(t), \tag{5}$$

where $\boldsymbol{y} = [y_1(t), \ldots, y_n(t)]^T$, $\boldsymbol{A} = (a_{jk})$ is the matrix of basis coefficients, and $\boldsymbol{\phi}(t) = [\phi_1(t), \cdots, \phi_K(t)]^T$ is a K dimensional vector of basis functions. The basis coefficients may be fitted by least squares approximation with B-splines basis (De Boor 2001). A comparative study of different B-spline approaches for functional data was developed in Aguilera and Aguilera-Morillo (2013).

2.2 The FANOVA Model with Regularized Basis Expansions

Let $\{y_{ij}(t) : t \in \mathcal{T}, i = 1, \ldots, I; j = 1, \ldots, n_i\}$ be I independent samples of functions drawn from a second order stochastic processes $Y = \{Y(t) : t \in \mathcal{T}\}$, continuous in quadratic mean, whose sample functions belong to the Hilbert space $L^2(\mathcal{T})$ of square integrable functions. Assuming that there is a single factor with I different levels or groups ($i = 1, 2, \ldots, I$) and

n_i observations within each group; the model for the jth observation ($j = 1, 2, \ldots, n_i$) in the ith group can be expressed as follows:

$$y_{ij}(t) = \mu(t) + \gamma_i(t) + \epsilon_{ij}(t) \qquad j = 1, 2, \ldots, n_i; \;\; i = 1, 2, \ldots, I, \tag{6}$$

where $\mu(t)$ is the grand mean function; $\gamma_i(t)$ represents the specific functional effect of being in the ith group, subject to the sum to zero constraint for their unique identification:

$$\sum_{i=1}^{I} \gamma_i(t) = 0 \quad \forall t \in \mathcal{T}, \tag{7}$$

$\epsilon_{ij}(t)$ is the residual function, which expresses the unexplained variation specific to the jth observation within the ith group, and $\sum_{i=1}^{I} n_i = n$. The model in (6) can be written in matrix notation as follows:

$$\boldsymbol{y}(t) = \boldsymbol{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \tag{8}$$

where $\boldsymbol{y}(t) = [y_1(t), y_2(t), \ldots, y_n(t)]^T$ is a vector of functional observations of length n; $\boldsymbol{\beta}(t) = [\beta_1(t) = \mu(t), \beta_2(t) = \gamma_1(t), \ldots, \beta_Q(t) = \gamma_I(t)]^T$ is a vector of functional effects of length $Q = I + 1$; $\boldsymbol{\epsilon}(t) = [\epsilon_1(t), \epsilon_2(t), \ldots, \epsilon_n(t)]^T$ is a vector of n residual functions and $\boldsymbol{Z}$ is a $(n \times Q)$ design matrix, coding the group membership. In particular, the first column of $\boldsymbol{Z}$ consists entirely of ones to represent the overall mean, whereas the other columns represent the group membership with values one if the jth observation belongs to the ith group and zeros otherwise. The condition in (7) for the unique identification of the functional effects, is implemented in model (8) by adding a row to the matrix $\boldsymbol{Z}$ representing an additional observation for which $y_{n+1}(t) = 0, \forall t \in \mathcal{T}$ (Ramsay and Silverman 2005).

The FANOVA model is equivalent to a standard ANOVA model, with the difference that the parameters $\boldsymbol{\beta}(t)$, and hence the predicted observations $\widehat{\boldsymbol{y}}(t) = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}(t)$, are functions rather than vectors of numbers.

