Big Data Research 30 (2022) 100354

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Augmented Functional Analysis of Variance (A-fANOVA): Theory and Application to Google Trends for Detecting Differences in Abortion Drugs Queries

Fabrizio Maturo [a,*], Annamaria Porreca [b]

[a] University of Campania Luigi Vanvitelli, Department of Mathematics and Physics, Viale Abramo Lincoln, 5, 81100 Caserta CE, Italy
[b] Department of Medical, Oral and Biotechnological Sciences, University "G. d'Annunzio" Chieti-Pescara, Chieti, Via dei Vestini 31, Chieti, 66100, Italy

ARTICLE INFO

Article history:
Received 20 March 2022
Received in revised form 14 September 2022
Accepted 7 October 2022
Available online xxxx

Keywords:
Augmented functional analysis of variance
A-fANOVA
Google Trends
Functional means difference
Cytotec
Abortion

ABSTRACT

The World Wide Web (WWW) has become a popular and readily accessible big data source in recent decades. The information in the WWW is offered in many different types, e.g. Google Trends, which provides deep insights into people's search queries in the Google Search engine. Analysing this kind of data is not straightforward because they usually take the form of high-dimensional data, given that the latter can be collected over extensive periods. Comparing Google Trends' means of different groups of people or Countries can help understand many phenomena and provide very appealing insights into populations' interests in specific periods and areas. However, appropriate statistical techniques should be adopted when inspecting and testing differences in such data due to the well-known curse of dimensionality. This paper suggests an original approach to dealing with Google Trends by concentrating on the search for the "*Cytotec*" abortion drug. The final purpose of the application is to determine if different Countries' abortion legislation can influence the research trends. This research focuses on Functional Data Analysis (FDA) to deal with high-dimensional data and proposes a generalisation of the classical functional analysis of variance model, namely the Augmented Functional Analysis of Variance (A-fANOVA). To test the existence of statistically significant differences among groups of Countries, A-fANOVA considers additional curves' characteristics provided by the velocity and acceleration of the original google queries over time. The proposed methodology appears to be intriguing for capturing additional information about curves' behaviours with the final aim of offering a monitoring tool for policy-makers.

1. Introduction

The World Health Organization (WHO) named the alternative methods to surgical abortion "medical abortions" or sometimes "non-surgical abortions". WHO included the association of Misoprostol and Mifepristone in its list of essential medicines for abortion, and in particular, Misoprostol is one of the most used approaches for the treatment of incomplete abortion and postpartum bleeding [1–4]. Although abortion is illegal in some countries, many people try to terminate their pregnancy outside of a controlled clinical environment [1–3]. Access to abortion is conditioned both by ethical-religious, legislative and socio-economic factors and by a lack of information and low level of education. For these reasons, many people are consulting the World Wide Web (WWW) to look for information on self-abortion.

Recent research suggests that it is effortless to purchase Mifepristone and Misoprostol on the WWW; these are usually used in early abortions in controlled clinical settings. In the practice of "non-surgical abortions", the most common abortion drugs are Misoprostol, mainly in the form of the trade name "*Cytotec*", and Mifepristone, known under the trade name "*RU*486". Depending on the country and its legislation, these drugs are used alone or in combination based on supply conditions, availability, and use, with consequences on treatments' effectiveness [5]. In recent years, the use of drugs to practice abortion has expanded considerably, both in Countries where it is legal and not. Where access to abortion is limited, medical abortion has become widespread, and Cytotec has acquired a certain notoriety for its abortion properties, whether administered orally or vaginally [6]. Even when health authorities do not officially allow Cytotec, women show interest in seeking information about self-abortion through informal

* Corresponding author.
E-mail addresses: fabrizio.maturo@unicampania.it (F. Maturo), annamaria.porreca@unich.it (A. Porreca).

https://doi.org/10.1016/j.bdr.2022.100354
2214-5796/

networks or online pharmacies, mainly where the Misoprostol is sold for use related to other conditions. In Brazil, after its arrival on the market in 1986, the demand for Cytotec has undergone a drastic expansion [7]. Moreover, since 1988, it has been more used for abortion than for treating ulcers, i.e. its primary indication. Since 1991, the Brazilian government has limited the sale of Cytotec to reduce its use as an abortion drug, allowing it to be sold only on prescription or given in specifically authorised locations. As a result of this policy, there has been an effective decrease in official Cytotec sales but also a disproportionate price increase in parallel markets such as the Dark Web [7]; indeed, recent research found that, in Brazil, Cytotec is still the primary method of abortion [8].

Furthermore, the use of Cytotec has become widespread also in other counties where applies restrictive legislation on abortion, such as in Latin American Countries [9], Asian countries such as the Philippines [10], and in Africa, e.g. Uganda, Ghana, and Nigeria [11–13]. It has also gained ground in Spain and Italy through Latin American migrants who have spread the word about its abortifacient properties [14]. In contexts where abortion is legal, such as Uruguay, France, Germany and others, medical abortion replaces surgery-based methods and is now the primary adopted method [15]. In some countries, such as France, medical abortion is under the strict control of health professionals, with two medical appointments necessary to obtain the drug. In contrast, in other countries, the practice is less regulated.

In the era of Big Data, Internet search data can provide valuable information on disease models and population behaviour [16]. A tool that allows users to interact with Internet search data is Google Trends, a service accessible for free online. Google Trends analyses a portion of Google's searches and provides information on geospatial and temporal data in search volumes for user-specified terms [17]. Google Trends is a tool that lets users graph searches for a defined word or string of multiple words or a sentence. The output data used to generate trends are resized to the average search traffic for the selected term and normalised on a relative basis. These charts can also be manipulated to limit specific time intervals and geographic regions.

Since Google provides an enormous amount of data over time, traditional statistical techniques fail where a limited number of statistical units and many variables are present, i.e. a large number of temporal observations, in this case. The latter issue is known as the curse of dimensionality.

For this reason, the use of Functional Data Analysis (FDA) has recently been proposed to investigate this type of data that is comparable to the so-called data-streaming, which is characterised by data that flows continuously and in large quantities. The storage of these data types is very complicated, and thus the need to analyse them with sophisticated techniques arises.

This study proposes an original approach to learning the dynamics of "*Cytotec*" Google Trends worldwide. FDA solves the curse of dimensionality concern and analyses the query trends as a functional object. Moreover, the Functional Analysis of Variance (fANOVA) is proposed as a tool to understand if groups' functional means are statistically different. Furthermore, this study suggests an extension of the classical fANOVA approach by augmenting information on curves' characteristics through additional insights such as the velocity and acceleration of the queries over time via their derivatives, namely Augmented Functional Analysis of Variance (A-fANOVA).

The article is structured as follows. Section "Materials and Methods" presents the basic information about google trends data, functional data analysis, the main fANOVA approaches proposed in the statistical literature, and the A-fANOVA approach. Section "Application and Results" shows the findings of the proposed method extended to the "*Cytotec*" Google Trends and a comparison using other one-way fANOVA strategies. The last Section illustrates the discussion and conclusions.

2. Material and methods

2.1. Google Trends data

Google Trends is an investigation instrument that offers real-time data represented in a selected period regarding interest in a research query volume in a specific geographical location. In particular, Google Trends display how often search terms are entered in Google regarding the total search volume. The search query index does not depict the total number of search queries but rather a relative search volume index that is given by:

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

where $s(q,r,t)$ is the number of search queries for the specific keyword q in a particular geographical area r at time t, and $Q(r,t)$ is the set of all the search queries from location r at time t.

RSV is then normalised by the highest query share of that term, over the time series, as follows:

$$\mathrm{GRSV}(q,r,t) = \frac{\mathrm{RSV}(q,r,t)}{\max_{t\in[1,\dots,T]} RSV(q,r,t)} \times 100 \in [0,100] \tag{2}$$

2.2. Google Trends data in a functional framework

When we analyse this data, the fundamental difficulty is that we could run into the curse of dimensionality if we focus on a considerable time interval. FDA [18] is one of the most used methodologies to solve the latter problem because, in this context, the whole function becomes the reference object. In general, FDA addresses issues in which the observations are functions rather than vectors and can be represented as curves [18]. Because Google Trends data constantly flow, they can be expressed as functions in a continuous domain rather than scalar observations. Thus FDA is the most natural approach to deal with this data type and reduce its dimensionality.

