Assessing Stationarity in Web Analytics: A study of Bounce Rates

Marios Poulos*

Faculty of Information Science and Informatics, Department of Archives and Library Science, Ionian University, Ioannou Theotoki 72, 49100, Corfu, Greece.

Nikolaos Korfiatis

Norwich Business School, University of East Anglia. Elizabeth Fry Building, NR47J, Norwich, United Kingdom

Sozon Papavlassopoulos

Faculty of Information Science and Informatics, Department of Archives and Library Science, Ionian University, Ioannou Theotoki 72, 49100, Corfu, Greece.

This paper should be cited as:

Poulos, M., Korfiatis, N., & Papavlassopoulos, S. (2020). Assessing Stationarity in Web Analytics: A study of Bounce Rates. *Expert Systems (Forthcoming)*

* Corresponding Author, E-mail: mpoulos@ionio.gr; Ioannou Theotoki 72, 49100, Corfu, Greece

Abstract

Evidence-based methods for evaluating marketing interventions such as A/B testing have become standard practice. However, the pitfalls associated with the misuse of this decision-making instrument are not well understood by managers and analytics professionals. In this study, we assess the impact of stationarity on the validity of samples from conditioned time series, which are abundant in web metrics. Such a prominent metric is the *bounce rate*, which is prevalent in assessing engagement with web content as well as the performance of marketing touchpoints. In this study, we show how to control for stationarity using an algorithmic transformation to calculate the optimum sampling period. This distance is based on a novel stationary ergodic process that considers that a stationary series presents reversible symmetric features and is calculated using a dynamic time warping (DTW) algorithm in a self-correlation procedure. This study contributes to the expert and intelligent systems literature by demonstrating a robust method for subsampling time series data, which are critical in decision making.

Keywords: Time Series, Numerical Analysis, Stationary Progress, Bounce Rate, DTW

1. Introduction

The proliferation of analytical methods and tools has led to a critical paradigm shift in managerial decision making, highlighting the importance of evidence-based evaluations of the impact of interventions across a spectrum of business practices (e.g., marketing). As in other areas, such as medicine, evidence-based methods in management practice (Marr, 2010; Pfeffer & Sutton, 2006) seek to evaluate not only whether the effect of an intervention is observable but also the reliability and validity of the results presented by the evaluation criterion used. The expert and intelligent systems' literature has necessitated the need for correct data input across a variety of methods and application domains. A very prominent case (as regards to measurable economic significance) is the evaluation of the impact of interventions in presentation/interface elements in various marketing functions, such as e-commerce and display advertising. The former has given rise to so-called "*customer-driven*" development (Edvardsson et al., 2012), in which real customers (or users) evaluate features of a particular medium under realistic marketing mix conditions.

Such a problem considers evaluating the performance of media use and consumption in terms of easy-to-understand metrics to guide budget allocation (Danaher & Rust, 1996). Considering a typical application scenario of an online retailer or an advertising agency, performance metrics capture consumer engagement with the medium and its effectiveness in attracting consumers' attention. A very prominent metric, which is the focus of this study, is the bounce rate, which is defined as the ratio of single-page user sessions to the total sessions within a given time duration (Sculley et al., 2009). A high bounce rate can lead to a poor retailer/advertiser return on investment (ROI) and suggests that users may have a poor experience once they land on a particular page through a referral link (e.g., by clicking on an ad or by finding the page through a web search). The former is commonly referred to as a marketing touchpoint.

A typical way to address such a deficiency is to intervene in the interface elements to find the combination that leads to the best performance metric (e.g., the lowest bounce or click-through rate) and measure the effect of this intervention. This approach is known as A/B testing and considers splitting the visitor traffic into two streams, which are assigned to the baseline condition (*B*) or its alteration (*A*). When considering multiple alternations wherein the comparison of the difference considers more than two groups, M/N or multivariate testing is performed (Kohavi et al., 2009). The evaluation of the effect of these interventions is performed through a typical test of the mean differences using either two-sample parametric tests or ANOVA when considering various alternative interventions. This capability is integrated into web analytics tools, which are prevalently used to guide decision making by analytics professionals. With the ever-increasing dimensionality of the test features and attributes of testing, expert input has become limited and biased (Sauter, 2014).

A typical example of such a bias is the decision concerning the duration of A/B tests and whether enough statistical power has been accrued to declare a *winner* or best-performing configuration. Deciding the length of a test is critical since, in the case of the worst performance in the post-hoc period of such a test, opportunity costs arise from lost conversions. Seasonal and cyclical variations of demand have been demonstrated to affect several aspects of economic activity, with online shopping being no exception.

In this study, we aim to address a typical question that is abundant in this type of test, which is "*how long we should sample a session in order to extract results that capture an adequate level of periodicity for an effect to be observed?"* An affirmative and non-biased answer would allow analytics professionals (e.g., those active in search engine optimization) to safely evaluate the economic significance of their intervention, avoiding Type I and Type II errors that typically accompany such undertakings and may result from inadequate sampling.

Such a challenge, while approachable by a set of standard statistical practices, has the characteristic of considering the evaluation of a metric that is of a longitudinal rather than a cross-sectional nature. The former assumes that the time series of the evaluation criterion used in a typical A/B testing scenario corresponds to the aggregated metrics of an entire source and is free of any precondition, and the condition of the time series is inherent in its data structure (e.g., the way the metric is calculated). In this case, the time series is also referred to as a *conditioned* time series (Hamilton, 1994). In some embodiments, a source contains a number of conditioned time series, such as metrics, including visits, page views, bounce rates, pages/visits, new visits, average time on site, etc. (Vaughan & Yang, 2013).

Considering that such time-series data are relatively large or high-frequency, approaches related to sampling and periodicity pose a challenge to standard analytics tools (Varian, 2014). From a statistical viewpoint, the problem that we are looking to address here is more specifically discussed in the work of Downing, Fedorov, Lawkins, Morris, and Ostrouchov (2000). Because of size, there is the assumption that the dataset cannot be analysed at once and should be analysed in segments. The strategy adopted in our study considers the segmentation of a large data series into a series of segments of arbitrary length and then an examination of one part of the division at a time to allow unequal segments to reach an optimal segment length. In this way, the variation of the stationarity per period is investigated to ascertain whether there is a stable periodical pattern of this variation, which in turn, can be a guiding heuristic of sample size. Building on previous work (Poulos, 2016), our methodology provides a simple but robust approach to dealing with the segmentation and periodicity estimation of time series data representing conditioned metrics. Considering that such metrics are abundant in web analytics and marketing practices, our work also has practical implications.

Our study responds to several points of interest already outlined in the literature, such as that of Mortenson, Doherty, and Robinson (2015), regarding the integration of operational and computational intelligence methods with the emerging field of data analytics and in particular high-frequency data from digital trails of customer activity. From the perspective that sampling periods can alter the significance of marketing interventions, such as those measured in A/B or M/N testing scenarios, our paper also contributes to the practice of web analytics by incorporating research with real-world data captured through analytics tools that are considered standard in the industry (Google Analytics). To this end, this paper is structured as follows. Section 2 discusses related work and the background of the bounce rate definition and the use of A/B and M/N testing methodology in evaluating the significance of marketing interventions. We provide an analytical formulation and explanation of the algorithmic process in Section 3, where the problem of identifying the optimal sampling period for a conditioned time series is discussed. A benchmark evaluation using data from an online retailer is discussed in Section 4, along with implications for practice in Section (5). The paper concludes with Section 6, discussing limitations and future research directions.