The parameter vector $\boldsymbol{\beta}(t)$ in Eq. (8) can be estimated using the standard least squares criterion; thus, minimizing the residual sum of squares:

$$LMSSE(\boldsymbol{\beta}) = \int [\boldsymbol{y}(t) - \boldsymbol{Z}\boldsymbol{\beta}(t)]^T [\boldsymbol{y}(t) - \boldsymbol{Z}\boldsymbol{\beta}(t)]\, dt, \tag{9}$$

subject to the constraint (7). If there are no particular restrictions on the way in which $\boldsymbol{\beta}(t)$ varies as a function of t, it is possible to minimize the discrete version of (9) individually for each $t \in \mathcal{T}$, obtaining pointwise least squares estimates of the functional parameters (Ramsay and Silverman 2005):

$$\widehat{\boldsymbol{\beta}}(t) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}(t) \quad \forall t \in \mathcal{T}. \tag{10}$$

However, this pointwise minimization might result in losing some potential useful information because it ignores the continuous nature of the parameters.

To fit the model (8), it is usual to assume that the sample paths and the parameter functions belong to the same finite space. That is,

$$y_{ij} = \sum_{k=1}^{K} a_{ijk}\phi(t), \qquad \beta_q(t) = \sum_{k=1}^{K} b_{qk}\phi_k(t), \tag{11}$$

so that the vector of functional parameters is given by $\boldsymbol{\beta}(t) = \boldsymbol{B}\boldsymbol{\phi}(t)$ where $\boldsymbol{B} = (b_{qk})$ is the $Q \times K$ matrix of basis function coefficients. In this context, the least squares fitting criterion in (9) can be defined as follows:

$$LMSSE(\boldsymbol{\beta}) = \int \left[\boldsymbol{A}\boldsymbol{\phi}(t) - \boldsymbol{Z}\boldsymbol{B}\boldsymbol{\phi}(t)\right]^T \left[\boldsymbol{A}\boldsymbol{\phi}(t) - \boldsymbol{Z}\boldsymbol{B}\boldsymbol{\phi}(t)\right] dt, \tag{12}$$

which leads to the following estimation of the functional effects:

$$\widehat{\boldsymbol{B}}(t) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{A}, \tag{13}$$

where $\boldsymbol{A}$ has an additional row of zeros to satisfy the constraint on the functional effects (Sayes et al. 2008). Therefore, the one-way FANOVA model has been transformed into a one-oy MANOVA model on the matrix $\boldsymbol{A}$ of vectors of basis coefficients of the sample curves.

2.3 The FANOVA Testing Problem

The FANOVA testing problem aims to verify the possible differences in functional responses according to various treatment conditions. This problem is known as the I-sample testing problem or the one-way ANOVA problem for functional data and can be expressed as follows:

$$H_0 : \gamma_1(t) = \gamma_2(t) = \cdots = \gamma_I(t) = 0, \quad t \in \mathcal{T}, \tag{14}$$

against the alternative that its negation holds for at least one t and $i \neq i'$.

Similarly to the case of MANOVA tests for multivariate variables, the functional tests are based on the within-subject and between-subject variations given by the following matrices $\boldsymbol{E}$ and $\boldsymbol{H}$, respectively:

$$\begin{aligned} \boldsymbol{E} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\boldsymbol{a}_{ij} - \bar{\boldsymbol{a}}_i\right)\left(\boldsymbol{a}_{ij} - \bar{\boldsymbol{a}}_i\right)^T dt, \\ \boldsymbol{H} &= \sum_{i=1}^{I} n_i\left(\bar{\boldsymbol{a}}_i - \bar{\boldsymbol{a}}\right)\left(\bar{\boldsymbol{a}}_i - \bar{\boldsymbol{a}}\right)^T dt, \end{aligned} \tag{15}$$

where $\bar{\mathbf{a}}$ and $\overline{\mathbf{a}}_i$ are the corresponding unbiased estimators of the grand mean vector and the group mean vector associated with the vectors of basis coefficients of each sample curve $\boldsymbol{a}_{ij} = (a_{ij1}, \ldots, a_{ijK})'$. Let us observe that, under the basis function representation of the sample curves, the sample group mean and the grand mean functions are estimated as

$$\begin{aligned} \bar{\boldsymbol{y}}_i(t) &= \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{y}_{ij}(t) = \overline{\boldsymbol{a}}_i'\boldsymbol{\phi}(t), \\ \bar{\boldsymbol{y}}(t) &= \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\boldsymbol{y}_{ij}(t) = \overline{\boldsymbol{a}}'\boldsymbol{\phi}(t). \end{aligned} \tag{16}$$