Despite the continuous nature of functional data, in real applications, sample curves are observed with errors in a discrete set of sampling points, $t_1 < t_2 < \cdots < t_L$ of T. Specifically, let $y_i(t) = \{y_i(t_l)\}_{l=1}^{L}$, $i = 1,2,\dots,n$, be a functional variable observed in a discrete set of sampling points, $l = 1,2,\dots,L$, in the temporal domain $\mathcal{T}$. We 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 $\langle y,g\rangle = \int_\tau y(t)g(t)dt$, $\forall y,g \in L^2(\mathcal{T})$ and the L^2-norm $\|y\| = \langle y,y\rangle^{1/2} < \infty$. Thus, the observed data satisfy the following statistical model:

$$y_{il} = y_i(t_l) + \varepsilon_{jl} \quad l = 1,\dots,L;\ i = 1,\dots,n \tag{3}$$

with y_{il} being the observed value for the i-th sample at the sampling point t_l.

The first step in FDA is to convert the original data into functions using a suitable smoothing technique. In other words, we need a mathematical description of a curve with some flexibility because we usually do not know in advance how complex the phenomenon will be. Thus, choosing a set of basic functional building blocks that can be stacked is the starting point in FDA. The simplest way to reconstruct curves is via polynomials, but they are not very flexible. For this reason, many basis functions have been proposed in the FDA literature. Depending on the characteristics of the curves, various basis systems can be adopted [18]; however, the usual basis in L_2 can be separated into two main groups:

fixed basis and data-driven basis. The most common fixed bases are: b-spline functions, i.e. set of polynomials of order m defined in subintervals and often adopted for non-periodic data; fourier basis, i.e. functions consisting of a series of pairs of sines and cosines of increasing frequency and usually implemented for periodic data; piecewise constant functions that can be used for counting processes; and wavelets bases, i.e. a basis system based on translation and/or dilatation of a function called mother wavelet and often used to deal with curves with strong local behaviours.

On the other hand, a data-driven basis is not fixed in advance and depends on the data. The most used strategies are Functional Principal Components (FPCs) which optimise the amount of variability in the data, and Functional Partial Least Squares (FPLS), which optimise the amount of explanation between the data and response variable.

Let $y_i(t) \in L_2$, a basis system is a set of K known functions $\theta_j(t)$ linearly independent. Hence, a function $y_i(t)$ can be expressed by a linear combination of these basis functions as follows:

$$y_i(t) = \sum_{j=1}^{K} c_{ij}\theta_j(t) \quad j = 1, 2, \ldots, K \tag{4}$$

where c_{ij} is the coefficient defining the i-th curve via the j-th basis function, and $\theta_j(t)$ is the j-th basis function. The basis coefficients can be obtained from the discrete observations either by interpolation (when data are observed without error) or least square approximation. This study focuses on a linear combination of b-splines and least square approximation.

2.3. The functional analysis of variance (fANOVA)

Starting from the original dataset, each group (treatment) identified according to a categorical variable (factor) can be represented by a set of curves defined via the b-spline linear combination in Equation (4). In this framework, to quantify the effects exerted at multiple levels on the functional observation by some factors, the functional analysis of variance (fANOVA) can be used [18–34].

In the last twenty years, the statistical literature has been very lively on the subject of fANOVA and many scholars have proposed different approaches. First, it is necessary to differentiate between one-way and multi-way fANOVA. The former deals with a functional response and one factor, i.e. a categorical variable with one or more treatments. Although the one-way fANOVA is the most discussed in the literature, multi-way fANOVA is also quite interesting because it considers that many factors and their combined effect can impact the functional response.

Another important distinction, as in the case of the classical ANOVA, is given by "between" and "within" fANOVA. Between-groups fANOVA aims to understand whether there is a statistically significant difference among groups of independent curves. Instead, within-groups fANOVA seeks to determine whether the functional means of groups analysed at distinct instants of time (or under diverse conditions) are significantly different. In the latter case, the observations are not independent because the observed sample is always the same. Even though between-subjects fANOVA is much more present in the literature, within-groups fANOVA has recently received attention, e.g. [22].

Given these fundamental distinctions, this study focuses on between-groups fANOVA. Recently, several statistics have been proposed to evaluate the null hypothesis of equality between functional means. Moreover, many functions have also been implemented in worldwide spread R packages, e.g. *fdANOVA* [23], *fda* [25], and *fda.usc* [26].

In 2004, Shen and Faraway proposed different methods to deal with one-way ANOVA for functional data and suggested a new test based on the b-spline decomposition [27]. Initially, in 1997, in one of the most famous books on FDA, and then in subsequent editions of the same, Ramsay and Silverman suggested an extension of the classical ANOVA F-test for real variables to the context of FDA using the point-wise F-test [18,28]. The latter approach is often adopted in other research [18–21] and it is used in this paper as the primary methodology for our proposal. Cuevas et al. proposed the V statistic based on the sum of the integral differences between the functional means of different pairs of groups. In their work, they find the distribution of the suggested statistic and derive the asymptotic properties that lead to using their one-way fANOVA test (for more details, see [29]). Cuesta-Albertos and Febrero-Bande suggested extending the fANOVA approach to the Multiway fANOVA, in which different conditions may affect the functional outcome [30]. Finally, Zhang and Liang, in 2014, proposed a one-way fANOVA via globalising the point-wise F-test. A very detailed and comprehensive review on fANOVA is conducted in the work of Gorecki and Smaga [31], who also describe the contributions of Flores et al. [32] and Zhang [33], who are of great significance in the vast literature on fANOVA. For further information on the most considerable contributions to the literature on fANOVA and their methodological details, we refer to Gorecki and Smaga [31] and their R package *fdANOVA* [23], which includes the R code for dealing with the most famous approaches to the topic.

In the rest of this article, we focus on the Ramsay and Silverman methodology [18] and employ other methods to assess our results. In particular, we concentrate on the techniques of Cuesta-Albertos and Febrero-Bande [30], Gorecki and Smaga [31], and Zhang and Liang [34].

Ramsay and Silverman's proposal looks for differences among mean curves under V treatments over the whole functional domain like the other methodologies. The null hypothesis is that the groups have the "same" functional means against the alternative that there is some differences among them. We assume that there is a single factor with V different groups ($v = 1, 2, \ldots, V$) and $N = \sum_{i=1}^{V} n_v$ observations, with n_v observations within each group v. Thus, the model for the i-th functional observation ($i = 1, 2, \ldots, \mathrm{N}$) in the v-th group can be expressed by:

$$\mathrm{y}_{iv}(t) = \mu(t) + \gamma_v(t) + u_{iv}(t) \tag{5}$$

where $y_{iv}(t)$ is the functional response of the i-th curve in the v-th group, $\mu(t)$ is the grand mean function, $\gamma_v(t)$ is the functional effect of being in a specific group v, and $u_{iv}(t)$ is the residual function, i.e. the unexplained variation for the i-th observation within the v-th group.

The null hypothesis that the groups have the same functional means, against the alternative hypothesis that there is a difference, can be expressed as follows:

$$H_0: \mu_1(t) = \mu_2(t) = \cdots = \mu_V(t)$$
$$H_1: \mu_{v'}(t) \neq \mu_{v''}(t) \quad \text{for at least one couple } (v', v'') \text{ with } v' \neq v''$$

Suppose $SSA(t)$ represents the variance between groups in the functional context and $SSE(t)$ is the functional variance within groups. Hence, we can express the point-wise functional F-statistics as follows:

$$F(t) = \frac{\frac{SSA(t)}{V-1}}{\frac{SSE(t)}{N-V}} \tag{6}$$

where V-1 and N-V identify the degrees of freedom of two components in the ratio of Equation (6) that is obtained by calculating the Fisher test statistic for each point of the time domain.