2. Related work

2.1 Bounce rates

Bounce rates represent a significant benchmark for the assessment of the engagement value of interactions—so-called touchpoints—in various areas of content authoring and advertising (Murthy & Mantrala, 2005). In their simplest form, bounce rates can be defined as the ratio of extremely short-lived sessions (generally defined as single-page sessions) established either by direct entry (when the user types the URL into the browser) or by referral entry (by clicking on a hyperlink) and its correspondent landing. Several established industry tools, such as Google Analytics (Clifton, 2012; Plaza, 2011), define bounce rates as sessions in which either immediate back-button clicks have been initiated once the user loads the page or as abandoned clickstreams in which no further action has been taken after the user initiates a session.

Considering the universe S_1^n of *n* sessions initiated on a display space (e.g., website, banner, etc.) with each session corresponding to an event time clickstream of *k* length:

$$S_t = t_{i=1}..., t_{i=k} \qquad (1)$$

the bounce rate (*BR*) is defined as the ratio of sessions in which the depth of the clickstream is singular to the overall number of sessions, such as:

$$BR = \frac{\sum S_{k=1}}{\sum S_{k=1}, S_{k>1}} \qquad (2)$$

Due to its simplicity, the bounce rate has been a standard benchmark for evaluation of the performance of entry points (or referrals) in web analytics. In the case of display or sponsored search advertising, bounce rates can be used to measure the performance of an ad and provide input for decision making in advertising budget allocation (Jeziorski & Moorthy, 2017). For example, if a landing page (the part of the website to which the click-through action leads) has a bounce rate of 80%, this suggests that only 20% of the users that clicked on the ad or sponsored search result were engaged with the action encapsulated in the landing page. Considering that click-through rates are linearly dependent on the cost per click (which, in the case of sponsored search results, varies and is the result of an auction), then an 80% abandonment of the landing page corresponds to a significant loss of the investment provided in the advertising budget.

Nevertheless, while optimizing bounce rates is an obvious approach, several practitioners consider high percentages to be the results of induced demands that can be driven by other factors and not necessarily by user attention (e.g., accidental landings, technical errors, user interruptions, etc.). Industry reports suggest that an average bounce rate of 40% is nominal for particular sectors (e.g., retailing), and as such, more resources should be directed toward the optimization of user trajectories regarding $k \geq 2$ actions in the clickstream (eCommerce Europe, 2016). Furthermore, due to its inherent behavioural nature, the bounce rate depends on the targeting that the ad initiates. Entries initiated through sponsored search advertising (e.g., Google Adwords) tend to have lower bounce rates than do entries initiated through display advertisements (e.g., banners) due to the inherent information targeting that the advertising mechanism uses (Yang & Ghose, 2010).

In the academic literature, researchers have associated increased bounce rates with the engaging nature of the informational content contained in the website or the visual attributes of the content (Lindgaard, Fernandes, Dudek, & Brown, 2006), including audio features (e.g., in the case of disruption). However, our understanding of bounce rate characteristics and whether they can be predicted is somewhat limited (Wells, Valacich, & Hess, 2011), and content optimization techniques, such as A/B testing, have become prevalent as standard tools in the industry.

2.2 A/B testing and sample size

In its simplest form, an A/B test is a randomized controlled experiment technique that involves the experimental evaluation of an overall evaluation criterion OEC (e.g., the performance of an alteration of a web page) against a baseline. From an analytical point of view, it considers a hypothesis test of two samples, with the null hypothesis corresponding to the baseline variant, resembling a between-subjects design from an experimental point of view. It has been adopted by content designers and marketing analysts for the evaluation of different stages of the purchase funnel in e-commerce scenarios (Hoban & Bucklin, 2015). Typically, content designers select a feature that has a level of uncertainty regarding its effect on a performance metric (e.g., bounce rates, click-through rates, etc.). Then, a new page is created (Version B), and a visitor is randomly assigned to either page A (or the baseline), which is the unaltered version of the website, or page B, which represents the altered version of the page. The subject assignment procedure is performed through a randomized mechanism (a so-called splitter), which is typically executed on a server using a cookie assignment to the visitor. This procedure is performed to ensure that for the duration of the experiment, repeat visits are assigned to the same version of the page.

Since the evaluation of the altered version against the baseline is performed with a parametric test, assumptions of normality are followed for all parameters of the problem, including confidence intervals and statistical powers. For several categories of web analytics metrics, for which the underlying distribution is not normal (e.g., Gaussian or Poisson), appropriate non-parametric tests are used. For example, if we consider the evaluation of the effect of an intervention on click-through rates, which has been shown to follow a binomial distribution (REF), Fischer's exact test is used, while non-parametric tests, such as the Mann-Whitney U-test, are dominant when no assumptions about the underlying distribution are made. The standard guiding principle behind the reliability of the test is the statistical significance of the difference between the sample means and the appropriate statistical power that the difference in the selected metric is going to exhibit. Several researchers in the literature have studied the issue from a statistics point of view, and the probability perspective (Brodersen, Gallusser, Koehler, Remy, & Scott, 2015; Varian, 2016) and alternative corrections and criteria have been proposed and adopted from the experimental literature. For example, Gibbs sampling may be appropriate for the selection of sample intervals for A/B testing if no direct data are available about the probability distribution of the chosen OEC.

Regardless of the evaluation approach, questions regarding the optimal sampling size and length are still debatable and subject to the sensitivity of the selected test, and the assumptions regarding the underlying distribution. Our aim in this study is not to delve into the mechanism used to compare the differences between the two samples but to direct our attention toward the issue of sub-sample selection to evaluate the OEC in the context of A/B testing. This issue is directly related to the question of the experimental duration and its time series specific nature. Building on prior work concerning time series stationarity detection (Poulos, 2016), our approach considers the extraction of the stationarity degree to guarantee equal likelihoods of activity captured by the OEC across the testing sample.

2.3 Our contribution

The problem that we tackle with is that the underlying assumption of the random assignment achieved with a split generator in an A/B testing scenario may not be enough to safeguard the validity of the test result, and as such, a more robust approach based on the time series characteristics of the targeted metric is needed.

This problem is of high economic significance for users of an advertising network and, in particular, retailers, since it is costly at two levels. First, the direct advertising cost involves the cost-per-click (CPC) associated with a bounced visit, and second and most importantly, lost opportunity results from missed activity of a potential client. Arguably, the problem of assessing the usability performance of a web space (e.g., an e-commerce site) considers not only the bounce rate but also the overall trackable activity until the point of checkout (and hence other elements of the purchase funnel, which can lead to an abandonment of the clickstream). However, concerning the question of decisions related to budget allocation (e.g., for sponsored-search or display advertising), the returns of these decisions may be harmful if the optimization strategy does not consider an accurate estimation of the time dependence. Inherent sources of error in this case, such as stationarity, have been known to influence the reliability of time-dependent metrics (Sculley et al., 2011), and our intention in this study is to address this issue by introducing an analytical process.

The method is based on an algorithm that detects the sampling stability of a time series (Poulos, 2016). The sampling stability is expressed by the discovery of some dominant periodicity extracted from the change of the stationarity degree within a particular time series segment. Therefore, the algorithmic contribution of the study could be applied beyond the bounce rate issue. The details of this contribution are discussed in section 5.1.