In this context, under the usual requirements for MANOVA, the following well-known test statistics for the classical MANOVA problem (Anderson 2003) can be used:

– The Wilks's lambda test statistic:

$$W = \frac{det(\boldsymbol{E})}{det(\boldsymbol{E}+\boldsymbol{H})} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}, \tag{17}$$

where s is the number of nonzero eigenvalues λ_i of the matrix $\boldsymbol{HE}^{-1}$.

– The Lawley–Hotelling's trace test statistic:

$$LH = trace\left(\boldsymbol{HE}^{-1}\right) = \sum_{i=1}^{s} \lambda_i. \tag{18}$$

– The Pillai's trace test statistic:

$$P = tr(\boldsymbol{H}\left(\boldsymbol{H}+\boldsymbol{E}\right)^{-1}) = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}. \tag{19}$$

– The Roy's maximum root test statistic:

$$R = \lambda_{max}\left(\boldsymbol{HE}^{-1}\right), \tag{20}$$

where $\lambda_{max}(\boldsymbol{HE}^{-1})$ denotes the maximum eigenvalue of a matrix $\boldsymbol{HE}^{-1}$.

There is enough evidence to reject the null hypothesis when W is close to zero and LH, P and R assume increasing values.

3 Applications: Functional Google Queries for the Keyword "Burnout" in Italy

In this section, Google Trends data have been used to evaluate the social interest in burnout argumentation among different macro geographic areas of Italy. Weekly search volumes for the terms "burnout" have been obtained from the website Google Trends (trends.google.com) by considering searches made in Italy. Search terms based on the corresponding national language (that is "stress da lavoro") and related terms suggested by Google Trend ("sindrome da burnout" and "tunnel del burnout") have been also examined. In both cases, search frequencies are closed to zero; thus only the keyword "burnout" has been considered. Data have been collected across a five year period, starting from January 2005 until May 2019 (for a total of 230 weeks) using the "Occupational Health and Safety" category and considering the twenty Italian regions. The region of Aosta Valley has the only one with zero frequencies for the search term under consideration, hence it has been removed from the study. We remark that Google Trends provides only data for popular queries and sets a value of zero if the search query volume falls below an unreported privacy threshold. Thus, a value of zero does not represent the absence of search for a specific keyword in a geographical area but low search volumes. Figure 1 shows the Google Trends series for the Italian regions, observed from February 1st 2019 to March 31st 2019. We remark that the number of queries for the term "burnout" is rescaled to a value between 0 and 100 as in Eq. (2), with 100 corresponding to the peak of relative search volume obtained for the keyword during the period of interest. Most regions show peaks of maximum interest in the 12th week of 2015 (2015-03-22); Trentino Alto Adige and Friuli Venezia Giulia have a value of 100 in December and October 2015 (2015-12-06 and 2015-10-04, respectively); Umbria at the beginning of 2016 (2016-01-31); Molise is distinguished by two peaks in June and July