As in the classical ANOVA, a large $F(t)$ value means that the variance described by the model is higher than the non-explained

variance. The main distinction between this method and the conventional univariate or multivariate ANOVA is that the value of $F(t)$ is not fixed, but it varies over the entire domain. The classical significance level was intended to be utilised for a single hypothesis rather than a continuum. Hence, we must defend ourselves against falsely claiming around in the interval. To this end, a viable solution is applying the permutation test [18], which is the functional equivalent of the univariate F-test statistic. It allows us to estimate any significant differences between the groups. The idea is to calculate the Fisher test statistic (Equation (6)) as a function created from the series of point estimates for each domain piece.

Nevertheless, a unique test statistic is required to formally test the null of no-relationship between the functional variables. Adopting the maximum of the observed F-statistics function, the distribution under the null hypothesis can be obtained by estimating the test-statistic several times, making random permutations of the curves. In particular, the idea is to assess the observed F-statistic function by getting the "*observed functional F*", whose maximum is needed to estimate the p-value of the test. We randomly re-label the curves with different curve numbers in the second step without modifying the grouping structure. For the set of re-labelled curves, we calculate the F-statistics for each point of the domain and the maximum of these functions. This re-labelling procedure is replicated many times, and for each, we estimate the point-wise F-statistic function and its maximum. In the third step, we find the point-wise 0.05 critical value of the null distribution at each domain point and calculate the 95-th percentile of the F-statistic values corresponding to that point. The last action provides us with the maximum 0.05 critical value of the null distribution, calculating the 95-th percentile of the distribution obtained by the permutations of the second step. Taking the 95-th percentile, we find the critical threshold value without needing the F distribution's classical statistical tables.

2.4. Augmented fANOVA with derivatives information

This paper proposes extending the fANOVA approach by using additional information on curves' characteristics. Many authors who contributed to the development of FDA (e.g. [18]) highlighted that fundamental information on the curves is often contained in the derivatives rather than in the original functions. Although this is known to all scholars, when it comes to differences between groups, only the original functions are considered without paying attention to the additional information, such as the speed and acceleration of the phenomenon. In this article, we propose an augmented fANOVA, i.e. not limited to testing differences between the means of the groups composed of the original functions but extended to comparing the means of derivatives. This knowledge is exciting because the derivatives can capture supplementary knowledge that the original functions miss and imperceptible differences that only occur in some parts of the time domain.

Let the functional derivative of order r for the i-th curve be given by:

$$y_i^{(r)}(t) = \sum_{j=1}^{K} c_{ijr}\theta_j^{(r)}(t) \quad j = 1, \ldots, K \tag{7}$$

where c_{ijr} is the coefficient of the i-th curve and j-th b-spline and $\theta_j^{(r)}(t)$ is the r-th derivative of the j-th basis function. The subscript r indicates the estimated value of the coefficient c_{ijr} for a specific r-th derivative; indeed, as pointed out in [18], this is an estimated value. As stressed by Ramsay and Silverman [18], the choice of the basis system is particularly important for the derivatives estimates. Indeed, we must assume that the basis for representing the object can support the order of derivative to be computed. For b-spline bases, this means that the order of the spline must be at least one larger than the order of the derivative to be calculated. In this context, we will focus on a four-order b-spline basis. The highest degree of derivative we analyse is the second; consequently, there are no problems estimating the basis functions' derivatives. In this work, the estimate is made using the *deriv.fd* function in the *fda* R package [25].

Consequently, the functional derivative could be in a different space spanned by a diverse basis system. A simple multivariate approach based on the joint use of the coefficients of the b-splines of various derivatives, used all together, must be done with caution. Effectively, the coefficients of the linear combinations of diverse derivatives have different orders of magnitude and variability. However, focusing on a point-to-point method, as in the functional F-statistic introduced below, the change of the basis system does not involve any harmful consequences.

Extending Equation (6) to derivatives involves considering the null hypothesis that the functional groups have the same functional r-derivative means against the alternative hypothesis that there are some differences:

$$H_0: \mu_1^{(r)}(t) = \mu_2^{(r)}(t) = \cdots = \mu_V^{(r)}(t)$$
$$H_1: \mu_{v'}^{(r)}(t) \neq \mu_{v''}^{(r)}(t) \quad \text{for at least one couple } (v', v'') \text{ with } v' \neq v''$$

Consequently, we can express the point-wise functional F-statistic related to the r-th derivative as follows:

$$F^{(r)}(t) = \frac{\frac{SSA(t)^{(r)}}{V-1}}{\frac{SSE(t)^{(r)}}{N-V}} \tag{8}$$

where $SSA(t)^{(r)}$ and $SSE(t)^{(r)}$ are the sum of squares among groups and the sum of squares of residuals based on the groups composed of the curves expressed via the r-th derivatives.

It follows that by limiting the attention on $r = 1$ and $r = 2$, we obtain as specific cases of Equation (8), the point-wise F-statistic based on the velocity of the curves, i.e. $F^{(1)}(t)$, and the point-wise F-statistic based on the acceleration of the curves, i.e. $F^{(2)}(t)$. The permutation tests and the thresholds obtained as in the classical fANOVA can be used to test for significant differences among groups considering the additional information provided by $F^{(1)}(t)$ and $F^{(2)}(t)$.

The final evaluation of the equality of groups means can be kept separate considering the different aspects and reasonings of existing differences over time. Alternatively, a final decision based on other reasoning can be considered. More details about this last aspect are presented in the discussion Section.

3. Application and results

We focused on the Google queries in 2018 for the term "*Cytotec*" in the application so as not to use a period that can overlap with the pandemic that could have altered the amount of research. In particular, we take the period from the beginning of January to the end of December and have one temporal observation for each week.

Fig. 1 shows the original raw data of Google queries in 2018 for the term "*Cytotec*". We can observe great confusion, and it isn't easy to understand what is happening. The point data is joined together via a graphical interpolation. Fig. 2 illustrates the smoothed functions of Google queries using Equation (4). For the analysis, sixty-three countries with three different legislations concerning abortion were used. All those countries in which the searches were in sufficient volume to be considered by the Google search engine

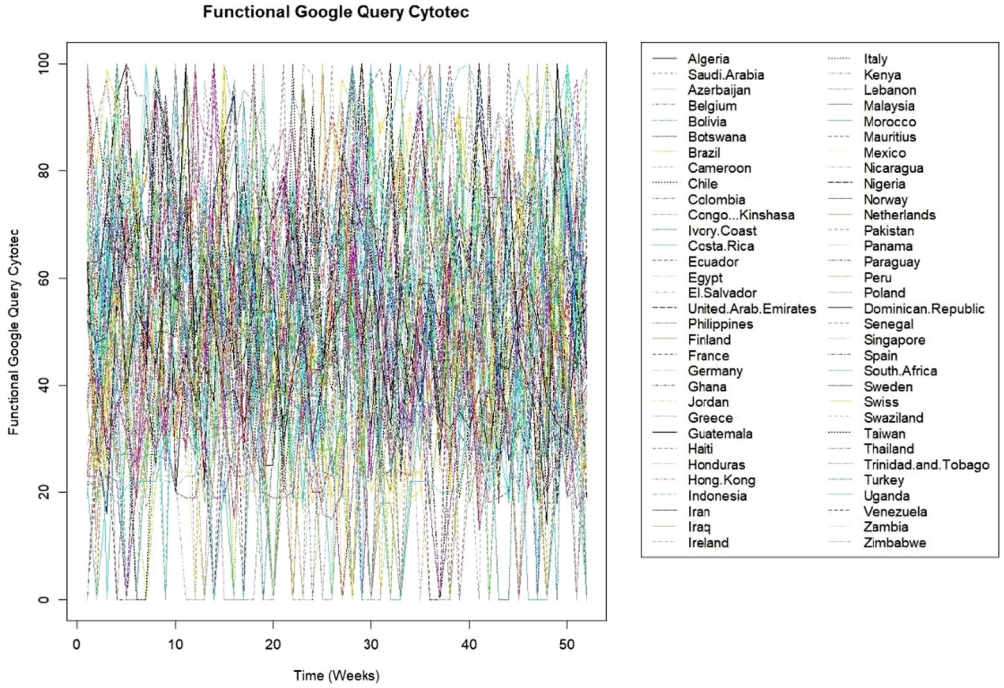

Fig. 1. Original raw data of Google queries in 2018 for the term "Cytotec". (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