3. Analytic formulation

2.4 Preliminaries

The extraction of the stationarity degree is based on previous work (Poulos, 2016; Sharifdoost, Mahmoodi, & Pasha, 2009), in which it has been defined that a discrete time stationary process $\{M_n\}$ with $i = 1,...: n$, is time reversible for every positive integer n if the following equation is satisfied:

$$(M_1, M_2,...,M_n) = (M_n,M_{n-1},...,M_1) \qquad (3)$$

Then, it is considered that a discrete time series $\{\overrightarrow{M}_n\}$ with $i = 0,...,n$ produces a mirror time series, which can be described as:

$$\{N_n = \overleftarrow{M}_n; i = 0, \ldots, n\} \qquad (4)$$

Thus, taking into account Equation 4, the degree of stationarity is based on in the following formulation:

$$\{\overrightarrow{M}_n = \overleftarrow{M}_n - error\} \qquad (5)$$

If $error = 0$, then the time series $\overrightarrow{M}_n$ consists of a stationary process based on the error estimation of the dissimilarity measure between the discrete time series $\overrightarrow{M}_n$ and the reversible $\overleftarrow{M}_n$. Then, using Euclidean and dynamic time warping (DTW) techniques, the local dissimilarity of the function f is defined between any pair of elements $M_n \wedge N_n$, with the shortcut:

$${}_{i,j=1}^{\;n}\left|d\left(i, j\right) = f\left(M_i, N_i\right) \geq 0 \qquad (6)$$

Then, if the path is the lowest cost path between two series, the corresponding dynamic time warping (DTW) technique (Salvador & Chan, 2007) provides the warping curve $\varphi(k)$, $\forall k = 1,2,\ldots,T$ as:

$$\begin{aligned} &\varphi(k) = \left(\varphi_\chi(k), \varphi_y(k)\right) \;\; with \\ &\varphi_\chi(k) \;\; \wedge \;\; \varphi_y(k) \in \left\{1,2,\ldots,n\right\} \end{aligned} \qquad (7)$$

The warping functions $\phi_x(k) \wedge \varphi_\psi(k)$ remap the time indices of $M \wedge N$ accordingly. Given ϕ and following Cortez, Rio, Rocha, and Sousa (2012), the average accumulated distortion between the warped time series $M \wedge N$ is calculated as follows:

$$d_\varphi(M,N) = \sum_{\kappa=1}^{\mathrm{T}} \frac{d\left(\varphi_x(k), \varphi_x(k)\right) m_\varphi(\kappa)}{M_\varphi} \qquad (8)$$

where $m_\phi(k)$ is a per-step weighting coefficient of the corresponding normalization constant (m_ϕ), which confirms that the accumulated distortions are comparable along different paths.

To ensure reasonable warps, constraints are usually imposed on ϕ. The basic idea underlying DTW is to find the optimal alignment ϕ such that:

$$D\left(M_n, N_n\right) = \min_\varphi \left\{d_\varphi\left(M_n, N_n\right)\right\} \qquad (9)$$

Therefore, one picks the distortion of the time axes of $M \wedge N$, which brings a couple of the time series as near to each other as possible.

2.5 Procedural definition

Graphically, this algorithm is described in Figure 1. We provide a more detailed analytical overview and the algorithmic steps below.

[Insert Figure 1 here]

Step1. Let us consider the matrix *M*, which contains the hourly bounce rate data set with the size (1 × *R*), $R \in N$ ${}_{j=1}^{i}\left|M(j,x) \subset M\right.$, where

$$ {}_{j=1}^{i}\left|M(j,x) = M(x+j : x+j+i), 0 < x < R-i \right. \quad (10)$$

Index *j* corresponds to the number of repetitions of the algorithm in the same window length each time, with a unit step of sliding. Additionally, the indicator *i* is the selected size of the investigated window, which is constant for each experiment, and *x* is the beginning point of the series. Then, the corresponding mirror data set is:

$$ {}_{i,j=1}^{i}\left|N(j,x) = N(x+j+i : x+j), 0 < x < R-i \right. \quad (11)$$

Subsequently, the extraction of the stationarity value according to Equation 7 is depicted in the following square matrix

$$ {}_{j=1}^{i}\left|Z(M(j,x), N(j,x)) = \right. \begin{bmatrix} Z_{1,1} & Z_{1,2} & . & . & Z_{1,\mathrm{i}-1} & Z_{1,i} \\ Z_{2,1} & Z_{2,2} & . & . & Z_{2,i-1} & Z_{2,i} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ Z_{j-1,1} & Z_{j-1,2} & . & . & Z_{j-1,i-1} & Z_{j-1,i} \\ Z_{j,1} & Z_{j,2} & . & . & Z_{j,i-1} & Z_{j,i} \end{bmatrix} \quad (12)$$

Step2. Then, the matrix ${}_{i=1}^{i}[A_i] = {}_{j=1}^{i}\left|\bar{Z}(M(j,x), N(j,x))\right.$ is produced, along with a second matrix using the same procedure

$$ {}_{i=1}^{i}[B] = {}_{j=1}^{n}\left|\bar{Z}(M(j,y), N(j,y))\right. \quad (13)$$

where $0 < y < R - n$ is produced to construct a correlated pair of matrices.

Step3. Thereafter, aiming to produce a smoothing procedure in the data of matrices ${}^{i}_{i=1}[A_i]$ and ${}^{i}_{i=1}[B_i]$, a cumulative moving average (*CMA*) procedure is submitted as follows:

$$CMA_{n+1} = CMA_n + \frac{{}_{i=n}^{i}[A_{n-1}] - CMA_n}{n+1}, 1 < n < i \quad (14)$$

and

$$CMB_{n+1} = CMB_n + \frac{\mathop{}\limits_{i=n}^{i}\left[B_{n-1}\right] - CMB_n}{n+1}, 1 < n < i \qquad (15)$$

Then, the new matrices are:

$$\mathop{}\limits_{i=1}^{i}\left[MA\right] = \left[A_1, \mathop{}\limits_{i=1}^{n-2}\left[\bar{A}_i\right], \mathop{}\limits_{i=1}^{n}\left[\bar{A}_i\right], \mathop{}\limits_{i=2}^{n+1}\left[\bar{A}_i\right], \ldots, \mathop{}\limits_{i=k-(n-1)}^{k}\left[\bar{A}_i\right], \mathop{}\limits_{i=k-(n-3)}^{k}\left[\bar{A}_i\right], A_k\right]$$

Step4. Then, $\mathop{}\limits_{c=1}^{p<i}\left[F\right] = \left[MA\right]' = 0$ and $\mathop{}\limits_{c=1}^{p<i}\left[G\right] = \left[MB\right]' = 0$ are calculated to extract the local maxima points of the graphs corresponding to the matrices [*MA*] and [*MB*].

Step5. Consequently, the differences between adjacent elements of the [*F*] and [*G*] matrices are calculated, i.e.,

$$\mathop{}\limits_{c=1}^{p-1}\left[T_1\right] = \mathop{}\limits_{c=1}^{p}\left[\left|F_c - F_{c-1}\right|\right] \wedge \mathop{}\limits_{c=1}^{p-1}\left[T_2\right] = \mathop{}\limits_{c=1}^{p}\left[\left|G_c - G_{c-1}\right|\right] \quad (16)$$

Step6. Then, the mean values of the matrices $\left[\bar{T}_1\right] \wedge \left[\bar{T}_2\right]$ are determined.