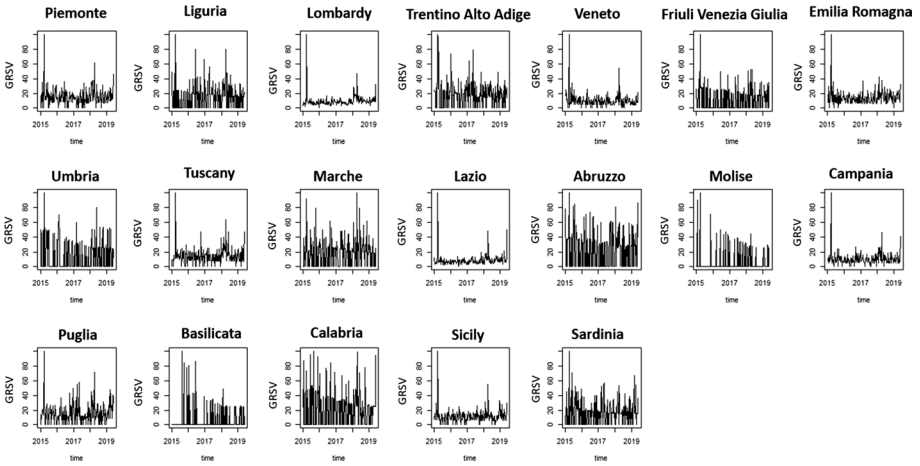

Fig. 1 Raw Google Trends data for the keyword "burnout" across Italian regions

2016 (2016-06-05 and 2016-07-31); while Abruzzo and Calabria have the maximum interest in the second week of April 2018 (2018-04-08).

Google Trends time series observed for each Italian region can be considered as raw functions observed in a discrete set of sampling points, that is 230 weekly observations in the period under consideration. Thus, the raw-data have been arranged in a (19×230) matrix; and then they have been converted into a sample of functions adopting a B-splines basis expansion as in Eq. (5). Specifically, the basis coefficients have been obtained by least square approximation with ten cubic B-splines basis, chosen by cross validation. Figure 2 shows the reconstructed functional queries, whereas Figure 3 displays the functional queries for each region individually. We can note that, in the first part of the domain (the first 50 weeks), Trentino Alto Adige and Friuli Venezia Giulia show the greatest interest; while Molise and Basilicata present the opposite attitude. In the last part of the domain, Abruzzo, Marche and Sardinia have the highest peaks, while Molise continues to show the lowest number of search queries. To assess whether different socio-economic characteristics have an effect on the interest in burnout, the Italian regions have been arranged according to the Istat classification into three macro areas: North, Centre and South. The latter present a socio-economic level that decreases passing from the North to the South of the country. Hence, we are dealing with a one-way ANOVA problem for functional data with