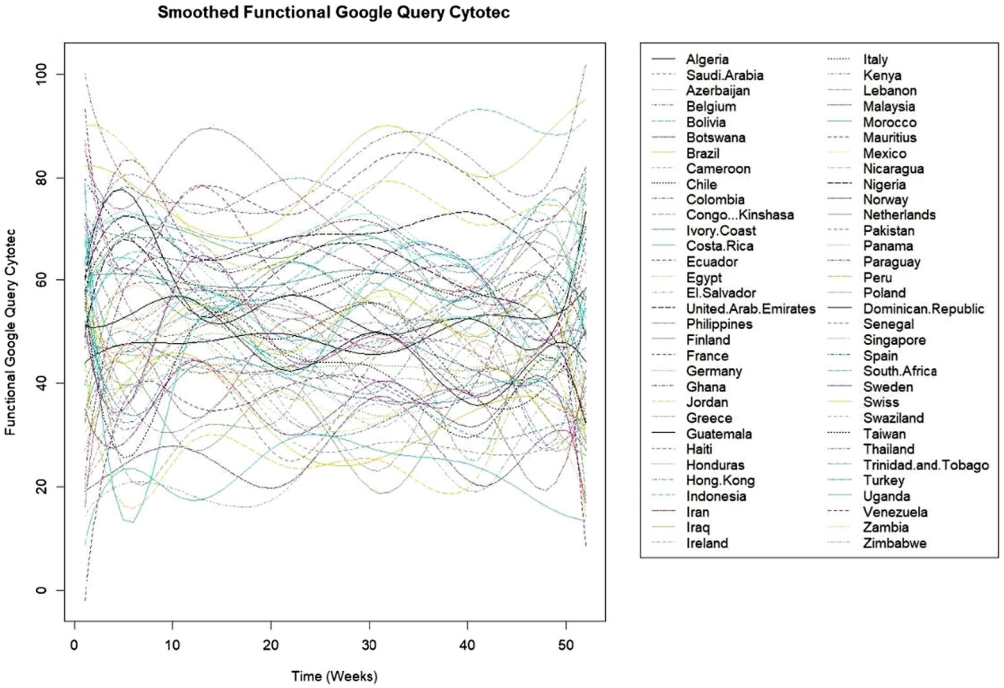

Fig. 2. Smoothed functions of Google queries in 2018 for the term "Cytotec".

were selected for the analysis. As a result, countries with too low search volumes were excluded. Figs. 3 and 4 present the first and second derivatives of Google queries, respectively. The two images highlight very different characteristics compared to Fig. 2 because they are susceptible to the speed and acceleration of the original curves at specific points of the domain.

Fig. 5(a), Fig. 5(b), and Fig. 5(c) illustrate the functional means of the group for the original curves, first derivatives, and second derivatives, respectively.

The solid black curve represents the functional average of drug research in countries where abortion is not allowed in each image. The dotted red curve denotes the functional average of countries where legislation imposes limits on abortion. Finally, the green dot curve characterises the functional average of searches in countries where abortion is allowed and legal.

Fig. 6(a), Fig. 6(b), and Fig. 6(c) show the permutation functional F-tests based on the original curves, first derivatives, and second derivatives, respectively.

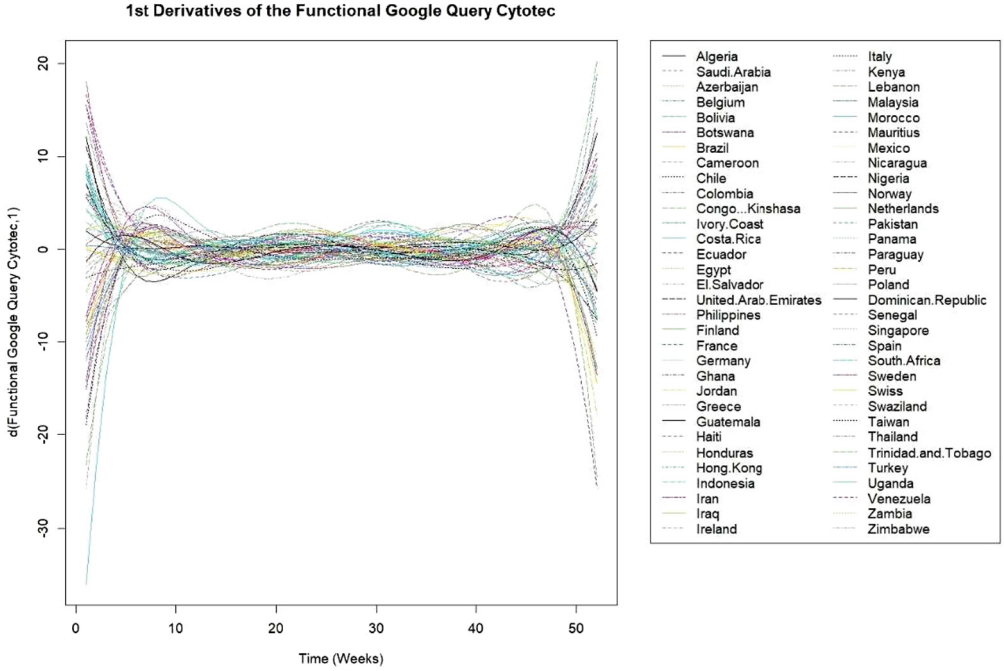

Fig. 3. Smoothed first derivatives of Google queries in 2018 for the term "Cytotec".

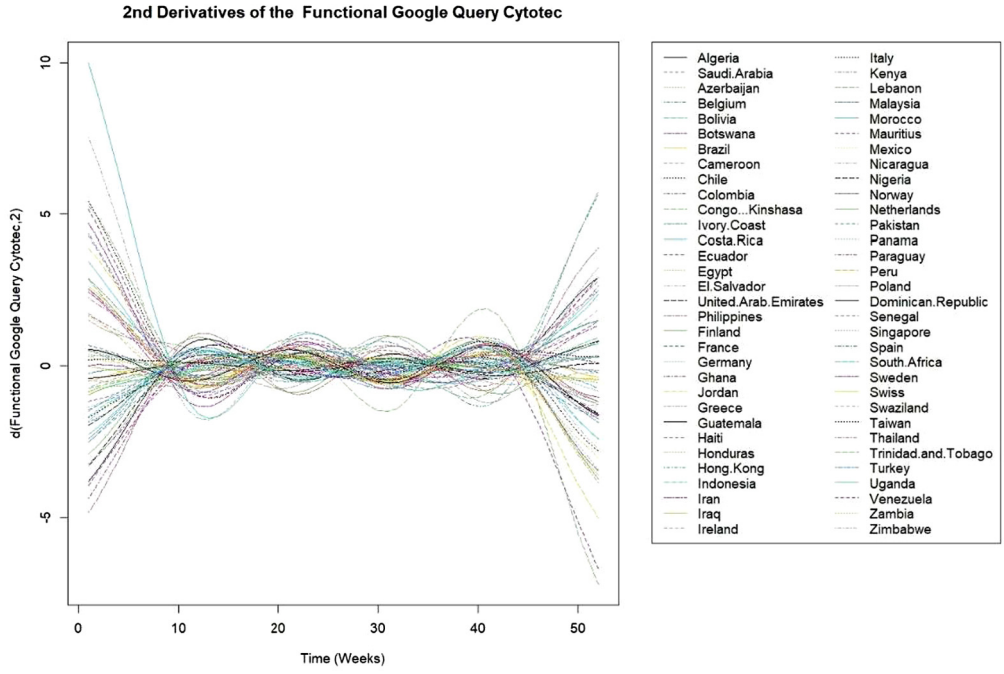

Fig. 4. Smoothed second derivatives of Google queries in 2018 for the term "Cytotec".