Step7. Then, the matrix $\left[W\right] = \left[\bar{T}_1, \bar{T}_2\right]$ is determined, and the standard error of the mean of the matrix [*M*] is calculated as follows:

$$s_{error} = \frac{\sqrt{\frac{\sum_{i=1}^{2}\left(W_i - \bar{W}\right)^2}{2-1}}}{\sqrt{2}} \qquad (17)$$

Step8. Finally, using a two-tailed *t*-test with *df* = 1 for $\left[\bar{W}\right]$, the below equation is obtained:

$$\left\{\begin{array}{l} lower\_\lim it = \left[\bar{W}\right] - s_{error} * t_{value} \\ upper\_\lim it = \left[\bar{W}\right] + s_{error} * t_{value} \end{array}\right\} \Rightarrow ll < \bar{W} < ul \quad (18)$$

where *ll* and *ul* are the lower and upper limits, respectively.

4. Experimental part

2.6 Data and methods

The experiment considered a dataset sourced from an online retailer active in the segment of consumer electronics[1]. The retailer's objective was to evaluate the performance of a search engine optimization intervention that was carried out to improve the overall bounce rate that the e-shop exhibits when visitors land on the website by clicking on an organic (non-sponsored) search result through Google or secondary search providers.

We gained access to the retailer's Google Analytics account and extracted data from the main landing page, which listed entry points for the different categories (e.g., digital cameras, laptops, etc.). Hourly data were obtained using the API provided by the Google Analytics backend and exported to CSV files for further processing. The resulting input data matrix corresponded to the click-stream for an approximate two-year period and had a sample size n=18288 visitor sessions. During this period, the retailer's website remained unchanged concerning visual cues and interface characteristics. We used the default computation for the bounce rates from Google Analytics and performed some preliminary analysis to ensure that during the period to be analysed, there was no technical failure (downtime) of the website that would interrupt the continuity of the time series. The graphical representation of the variation of the bounced sessions in our dataset is shown in Figure 2.

[Insert Figure 2 here]

Having acquired the data and prepared the input data series, we proceed with the implementation of the analytic procedure as described in Section 3.2. For clarity, we refer to the points of the time series by their index value, which is set from 1 to the maximum length of the data matrix (n = 18288). We outline the numerical computation of the steps that we used in the sections that follows.

[Insert Figure 3 here]

Step 1. For a random value x = 11813 with i = 60 and taking j = 1,2,3...60, a matrix M(60,11813) is constructed according to Equation 8 (see the blue line in Figure 3 and Table 1). In the same way, the mirror N of matrix M is produced:

$$\overset{i}{\underset{i,j=1}{}}\left|N\left(j,x\right)=N\left(x+j+1:x+j\right),0<x<R-i\right.$$

according to Equation 9 (see the red line in Figure 3 and Table 1). Then, the calculated degree of stationarity D_{11} (see Equation 10) is computed using the dataset D11 = 0.0209. Similarly, the other values of the matrix with its the corresponding dimensions (60 × 60) are obtained.

Step 2 Thereafter, the matrix $\overset{i}{\underset{i=1}{}}\left[A_i\right]=\overset{i}{\underset{j=1}{}}\left|\bar{D}\left(M\left(j,x\right),N\left(j,x\right)\right)\right.$ of size (1x60) is obtained. In the same way, the matrix $\overset{i}{\underset{i=1}{}}\left[B_i\right]=\overset{n}{\underset{j=1}{}}\left|\bar{D}\ \left(M\left(j,y\right),N\left(j,y\right)\right)\right.$ of size (1x60) is obtained.

Step 3. Using a cumulative procedure with a 5-point (n=5) moving average, the matrices $\overset{i}{\underset{i=1}{}}\left[MA\right]$ and $\overset{i}{\underset{i=1}{}}\left[MB\right]$ are obtained.

Step 4. According to Equation 11, the local maxima points of the graphs of the matrices [MA] and [MB] are calculated in the new matrices [F] ∧ [G] (see Table 1).

Step 5. Consequently, the differences between the adjacent elements of matrices [F] and [G] are calculated in the new matrices [T_1] ∧ [T_2].

Step6. Then, the mean values of the matrices $\left[\bar{T}_1\right]\wedge\left[\bar{T}_2\right]$ are determined (see Table 1, column: Mean).

Step7. Then, the matrix $\left[W\right]=\left[\bar{T}_1,\bar{T}_2\right]$ is determined, and the standard error of the mean of the matrix is calculated (see Table 1, column: Error).

Step8. According to Equation 14, for a 98% confidence interval with t=31.820, a p-value of 0.02 for 2% significance and W^{-} = 25.2767 (see Table 1, cell: Mean Error) which is transformed as follows:

$$\left\{\begin{matrix} lower\_\lim it = 25.2767 - 0.0222 * 31.821 \\ upper\_\lim it = 25.2767 + 0.0222 * 31.821 \end{matrix}\right\} \Rightarrow \quad 24.5703 < 25.2767 < 25.9831 \qquad (19)$$

2.7 Results

According to the experimental procedure and taking into account the results presented in Table 1 and the resulting transformation of the time series depicted in Figure 4, an apparent periodicity of the applied processing is observed. This observation is focused on the measure of the differentiated positions of the local maxima points and puts great emphasis on the dominant query, by subjecting the task on finding the necessary sample size that significantly captures the observed periodicity of the time series.

[Insert Table 1 here]

The results of the experimental procedure (step 8), provide a sample size s=25, which can be interpreted as that the variation of the stationarity degree has stable maxima periodically for a set of 25 data points bounce rate samples. Considering that our data represents hourly bounce rates, the benchmark data suggest a window of 25 hours for the evaluation of interventions for content optimization.

[Insert Figure 4 here]

In more detail, the above calculations are achieved via Equations 4-18 using the mean difference between the matched pairs technique. The matched data pairs were obtained in a random way according to Equation 13 and had a scalable range from 60-200 with the step increment of 5, that is, 30 matched data pairs were created in total (see Table 1). Therefore, for each matched pair—for example, the values at length 60, which are depicted between the data x=[11813, 11873] and y=[5391, 5451]—the local peaks are calculated (see Figure 1, start). Then, as the mean matched data pairs are determined, the mean value of the distance between each local peak value is obtained. In the start case, the number of peaks for the pair x and y data set is two (2), and the mean distances are 28 and 29, respectively. Therefore, the M.D.B.M.P is calculated (see Equation 17) from the difference between the above means, which, in Table 1, is depicted as the error (e=0.5). In the same way, the M.D.B.M.P results are calculated through the last data set (see Figure 1, end). Additionally, in the Appendix, the graphical transformations of the above calculations are depicted for the 30 matched data pairs.

5. Discussion

2.8 Theoretical Implications

This study contributes to the expert and intelligent systems literature by demonstrating a robust method for subsampling time series data, which are critical in decision making. In particular the application of the stationarity detection algorithm as demonstrated in previous work (Poulos, 2016) allows for evaluating more complex problems in business practice such as measuring website prominence (Papavlasopoulos, 2019; Poulos, Papavlasopoulos, Kostagiolas, & Kapidakis, 2017) as well as dimensionality reduction in text analytics (Poulos, 2017).