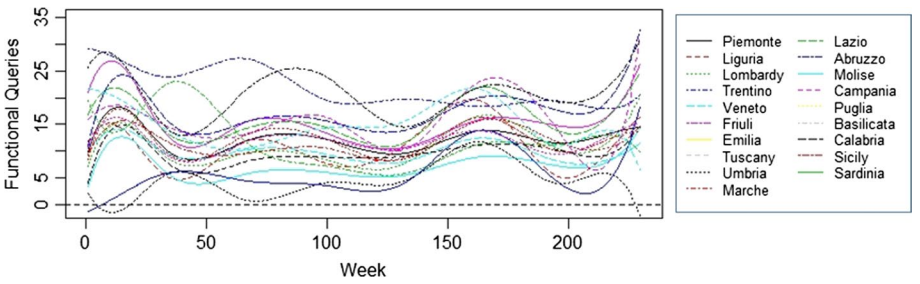

Fig. 2 Functional queries for "burnout" in the Italian regions

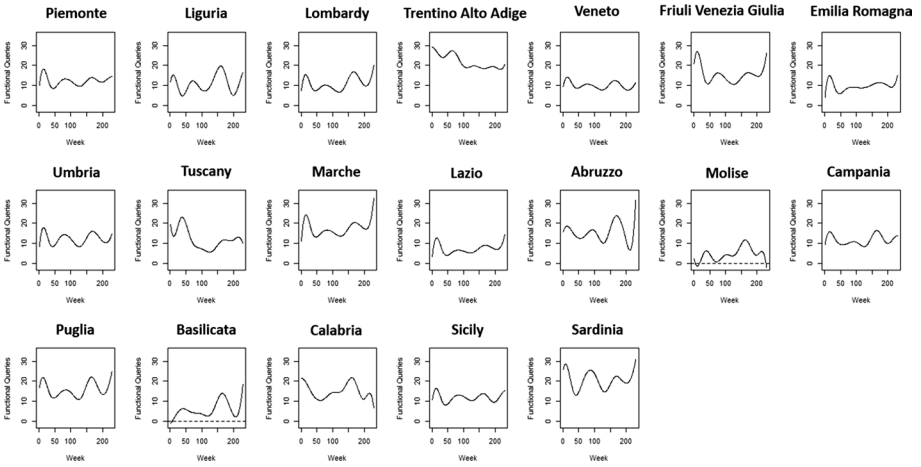

Fig. 3 Functional queries for "burnout" for each Italian region

$I = 3$ groups with sample sizes equal to $n_{North} = 7$ (excluding the region of Aosta Valley), $n_{Centre} = 4$ and $n_{South} = 8$. Figure 4 displays the estimated functional effects for the three macro areas and the intercept term, which represents the Italian mean. It can be noted that the South area has a negative effect on the interest in burnout; contrary to the North area, which has a positive effect on burnout Internet research for almost the entire domain except for the last weeks. Also the effect of the Center area is mainly positive. Table 1 provides the results of all the tests introduced in Sect. 2. The p values of the tests are all greater than the significance level 0.05, hence it can be concluded that the macro geographic areas do not have an effect on the mean functional query for the keyword 'burnout".

3.1 Implementation in the R Environment

In the following, we describe the implementation of the proposed method in the R program. Available functions of different R libraries have been used.

- Google Trends data have been obtained with the function "gtrends" of the R package "gtrendsR";
- Google Trends data have been arranged in a $n \times L$ matrix of raw observations, where n is the number of the statistical units and T the number of points of the temporal domain;
- The basis functions representation of the sample paths has been obtained with the function "create.bspline.basis" of the "fda" package;
- Raw data have been converted into a functional object through the function "Data2fd" of the "fda" package;
- The matrix of basis coefficients $\boldsymbol{A}$ has been obtained by means of the value 'coefs' provided by the function "Data2fd" of the "fda" package;
- MANOVA on the B-splines coefficients has been computed by fitting a multivariate linear model with the "lm" function of the "stats" package where the input data are the matrix of basis coefficients $\boldsymbol{A}^T$ of dimension $n + 1 \times K$ and a single column with category labels (corresponding to the samples) both stored in a data frame;