In each of the last pictures, a separate one-way fANOVA are performed to test the null hypothesis that there are no significant differences between the mean group functions. The solid line represents the observed F-statistic. The dashed curve indicates the 0.05 point-wise critical value computed with the permutation test. The dotted line is the 0.05 maximum critical value. This test is based on the null distribution constructed using 1,000 random permutations of the curve labels. Fig. 6(a) highlights that the observed F-statistic crossed the 0.05 point-wise critical value one time after a few weeks but did not intersect and always lies below the 0.05 maximum critical values. This denotes a statistically significant difference between the groups only during a week. Ramsay [18] suggested that to protect against false declarations, we must observe the 0.05 of the maximum critical level to decide on the whole domain. Thus, limiting the attention to the entire domain, we can conclude that there are no significant differences between the groups in terms of their means based on the original functions. Focusing on the test based on the first derivatives, we can observe a different behaviour because there is an intersection in the central part of the time domain.

For this reason, the conclusion similar to the previous one is that the test supports the null hypothesis of equality between functional means of the first derivatives on the whole domain. However, a significant difference in a small part can be due to

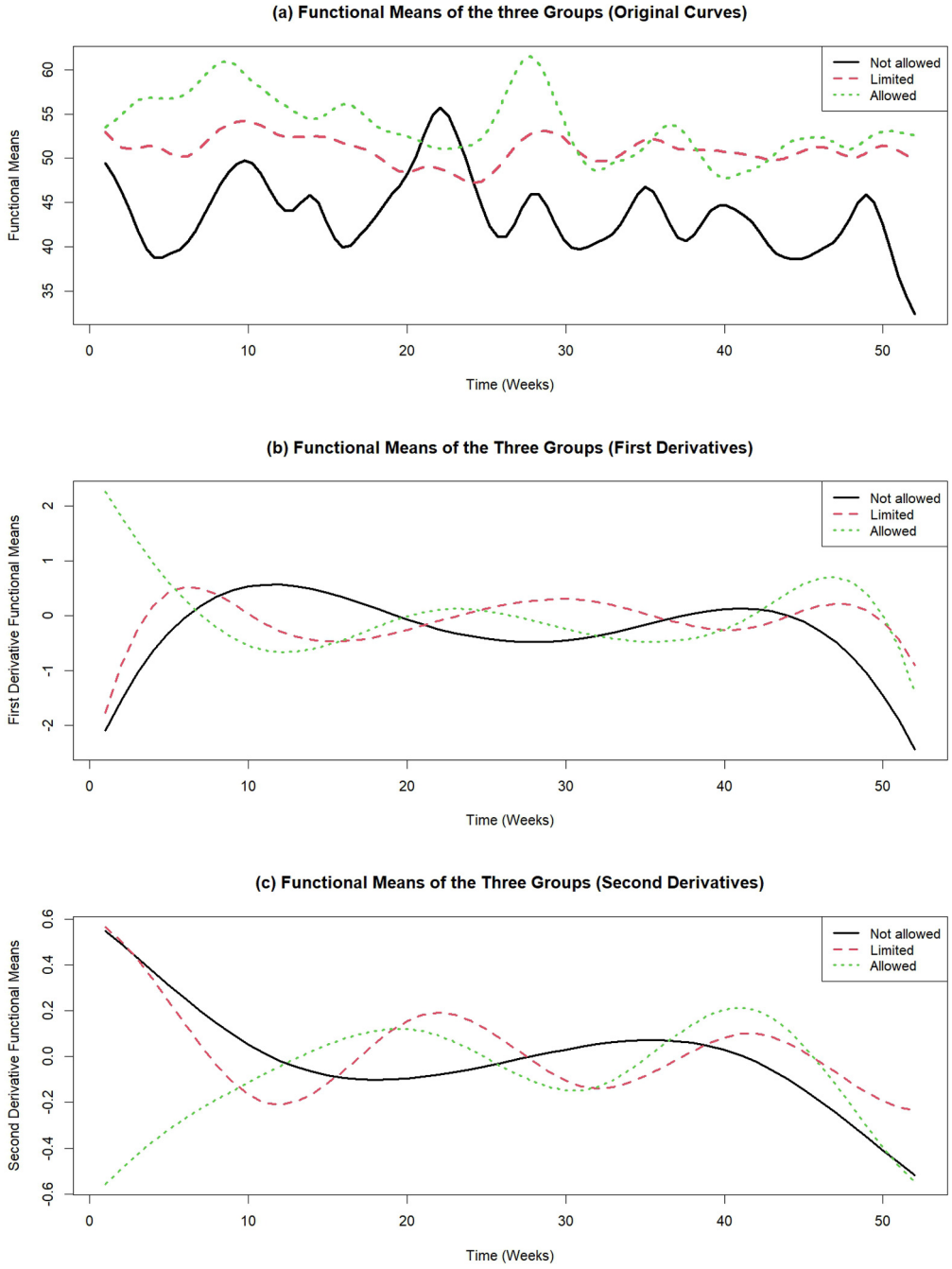

Fig. 5. Functional means of the group for the original curves, first derivatives, and second derivatives.

chance. Regarding the test based on second derivatives, the observed F-statistic does not cross the 0.05 point-wise critical value and the 0.05 maximum critical values; therefore, we can conclude that there are surely no significant differences between the groups in terms of their mean functions. In other words, the legal possibility to abort has not significantly impacted the average trends of the three groups.

The proposed approach is based on the functional F-statistic presented by Ramsay and Silverman [18]. However, A-fANOVA can be easily extended to other procedures existing in the literature. Thus, to compare our results with competing strategies, we integrate the idea of A-fANOVA into the tests proposed by Cuesta-Albertos and Febrero-Bande [30], Gorecki and Smaga [31], and Zhang and Liang [34] available in the R packages *fda.usc* and *fdANOVA*.

Figs. 7, 8, and 9 show the results of the one-way fANOVA suggested by Albertos and Febrero-Bande [30]. The test is performed using the function *fanova.onefactor* [29] that is available in the R package *fda.usc* [26] to the original curves and the first and second derivatives, respectively. All results show a p-value greater than 0.05 and are consistent with the approach in [18] because they suggest that the functional means are not statistically significantly different. Since the result agrees on all dimensions, we can confirm that the groups are similar for each dimension.

Table 1 presents the results of the approaches of Gorecki and Smaga [31] (permutation test based on basis function representation - FP) and Zhang and Liang [34] (globalising the point-wise F-test - GPF) available in the R package *fdANOVA*.

Both tests are based on one-thousand permutations, four order b-splines, and the Bayesian information criterion (BIC) criterion.