The particular implication in researching patterns in high-frequency time series data can also be applied in patterns of web queries such as those in Google Trends. This extraction of the periodical non-stationarity features of time series can complement existing approaches for novelty detection in scientific literature utilizing the patterns on prominent keywords appearing in scientific publications (Papavlasopoulos, 2019). While the application in this context concerns consumer activity it confirms previous results that proxy a visitor’s activity using the search queries and the related keywords that have been utilized. This method is implemented via the same algorithm, with the only exception being that parameter M is fed with a multidimensional data structure (see Equation 10). While this study aims to assess when periodicity can distort the outcomes of a marketing intervention, it confirms similar results with the study of Papavlasopoulos (2019) which investigates when a keyword time series gives non-stationarity peaks. As such, asserting the condition of a non-stationary categorical time series, yields goodness of fit in the prediction issue.

A further implication that can be investigated further in future studies comes from the aggregation of individual time series using grouping factors such as product category and brand.Poulos et al., (2017) demonstrated that asserting stationarity of aggregated time series of search keywords using Google Trends can be achieved, using an example of publishing houses and their corresponding publications. In a similar manner, the data type for parameter M (Equation 10) needs to modified to represent 2-dimensional groupings.

The algorithmic process presented here can also be used in the context of text analytics (Poulos, 2017), where the possible relationship between the syntactic property of a text sample and the stationary variation of the time series that produces the text, can be asserted. This can inform additional dimensions, such as the case of recommendations based on semi-structured data such as those on online reviews (Korfiatis and Poulos, 2013).

Therefore, application of Equations 1-13 to the data type (M) yields the new modified time series A and B (see Equation 12 and 13), which in turn, leads us to the technique for calculating the periodicity of various type of high-frequency data as the ones described above (see Equations 14-18).

2.9 Implications for practice

Trust in evidence-based methods for evaluating marketing interventions, such as A/B testing, is gaining momentum for both managers and marketers. However, the pitfalls associated with misuse of this decision-making instrument are not well understood by managers and analytics experts since the prevalence of software tools provides an out-of-the-box solution, which may not be optimal (Dmitriev, Frasca, Gupta, Kohavi, & Vaz, 2016). Anecdotal examples of negative results induced by Type I and II errors are known in the industry, and careful consideration of the time-dependent properties of marketing metrics (e.g., stationarity) by decision-makers is important. Making a healthy choice between alternative interventions guided by customer-driven interactions is an important example of analytical maturity (Davenport & Harris, 2007) and is independent of the organizational size. Several examples of A/B testing scenarios consider interventions on web spaces owned by small- and medium-sized companies. As such, being able to reliably ascertain the impact of these interventions on conditioned time series can also give a competitive advantage in capturing consumer attention. However, as Kohavi et al. (2012) state, experimentation is not a *panacea* for everyone, and its assumptions should be well understood when interpreting results of high economic significance. In this study, an experimental method is attempted to override the aforementioned unsafe decision assumptions of the A/B method. To achieve the above objective, the data were applied to the algorithm in Equations 4-18 using matched data pairs as described in section 4.2, which yields a statistical significance test. In the analysis of the results, a 25-hour sample time period has been derived for the data set. Furthermore, this study places a large emphasis on examining the nature of the time series and the stage from which the data are retrieved. Practical considerations such as the assumptions that accompany the time series data retrieved at the initial stage, or the impacts of any demand peaks (e.g., due to marketing campaigns running in parallel) can be validated through the procedure outlined here.

As discussed in the previous section, data sets from Google Trends and bounce rate could yield this degree of periodicity. This consideration is based on the assumption that the nature of the data is depicted in the local peaks of the transformed time series that come from the stationarity process.

6. Conclusions, Limitations and Future Research

In this paper, the potential to extract periodical stationarity exhibited in a conditioned time series of bounce rates was investigated and evaluated using a benchmark dataset. Controlling for stationarity is a significant problem in analytics and forecasting, in which a time series is analysed for the levels of differences. Using the appropriate transformations with a new algorithm for calculating the stationary distance, our approach can be useful in the evaluation of marketing interventions, such as those in A/B testing scenarios. This distance is based on a novelty stationary ergodic process, which rests on the consideration that the stationary series presents reversible symmetric features and is calculated using the dynamic time warping (DTW) algorithm in a self-correlation procedure. The results of the benchmark test performed in the experimental part of this paper present the very clear and logical periodicity of the discussed method by utilizing the measures of differences in the positions of local maxima points during the segmentation of the conditioned series.

While our approach was operationalized for a conditioned time series, our method does not take into account causal influences from other time-dependent processes that may affect the behaviour of the evaluated metric (in our case, bounce rates) or psychological cases related with shopping cart abandonment (Huang et al., 2018). Such a case could arise when transitions from stages are considered (e.g., bounces after the second click). In this aspect, our analysis is therefore agnostic to important user characteristics, such as repeated visits and view-through conversions, which require a higher-order data structure than that considered in this study.

In addition, our analysis places a large emphasis on the issue of finding the necessary sample size that significantly satisfies the observed periodicity of a time series of bounce rates in an e-commerce scenario. Future work on other types of conditioned time series represented in web analytics, such as page views, pages/visits, percentages of new visits, and average times on sites, is also important as well as demand patterns in supply chains (Zissis et al., 2015). This work will involve studying more sophisticated time series processing and template matching techniques as well understanding the distributional characteristics of these metrics.

Data accessibility

No data are provided together with the manuscript

Notes

[*] For this study, the retailer has requested to remain anonymous.

References

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics, 9,* 247–274. doi:10.1214/14-aoas788

Clifton, B. (2012). *Advanced web metrics with Google analytics*. Indianapolis, Indiana: John Wiley & Sons.

Cortez, P., Rio, M., Rocha, M., & Sousa, P. (2012). Multi-scale internet traffic forecasting using neural networks and time series methods. *Expert Systems, 29*, 143–155. doi:10.1111/j.1468-0394.2010.00568.x

Danaher, P. J., & Rust, R. T. (1996). Determining the optimal return on investment for an advertising campaign. *European Journal of Operational Research, 95,* 511–521. doi:10.1016/0377-2217(95)00319-3

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business Press.

Dmitriev, P., Frasca, B., Gupta, S., Kohavi, R., & Vaz, G. (2016). Pitfalls of long-term online controlled experiments. In *2016 IEEE international conference on big data (big data)* (pp. 1367–1376). Washington, DC, USA: IEEE.

Downing, D. J., Fedorov, V. V., Lawkins, W. F., Morris, M. D., & Ostrouchov, G. (2000). Large data series: Modeling the usual to identify the unusual. *Computational Statistics & Data Analysis, 32*, 245–258. doi:10.1016/s0167-9473(99)00079-1

eCommerce Europe. (2016). E-commerce benchmark and retail report. Retrieved from https://www.ecommerce-europe.eu/app/uploads/2016/06/Ecommerce-Benchmark-Retail-Report-2016.pdf

Edvardsson, B., Kristensson, P., Magnusson, P., & Sundström, E. (2012). Customer integration within service development—A review of methods and an analysis of insitu and exsitu contributions. *Technovation, 32*, 419–429. doi:10.1016/j.technovation.2011.04.006

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hoban, P. R., & Bucklin, R. E. (2015). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research, 52*, 375–393. doi:10.1509/jmr.13.0277

Huang, G. H., Korfiatis, N., & Chang, C. T. (2018). Mobile shopping cart abandonment: The roles of conflicts, ambivalence, and hesitation. *Journal of Business Research, 85*, 165–174.