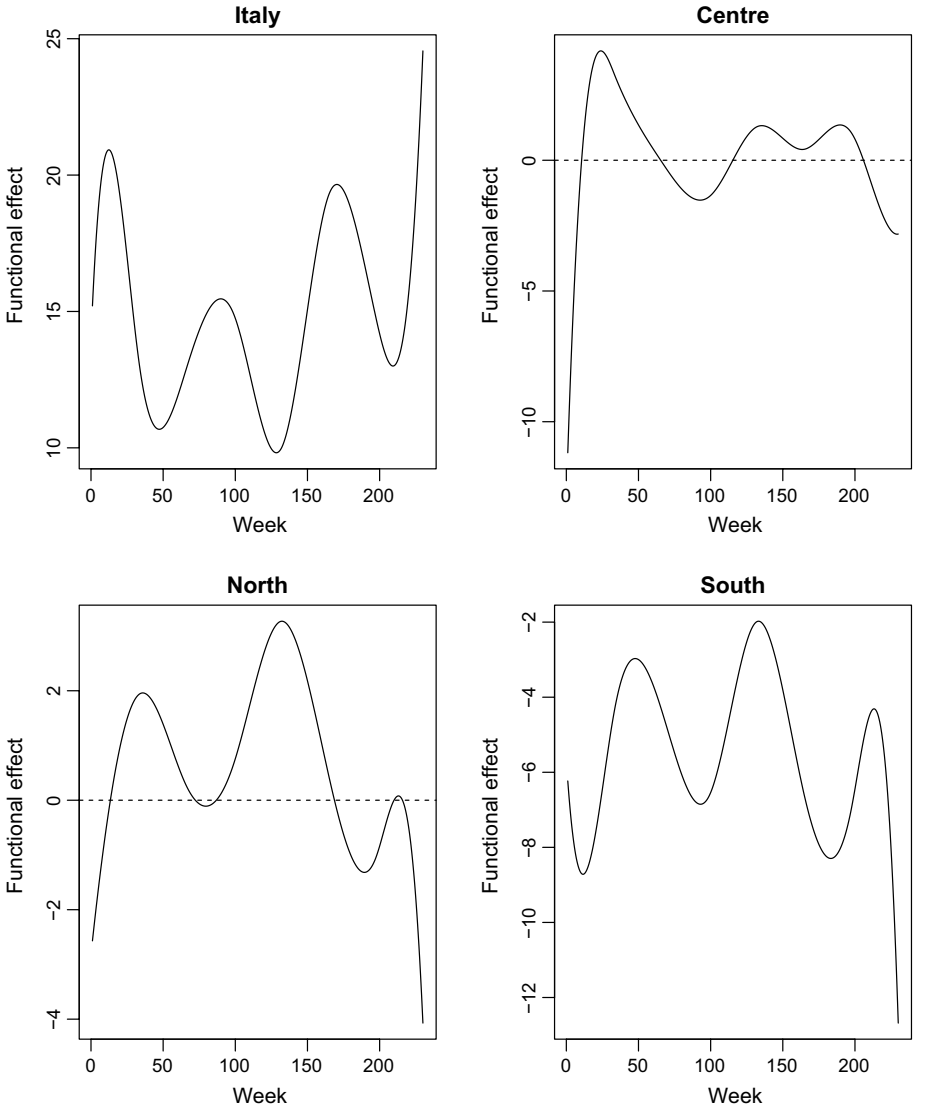

Fig. 4 Estimated functional effects for the three macro geographic areas and the overall mean function

Table 1 Values of the test statistics and p values of all tests for the GT data

Test	Test statistic	Approx F	p value
W	0.1405	0.6731	0.8433
LH	3.8377	0.7249	0.7854
P	1.2085	0.6071	0.9073
R	3.2123	2.8911	0.0627

– The estimates of the functional parameters have been obtained by creating a functional data object with the "fd" function of the "fda" package where the input arguments are the value of the coefficients provided by the "lm" function and the same basis of the sample paths;
– Different multivariate test statistic can be implemented with the function "manova" and the "summary" method of the "stats" package.

4 Discussions and Conclusions

Burnout is a growing problem in the modern society. It is usually thought of as an individual response to prolonged work related stress, which in turn, impacts on job satisfaction and thereafter, can affect the phycological, physiological, affective and behavioral well-being of workers (Suner-Soler et al. 2014; Dyrbye et al. 2011; Lizano 2015). The estimation of this phenomenon is essential to design social support for reducing its negative effects. However, in many countries, official statistics for the rates of burnout are not available. In this context, we propose the use of Google Trends data as proxies for the interest in burnout. Due to the popular use of the Internet search, people tend to seek data or information from the Internet and express opinions, moods and personal matters on social networks. Thus, keyword-driven search results of the Internet seems to be a good information base to create reliable proxies of social indicators (Di Bella et al. 2018).