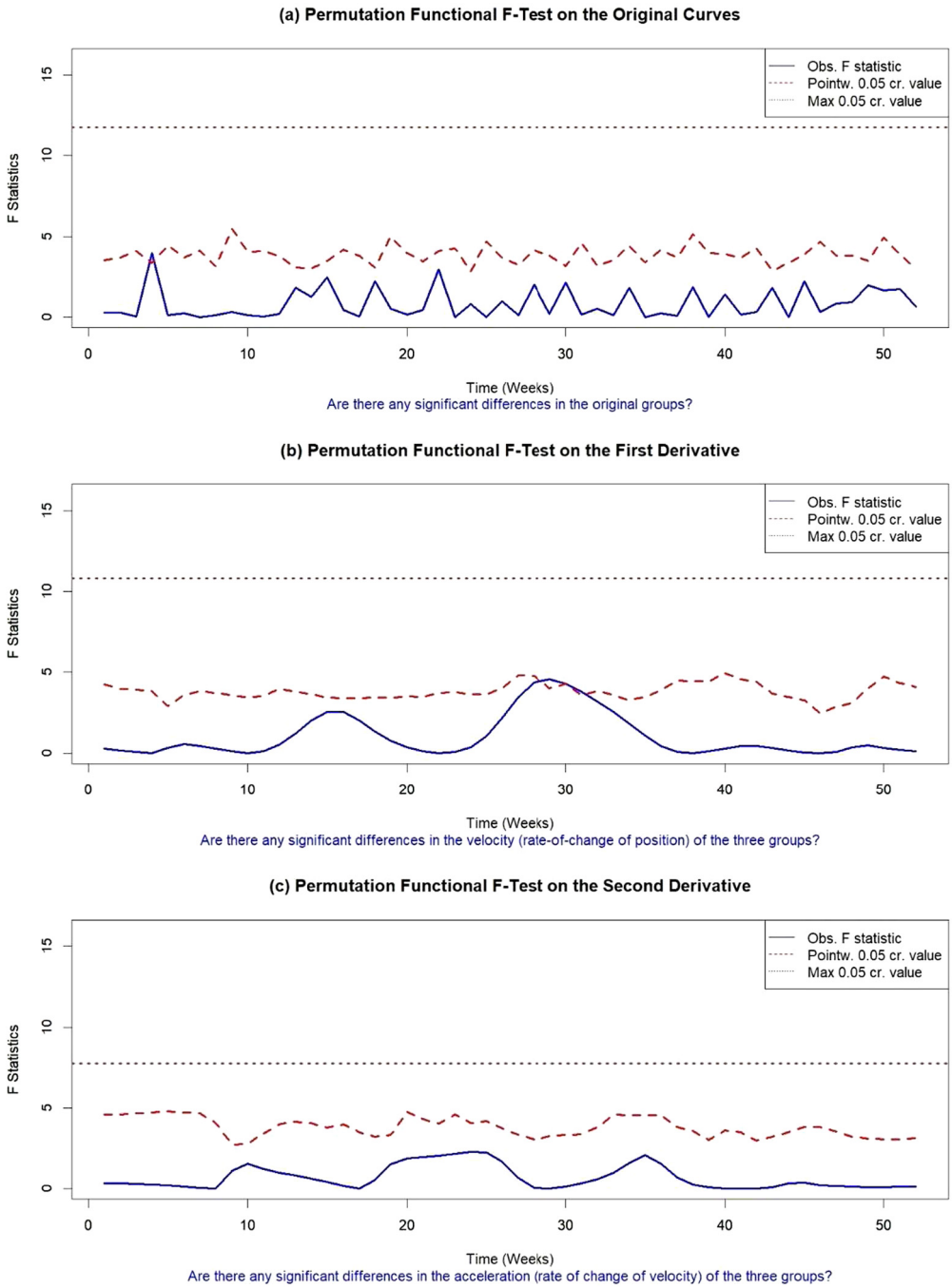

Fig. 6. Functional ANOVA according to different characteristics of the curves.

Table 1
fANOVA implemented using the function fanova.tests in the R package fdANOVA [29]: permutation test based on basis function representation (FP) and global pointwise F-test (GPF).

	Original curves	First derivatives	Second derivatives
FP	Test statistic = 0.872 p-value = 0.419	Test statistic = 0.568 p-value = 0.690	Test statistic = 0.464 p-value = 0.770
GPF	Test statistic = 1.033 p-value = 0.448	Test statistic = 0.917 p-value = 0.543	Test statistic = 0.629 p-value = 0.730

Again, consistent with the previous approaches, all tests suggest insufficient evidence against the null hypothesis. Detailed illustrations of the methodologies used for comparison purposes are available in [26,29,31,34].

4. Discussion

Because residents of more religiously conservative communities are more likely to be reticent toward abortion, abortion practices are less likely to be used in some countries. Moreover, because of fear, religion, privacy, and the judgment of others, people may also be less willing to mention where they can have an abortion. For these reasons, women seeking an abortion often conduct an extensive web search, even using tools such as the dark Web running into dangerous encounters.

In response to the current Covid 19 pandemic, many governments such as the U.S., Italy, the UK, Germany, Hungary and others have endorsed telemedicine for abortion care, thus overseeing and controlling “do-it-yourself abortion” [23]. Despite the progress made in the most open countries, there are still barriers to abor-

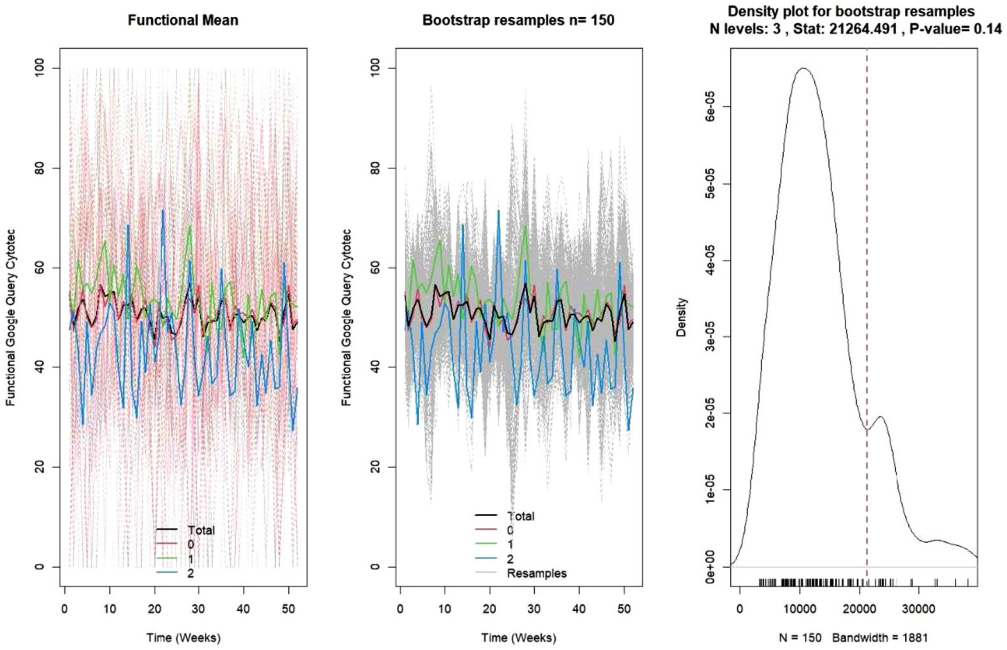

Fig. 7. fANOVA applied to the original curves using the function *fanova.onefactor* [29] available in the R package *fda.usc* [26].

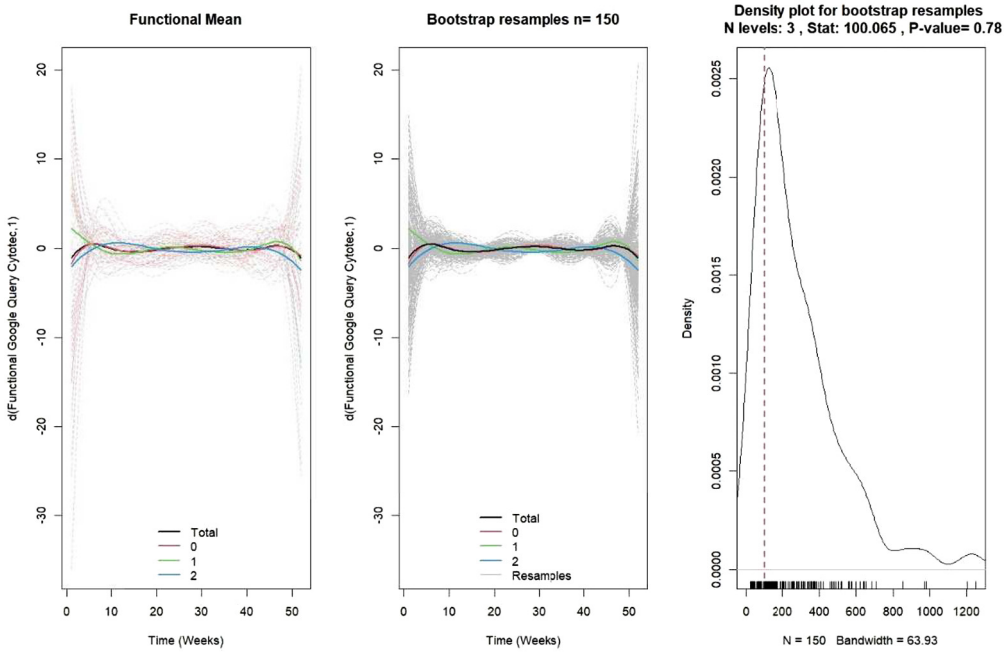

Fig. 8. fANOVA applied to the first derivatives using the function *fanova.onefactor* [29] available in the R package *fda.usc* [26].