Jeziorski, P., & Moorthy, S. (2017). Advertiser prominence effects in search advertising. *Management Science, 64*, 1365–1383. doi:10.1287/mnsc.2016.2677

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 786–794), Beijing, China: ACM.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery, 18*(1), 140–181. doi:10.1007/s10618-008-0114-1

Korfiatis, N., Poulos, M. (2013). Using online consumer reviews as a source for demographic recommendations: A case study using online travel reviews. *Expert Systems with Applications 40*, 5507–5515.

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology, 25*(2), 115–126. doi:10.1080/01449290500330448

Marr, B. (2010). *The intelligent company: Five steps to success with evidence-based management*. New York, NY: John Wiley & Sons.

Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *European Journal of Operational Research, 241*, 583–595. doi:10.1016/j.ejor.2014.08.029

Murthy, P., & Mantrala, M. K. (2005). Allocating a promotion budget between advertising and sales contest prizes: An integrated marketing communications perspective. *Marketing Letters*, *16*(1), 19–35. doi:10.1007/s11002-005-1138-6

Papavlasopoulos, S. (2019). Scientometrics analysis in Google trends. *Journal of Scientometric Research, 8*(1), 27–37. doi:10.5530/jscires.8.1.5

Pfeffer, J., & Sutton, R. I. (2006). Evidence-based management. *Harvard Business Review, 84*(1), 62.

Plaza, B. (2011). Google analytics for measuring website performance. *Tourism Management, 32*, 477–481. doi:10.1016/j.tourman.2010.03.015

Poulos, M. (2016). Determining the stationarity distance via a reversible stochastic process. *PLoS One*, *11*(10), e0164110. doi:10.1371/journal.pone.0164110

Poulos, M. (2017). Definition text's syntactic feature using stationarity control. In *2017 8th International conference on information, intelligence, systems & applications (IISA)* (pp. 1–5), Larnaca, Cyprus: IEEE.

Poulos, M., Papavlasopoulos, S., Kostagiolas, P., & Kapidakis, S. (2017). Prediction of the popularity from Google trends using stationary control: The case of STM publishers. In *2017 Fourth international conference on mathematics and computers in sciences and in industry (MCSI)* (pp. 159–163), Corfu, Greece: IEEE.

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis, 11*(5), 561–580. doi:10.3233/ida-2007-11508

Sauter, V. L. (2014). *Decision support systems for business intelligence*. Hoboken, NJ.: John Wiley & Sons.

Sculley, D., Malkin, R. G., Basu, S., & Bayardo, R. J. (2009). Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1325–1334), Paris, France: ACM.

Sculley, D., Otey, M. E., Pohl, M., Spitznagel, B., Hainsworth, J., & Zhou, Y. (2011). Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 274–282), San Diego, California, USA: ACM.

Sharifdoost, M., Mahmoodi, S., & Pasha, E. (2009). A statistical test for time reversibility of stationary finite state markov chains. *Applied Mathematical Sciences, 52*, 2563–2574.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28. doi:10.1257/jep.28.2.3

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences U S A, 113*, 7310–7315. doi:10.1073/pnas.1510479113

Vaughan, L., & Yang, R. (2013). Web traffic and organization performance measures: Relationships and data sources examined. *Journal of Informetrics*, *7*, 699–711. doi:10.1016/j.joi.2013.04.005

Wells, J. D., Valacich, J. S., & Hess, T. J. (2011). What signal are you sending? How website quality influences perceptions of product quality and purchase intentions. *MIS Quarterly, 35,* 373–396. doi:10.2307/23044048

Yang, S., & Ghose, A. (2010). Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science, 29,* 602–623. doi:10.1287/mksc.1090.0552

Zissis, D., Ioannou, G., & Burnetas, A. (2015). Supply chain coordination under discrete information asymmetries and quantity discounts. *Omega*,*53*, 21-29.

Tables

Table 1. Results of the experimental procedure. *l* corresponds to the sampling length and $N(X_i)$ and $N(Y_i)$ are the numbers of peak values of *X* and *Y*, respectively, with μ and e corresponding to the mean and standard error, respectively.

l	Range of x	Range of y	$N(X_i)$	$\mu(X)$	$N(Y_i)$	$\mu(Y)$	e
60	[11813,11873]	[5391,5451]	2	29	2	28	0.5
65	[16154,16219]	[586,651]	3	26.5	3	27	0.25
70	[7459, 7529]	[6487,6557]	3	27	3	25.5	0.75
75	[13014,13089]	[13519,13594	3	25	3	25.5	0.25
80	[12060,12140	[12830,12910]	3	25.5	3	26.5	0.5
85	[4693,4778	[11555,11640]	3	26	3	25.5	0.25
90	[11137,11227	[2765,2855]	4	25	4	25.33	0.17
95	[2023,2118	[8473,8568]	4	25.33	4	25.67	0.17
100	[16316,16416	[5787,5887]	4	26.33	4	25.33	0.5
105	[9950,10055	[3805,3910]	4	25.33	4	25	0.17
110	[12772,12882	[4337,4447]	4	25.67	4	25.67	0
115	[8602,8717	[11885,12000]	5	25	5	25.75	0.38
120	[15146,15266	[16308,16428]	5	25.25	5	25.25	0
125	[14293,14418	[4323,4448]	5	25	5	25	0
130	[13843,13973	[4140,4270]	5	24.75	5	25	0.13
135	[15798,15933	[5950,6085]	5	25.5	5	25.25	0.13
140	[3343,3483	[4269,4409]	6	25.2	5	24.8	0.2
145	[10473,10618	[8046,8191]	6	25	6	25	0
150	[5979,6129	[14125,14275]	6	24.8	6	24.6	0.1
155	[9950,10105	[9346,9501]	6	24.6	6	24.8	0.1
160	[15593,15753	[4860,5020]	7	24.67	7	24.83	0.08
165	[13554,13719	[13492,13657]	7	24.67	7	24.67	0
170	[6810,6980]	[10165,10335]	7	24.83	7	24.67	0.08
175	[1358,1533]	[966,1141]	7	24.5	7	24.67	0.08
180	[1768,18048]	[1011,1191]	7	24.67	7	24.67	0
185	[16719,16904]	[2326,2511]	8	24.57	8	24.71	0.07
190	[17000,17190]	[15200,15390]	8	24.57	8	24.57	0

195	[214,409]	[6035,6230]	8	24.86	8	24.57	0.14
200	[2904,3104]	[14218,14418]	8	24.57	8	24.57	0