In this scenario, we aim to provide an original methodological approach for the analysis of social indicators based on big data, through the FDA approach. The latter has the advantage of reducing the dimension of the huge amount of data with the conversion of vectors into functions. Under this framework, the FANOVA model can be used for testing a possible effect of different factors on the search queries. Focusing on free Google Trends data, this paper investigates the effect of different Italian macro areas on functional queries for burnout. Of course, additional information stored by big data sources could provide more evidence on social phenomena, but access to them is not free. Our results have shown that Italian macro geographic areas have not an effect on the Google search query "burnout". This result confirms the evidences found in the 6th European Working Conditions Survey (Eurofound 2016), which investigates the phenomenon of burnout in a random samples of workers from thirty-five European countries (for a total of 43,675 workers). To our knowledge, it is the first survey that analyzes differences in burnout across Europe using representative national samples. The 6th European Working Conditions Survey illustrates that burnout may not only be studied at the individual, psychological level, but also at the collective, national level as it relates in meaningful ways with various economic and socio-cultural indicators.

With reference to our results, it should be emphasized however that the three Italian macro geographic groups are characterized by small sample sizes. Of course, this aspect limits the statistical power of any test. Nevertheless, the Pillai's trace test statistic, P, is considered the most reliable of the multivariate measures and offers the greatest protection against Type I errors with small sample sizes (Seber 1984). In this regard, a future development of our approach could be to consider non parametric and bootstrap version of test statistics for the FANOVA problem to assess small sample size situations.

Although Google Trends data have been recognized as a valuable tool for official statistics and social indicators, they present several challenges that need to be addressed (Glasson et al. 2013). Firstly, Internet data are generated by processes not primarily aimed at

data collection, hence they do not have well-defined target population and they are often representative of particular segments of the population. Secondly, they generally are private process-produced data whose access by national statistical offices is rarely possible although the intrinsic value of the information contained in big data has a social importance that should be shared with the whole community (Di Bella et al. 2018). Thirdly, the number of queries for a given term is rescaled to a value between 0 and 100, potentially weakening the value of Google data in modelling, as the actual number of searches is not provided. Finally, the high frequency at which these data are available makes hard to apply standard multivariate techniques without first transforming the data. Therefore big data may be a big opportunity for the definition of traditional or new social indicators but their statistical reliability should be further investigated and their availability and use should be internationally coordinated.

Acknowledgements The authors Ana M. Aguilera and M. Escabias thank the support of the Spanish Ministry of Science, Innovation and Universities under project MTM2017-88708-P (also supported by the FEDER program).

References

Aguilera, A. M., & Aguilera-Morillo, M. C. (2013). Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling*, *58*(7–8), 1568–1579.

Anderson, T. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). London: Wiley.

Behseta, S., Kass, R., Moorman, D., & Olson, C. (2007). Testing equality of several functions: Analysis of single-unit firing-rate curves across multiple experimental conditions. *Statistics in Medicine*, *26*, 3958–3975.

Borritz, M., Rugulies, R., Christensen, K., Villadsen, E., & Kristensen, T. (2006). Burnout as a predictor of self-reported sickness absence among human service workers: Prospective findings from three year follow up of the PUMA study. *Occupational and Environmental Medicine*, *63*, 98–106.

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

Cuevas, A., Febrero, M., & Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, *47*(1), 111–112.

De Boor, C. (2001). *A practical guide to splines* (revised ed.). New York: Springer.

Di Battista, T., Fortuna, F., & Maturo, F. (2016). Parametric functional analysis of variance for fish biodiversity assessment. *Journal of Environmental Informatics*, *28*, 101–109.

Di Bella, E., Leporatti, L., & Maggino, F. (2018). Big data and social indicators: Actual trends and new perspectives. *Social Indicators Research*, *135*(3), 869–878.

Dyrbye, L., Shanafelt, T., Balch, C., Satele, D., Sloan, J., & Freischlag, J. (2011). Relationship between work-home conflicts and burnout among American surgeons: A comparison by sex. *Archives of Surgery*, *146*, 211–217.

Eurofound, (2016). *Sixth European Working conditions survey—Overview report*. Luxembourg: Publications Office of the European Union.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, *39*, 254–261.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis*. New York: Springer.