tion for the most vulnerable women, such as adolescents, those who have been raped or those in complex social and economic conditions. Hence, it becomes apparent that a large piece of the population will turn to the Internet to procure an unsupported unsupervised and sometimes illegal abortion. For these reasons, it is helpful to understand the frequency and characteristics of searches regarding abortion through web searches. Local authorities and politicians must then understand the magnitude of the problem and prepare new legislative tools that allow vulnerable people to protect their condition by reflecting on the religious barriers that endanger the lives of millions of women worldwide.

Analysing Google trends means dealing with a large amount of data over time. Thus, dealing with this kind of data is always a challenge due to the curse of dimensionality.

Section 3 illustrated an application to remark a potential implementation of the proposed approach. Even if there are few significant differences (in some parts of the time domain in the permutation F-test based on the original curves and that on the first derivatives), the comprehensive tests led to accepting the null hypotheses of equality of means for all the three fANOVAs. In our application, we conclude that abortion laws do not statistically affect the number of research people do on the WWW to obtain information about abortion medicines. This means that, regardless of the abortion laws in countries, people are interested in abortion drugs, even if they are outlawed in their country. This consideration should make policy-makers and those responsible for monitoring compliance with the law think carefully.

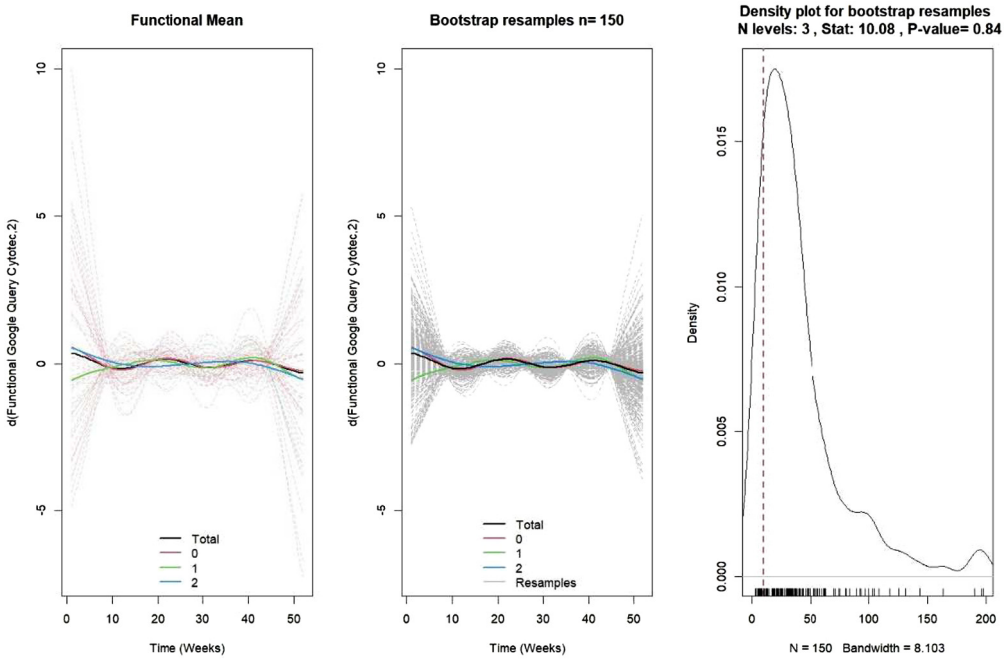

Fig. 9. fANOVA applied to the second derivatives using the function *fanova.onefactor* [29] available in the R package *fda.usc* [26].

Concerning the methodological proposal, the basic idea of the Augmented fANOVA is to consider all the characteristics of the functions to test if there are differences between the functional means of the groups based on distinct aspects. For this reason, we focused on three different one-way fANOVA tests: the first based on the original functions, the second on the speed of google trends, and the third on the search acceleration of the "Cytotec" query. Although the tests show the same results in favour of the null, in many applications, the results conflict when considering different functional derivatives.

Possible developments could contemplate a comprehensive test that considers all three tests with an appropriate adjustment of the p-value. Nonetheless, a global test would suffer from two fundamental limitations. The first drawback is that the dimensions are highly dependent. This latter aspect can be easily overcome with appropriate adjustments, which are very similar to those we can find in the fANOVA within subjects (repeated measures fANOVA), i.e. with statistical units observed under different conditions and thus not independent [22].

The second issue is of theoretical nature. Indeed, our idea is to keep the tests separate and not produce a final test that merge together things of a distinct nature. Our perspective is to maintain the characteristics disconnected and perform three different tests, one for each dimension.

If a final decision must be taken, we suggest two alternatives. The first is the majority vote, which consists in considering an odd number of tests, as in our application, and opting for the equality/difference between groups when the majority of the tests agree in one direction. The second possibility, which in our opinion is more reasonable, is that, if we really want to produce a final decision on all dimensions simultaneously, we can pick the unanimity rule. In other words, the groups can be defined as "equal" if they are equal in all respects and vice versa. This final solution looks pretty reasonable; indeed, it would make little sense to accept a null hypothesis of a comprehensive test where at least one dimension is against the null. Therefore, in case of conflicting results among different dimensions, no final decision would be satisfactory from a theoretical perspective.

With reference to the type of method to use for the one-way fANOVA test, as highlighted in the paper, there are many proposals in the literature. In general, all of them can basically be adapted to the case of A-fANOVA with appropriate adjustments. In this study, we used the Ramsey and Silverman approach as a benchmark, and then we considered other methods known in the literature to make a comparison of the results, but the idea can also be extended to other multidimensional methods and to a greater number of derivatives to consider.

5. Conclusions

In this study, we presented an original application to understand how the legal possibility to abort differently influences the trend of Cytotec searches in different countries. Regardless of the application submitted in this context, FDA in medical and public health studies in recent years is becoming widespread (see, e.g. [35–37]). In addition to this possible application, FDA offers numerous advantages, such as the possibility of using additional tools such as derivatives, curvature, and dimensionality reduction techniques when there are datasets with few statistical units and many temporal observations, as in this case. Moreover, recently many scholars have underlined the advantages of this approach concerning the analysis of time series because, in FDA, we do not need any assumptions and can easily be used even with scattered data.

All these possible future developments and their applications in statistics can be very interesting for policy-makers and institutions that intend to monitor phenomena on the WWW, but also the pharmaceutical companies that are interested, for marketing purposes and sales, to understand the evolution of people's interests in abortive drugs, and even other medicines. In this context, the FDA applied to Google Searches proves to be a valuable tool for finding and interpreting a population phenomenon. In this context, we used the functional analysis of variance to understand if there are significant differences between groups of countries that have different legislation regarding abortion. In any case, applying functional analysis to Google Trends can also be extended to other types of analysis, such as functional time series or regression analysis with different purposes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] R. Kumar, Misoprostol and the politics of abortion in Sri Lanka, Reprod. Health Matters 20 (2012) 166–174, https://doi.org/10.1016/S0968-8080(12)40652-8.