Index of Figures

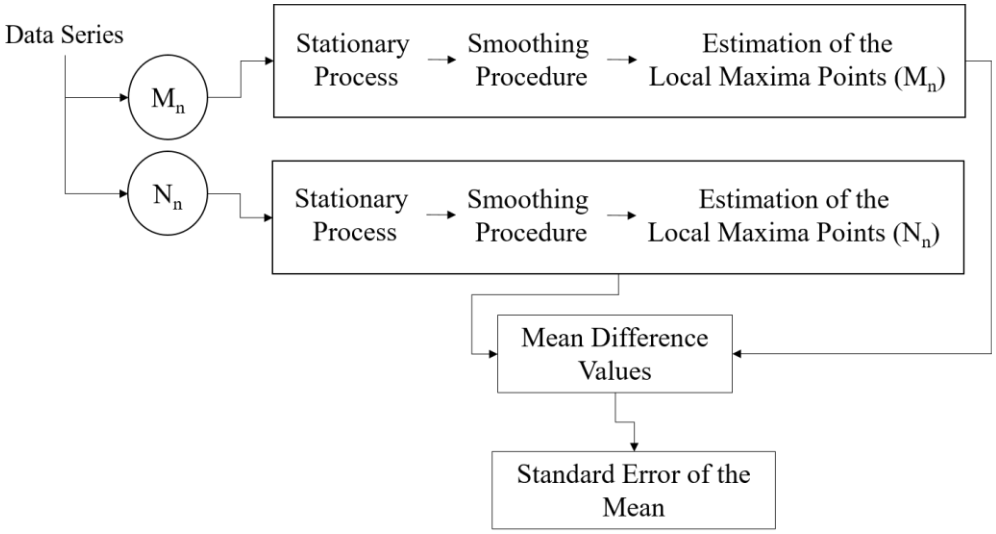

Data Series
M_n
N_n
Stationary Process
Smoothing Procedure
Estimation of the Local Maxima Points (M_n)
Stationary Process
Smoothing Procedure
Estimation of the Local Maxima Points (N_n)
Mean Difference Values
Standard Error of the Mean

Figure 1

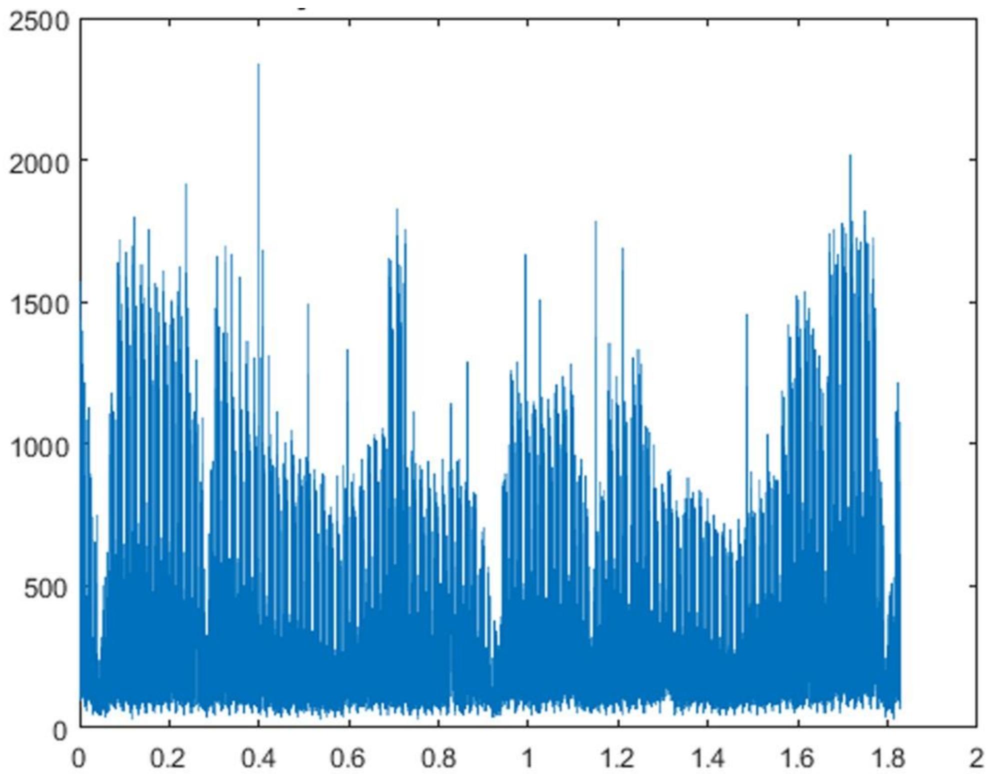


Figure 2

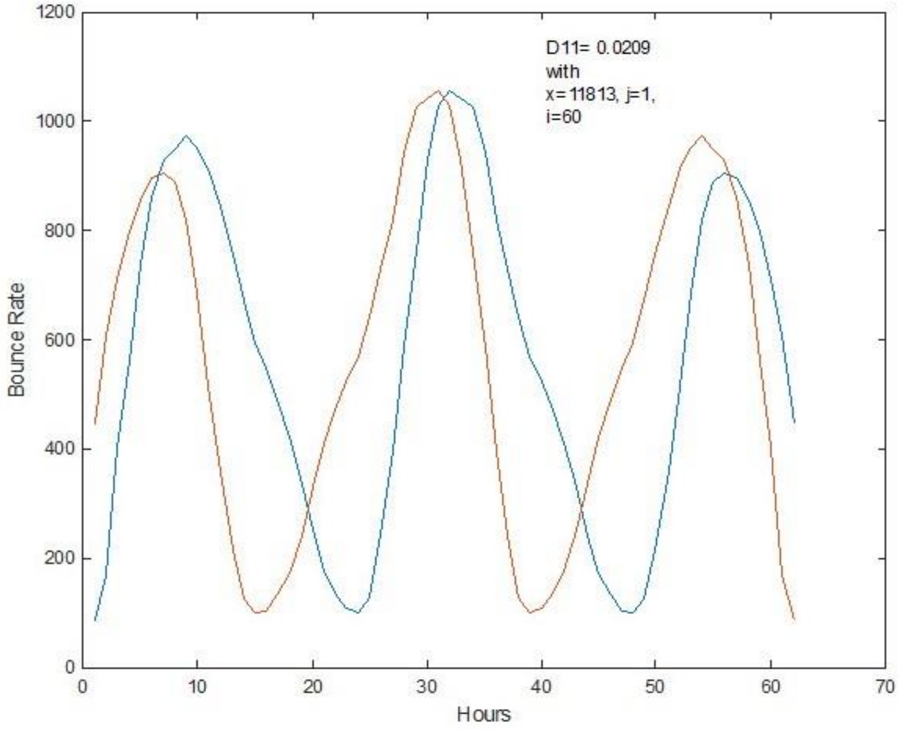
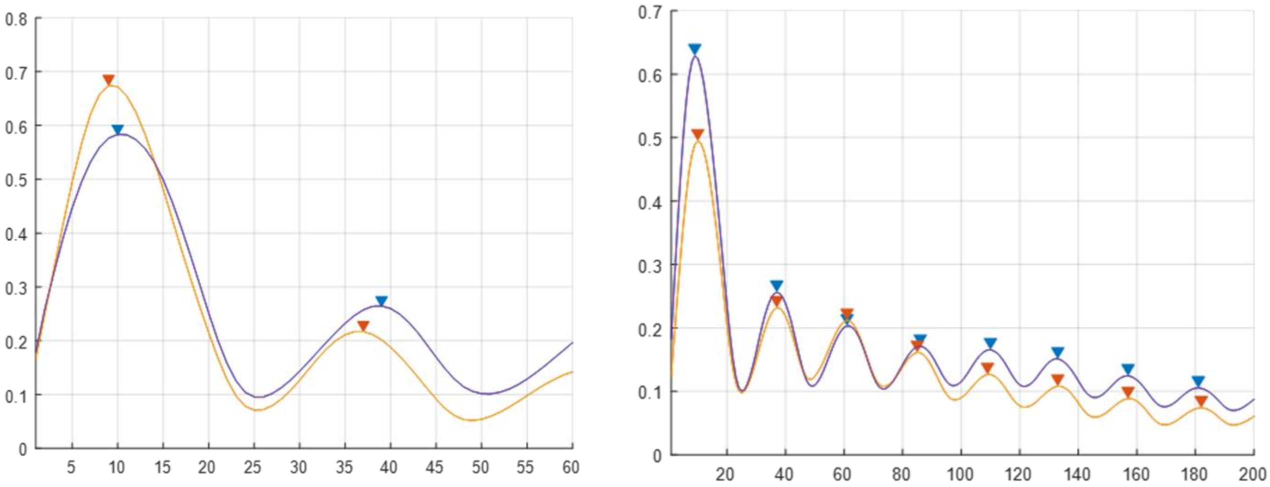