Ferraty, F., Vieu, P., & Viguier-Pla, S. (2007). Factor-based comparison of groups of curves. *Computational Statistics & Data Analysis*, *51*, 4903–4910.

Fortuna, F., Maturo, F., & Di Battista, T. (2018). Clustering functional data streams: Unsupervised classification of soccer top players based on Google trends. *Quality and Reliability Engineering International*, *34*, 1448–1460.

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M.,& Khan, A. (2013). What does "Big Data" mean for Official Statistics? In *Paper for the high-level group for the modernization of statistical production and services*, March 10.

Goel, S., Hofman, J., Lahaie, S., Pennock, D., & Watts, D. (2010). Predicting consumer behavior with web search. In S. Levin (Ed.), *Proceedings of the National academy of sciences, volume 107 of 41* (pp. 17486–17490). New York: Princeton University, National Academy of Sciences.

Hall, P., & Van Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, *17*, 1511–1531.
Honkonen, T., Ahola, K., Pertovaara, M., Isometsa, E., Kalimo, R., & Nykyri, e a E. (2000). The association between burnout and physical illness in the general population—Results from the finnish health 2000 study. *Journal of Psychosomatic Research*, *61*(59–66), 2006.
Kaufman, C., & Sainy, S. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, *5*(1), 123–150.
Leiter, M., & Schaufeli, W. (1996). Consistency of the burnout construct across occupations. *Anxiety, Stress, and Coping*, *9*, 229–243.
Lizano, E. (2015). Examining the impact of job burnout on the health and well-being of human service workers: A systematic review and synthesis. *Human Service Organizations: Management, Leadership & Governance*, *39*, 167–181.
Maslach, C., & Jackson, S. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, *2*, 99–113.
Maslach, C., & Jackson, S. (1986). *Maslach burn-out inventory*. Palo Alto: Consulting Psychologists Press.
Maslach, C., Leiter, M., & Schaufeli, W. (2008). Measuring burnout. In C. Cooper & S. Cartwright (Eds.), *The Oxford handbook of organizational well-being* (pp. 86–108). Oxford: The Oxford Handbook.
Maslach, C., Schaufeli, W., & Leiter, M. (2001). Job burnout. *Annual Review of Psychology*, *52*, 397–422.
Maturo, F., Fortuna, F., & Di Battista, T. (2018). Testing equality of functions across multiple experimental conditions for different ability levels in the IRT context: The case of the IPRASE TLT 2016 survey. *Social Indicators Research*, *146*, 19–39. https://doi.org/10.1007/s11205-018-1893-4.
Ramsay, J., & Silverman, B. (2005). *Functional data analysis* (2nd ed.). New York: Springer.
Sayes, W., De Ketelaerea, B., & Dariusa, P. (2008). Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics*, *22*, 335–344.
Seber, G. (1984). *Multivariate observations*. New York: Wiley.
Shanafelt, T., Balch, C., Bechamps, G., Russell, T., Dyrbye, L., Satele, D., et al. (2009). Burnout and career satisfaction among American surgeons. *Annals of Surgery*, *250*, 463–471.
Shen, Q., & Faraway, J. (2004). An F test for linear models with functional responses. *Statistical Sinica*, *14*, 1239–1257.
Suner-Soler, R., Grau-Martín, A., Flichtentrei, D., Prats, M., Braga, F., Font-Mayolas, S., et al. (2014). The consequences of burnout syndrome among healthcare professionals in Spain and Spanish speaking Latin American countries. *Burnout Research*, *1*, 82–89.
Wilson, S., Daar, D., Sinno, S., & Levine, S. (2018). Public interest in breast augmentation: Analysis and implications of google trends data. *Aesthetic Plastic Surgery*, *42*(3), 648–655.
Zeynalov, A. (2017). *Forecasting tourist arrivals in prague: Google econometrics*. Mpra paper, University Library of Munich, Germany.