[2] C. Millard, P. Brhlikova, A. Pollock, Social networks and health policy: the case of misoprostol and the WHO model essential medicine list, Soc. Sci. Med. 132 (2015) 190–196, https://doi.org/10.1016/j.socscimed.2015.03.011.

[3] I.H. Shah, M.B. Weinberger, Expanding access to medical abortion: perspectives of women and providers in developing countries, Int. J. Gynecol. Obstet. 118 (2012) S1–S3.

[4] World Health Organization, Safe Abortion: Technical and Policy Guidance for Health Systems, 2nd edition, WHO, 2013, p. 123.

[5] M.M. Fernandez, F. Coeytaux, R.G.P. de León, D.L. Harrison, Assessing the global availability of misoprostol, Int. J. Gynaecol. Obstet. 105 (2009) 180–186, https://doi.org/10.1016/j.ijgo.2008.12.016.

[6] R.M. Barbosa, M. Arilha, The Brazilian experience with Cytotec, Stud. Fam. Plann. 24 (1993) 236–240.

[7] H.L.L. Coêlho, A.C. Teixeira, A.P. Santos, E.B. Forte, S.M. Morais, C. LaVecchia, G. Tognoni, A. Herxheimer, Misoprostol and illegal abortion in Fortaleza, Brazil, Lancet 341 (1993) 1261–1263.

[8] D. Diniz, M. Medeiros, Itinerários e métodos do aborto ilegal em cinco capitais brasileiras, Ciênc. Saúde Colet. 17 (2012) 1671–1681, https://doi.org/10.1590/S1413-81232012000700002.

[9] N. Zamberlin, M. Romero, S. Ramos, Latin American women's experiences with medical abortion in settings where abortion is legally restricted, Reprod. Health 9 (2012) 34, https://doi.org/10.1186/1742-4755-9-34.

[10] J.D. Gipson, A.E. Hirz, J.L. Avila, Perceptions and practices of illegal abortion among urban young adults in the Philippines: a qualitative study, Stud. Fam. Plann. 42 (2011) 261–272, https://doi.org/10.1111/j.1728-4465.2011.00289.x.

[11] E.C. Atukunda, M.J. Siedner, C. Obua, G.R. Mugyenyi, M. Twagirumukiza, A.G. Agaba, Sublingual misoprostol versus intramuscular oxytocin for prevention of postpartum hemorrhage in Uganda: a double-blind randomized non-inferiority trial, PLoS Med. 11 (2014) e1001752, https://doi.org/10.1371/journal.pmed.1001752.

[12] R.F. Hess, Women's stories of abortion in southern Gabon, Africa, J. Transcult. Nurs. 18 (2007) 41–48, https://doi.org/10.1177/1043659606294191.

[13] F. Okonofua, O. Shittu, T. Shochet, A. Diop, B. Winikoff, Acceptability and feasibility of medical abortion with mifepristone and misoprostol in Nigeria, Int. J. Gynaecol. Obstet. 125 (2014) 49–52, https://doi.org/10.1016/j.ijgo.2013.10.009.

[14] S.D. Zordo, The biomedicalisation of illegal abortion: the double life of misoprostol in Brazil, Hist. Cienc. Saude-Manguinhos 23 (2016) 19–36, https://doi.org/10.1590/S0104-59702016000100003.

[15] V. Fiol, L. Rieppi, R. Aguirre, M. Nozar, M. Gorgoroso, F. Coppola, L. Briozzo, The role of medical abortion in the implementation of the law on voluntary termination of pregnancy in Uruguay, Int. J. Gynaecol. Obstet. 134 (2016) S12–S15, https://doi.org/10.1016/j.ijgo.2016.06.006.

[16] S.V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R.P. Dreyer, S.I. Chen, K. Murugiah, The use of Google trends in health care research: a systematic review, PLoS ONE 9 (2014) e109583, https://doi.org/10.1371/journal.pone.0109583.

[17] A. Seifter, A. Schwarzwalder, K. Geis, J. Aucott, The utility of "Google Trends" for epidemiological research: Lyme disease as an example, Geosp. Health 4 (2010) 135–137.

[18] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer-Verlag, New York, 1997.

[19] T.D. Battista, F. Fortuna, F. Maturo, Parametric functional analysis of variance for fish biodiversity assessment, J. Environ. Inform. 28 (2017) 101–109, https://doi.org/10.3808/jei.201600348.

[20] T. Di Battista, F. Fortuna, F. Maturo, Parametric functional analysis of variance for fish biodiversity, in: Proceedings of the International Conference on Marine and Freshwater Environments (IMFE 2014)-Our Water, Our Future, 2014.

[21] F. Maturo, S. Migliori, F. Paolone, Do institutional or foreign shareholders influence national board diversity? Assessing board diversity through functional data analysis, in: Mathematical-Statistical Models and Qualitative Theories for Economic and Social Sciences, Springer, 2017, pp. 199–217.

[22] C. Acal, A.M. Aguilera, Basis expansion approaches for functional analysis of variance with repeated measures, Adv. Data Anal. Classif. (2022), https://doi.org/10.1007/s11634-022-00500-y.

[23] T. Gorecki, L. Smaga, fdANOVA: analysis of variance for univariate and multivariate functional data, Comput. Stat. 34 (2019) 571–597, https://doi.org/10.1007/s00180-018-0842-7.

[24] T. Górecki, Ł. Smaga, fdANOVA: analysis of variance for univariate and multivariate functional data. R package version 0.1.2, https://CRAN.R-project.org/package=fdANOVA, 2018.

[25] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, fda: functional data analysis, R package version 2 (2014) 142.

[26] M. Febrero-Bande, M.O. de la Fuente, Statistical computing in functional data analysis: the R package fda.usc, J. Stat. Softw. 51 (2012) 1–28.

[27] Q. Shen, J. Faraway, An F test for linear models with functional responses, Stat. Sin. 14 (2004) 1239–1257.

[28] F. Maturo, F. Fortuna, T. Di Battista, Testing equality of functions across multiple experimental conditions for different ability levels in the IRT context: the case of the IPRASE TLT 2016 Survey, Soc. Indic. Res. 146 (1) (2019) 19–39, https://doi.org/10.1007/s11205-018-1893-4.

[29] A. Cuevas, M. Febrero, R. Fraiman, An ANOVA test for functional data, Comput. Stat. Data Anal. 47 (2004) 111–122, https://doi.org/10.1016/j.csda.2003.10.021.

[30] J.A. Cuesta-Albertos, M. Febrero-Bande, A simple multiway ANOVA for functional data, Test 19 (2010) 537–557, https://doi.org/10.1007/s11749-010-0185-3.

[31] T. Górecki, Ł. Smaga, A comparison of tests for the one-way ANOVA problem for functional data, Comput. Stat. 30 (2015) 987–1010, https://doi.org/10.1007/s00180-015-0555-0.

[32] R. Flores, R. Lillo, J. Romo, Homogeneity test for functional data, J. Appl. Stat. 45 (5) (2018) 868–883, https://doi.org/10.1080/02664763.2017.1319470.

[33] J. Zhang, Analysis of Variance for Functional Data, Monographs on Statistics and Applied Probability, 2014.

[34] J.-T. Zhang, X. Liang, One-way ANOVA for functional data via globalizing the pointwise F-test, Scand. J. Stat. 41 (2014) 51–71, https://doi.org/10.1111/sjos.12025.

[35] F. Maturo, R. Verde, Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers, Comput. Stat. (2022) 1–32, https://doi.org/10.1002/sim.9353.

[36] F. Maturo, R. Verde, Pooling random forest and functional data analysis for biomedical signals supervised classification: theory and application to electrocardiogram data, Stat. Med. 41 (2022) 2247–2275, https://doi.org/10.1007/s00180-022-01259-8.

[37] F. Maturo, R. Verde, Supervised classification of curves via a combined use of functional data analysis and tree-based methods, Comput. Stat. (2022) 1–41, https://doi.org/10.1007/s00180-022-01236-1.