D11= 0.0209
with
x= 11813, j=1,
i=60
Bounce Rate
Hours
0
200
400
600
800
1000
1200
0
10
20
30
40
50
60
70

Figure 3



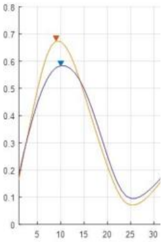
(1) Start

(2) End

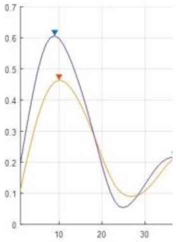
Figure 4

Appendix: Graphical Transformation for the experimental section

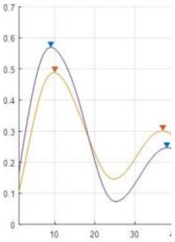
The segmentation of the resulted time series and the identification of the local maxima as presented in Table 1, is performed with a step of size of *s* = 5. For each step, the resulted length (*l*) transforms the data series and its reverse, as depicted in the subsequent panels.
Transformation sequence for steps 60 ≤ *l* ≤ 105



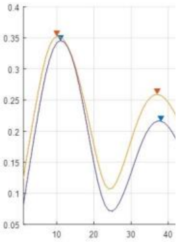
l = 60



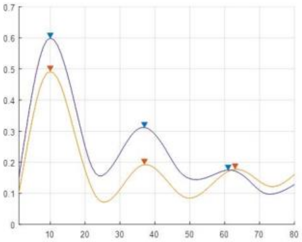
l = 65



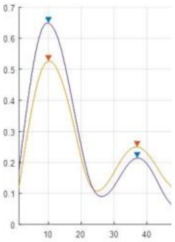
l = 70



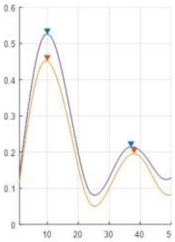
l = 75



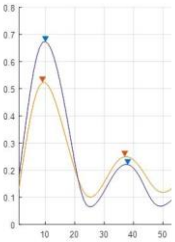
l = 80



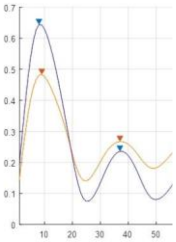
l = 85



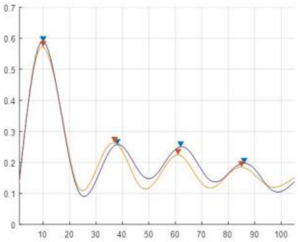
l = 90



l = 95

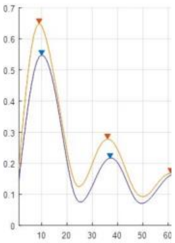


l = 100

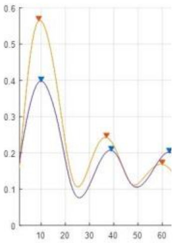


l = 105

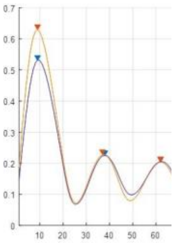
Transformation sequence for steps 110 ≤ *l* ≤ 155



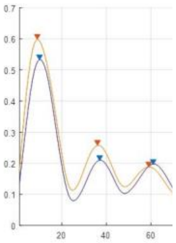
l = 110



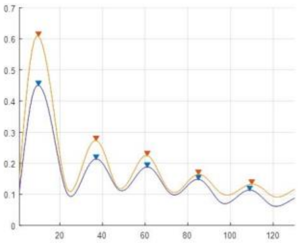
l = 115



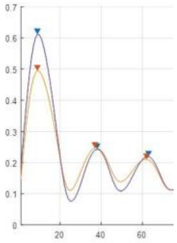
l = 120



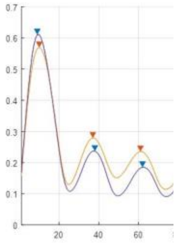
l = 125



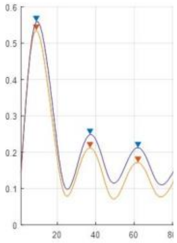
l = 130



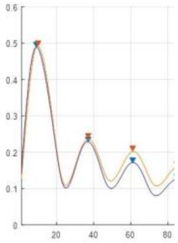
l = 135



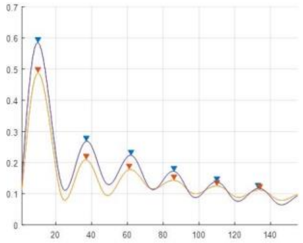
l = 140



l = 145

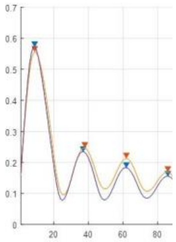


l = 150

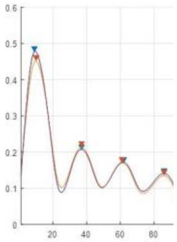


l = 155

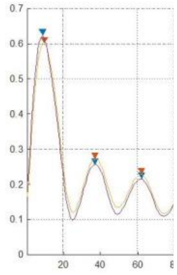
Transformation sequence for steps 160 ≤ *l* ≤ 200



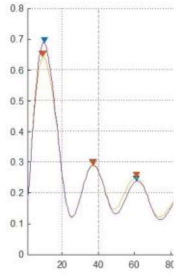
l = 160



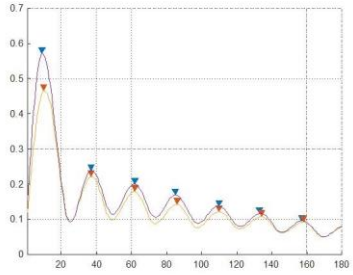
l = 165



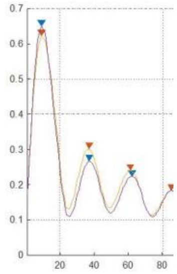
l = 170



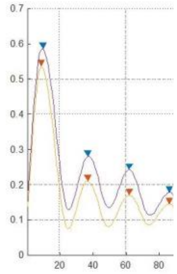
l = 175



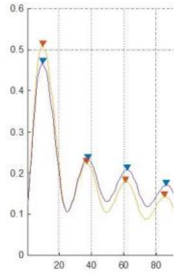
l = 180



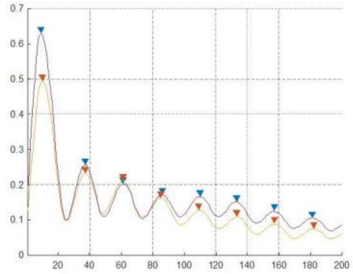
l = 185



l = 190



l = 195



l = 200