TELKOMNIKA, Vol.17, No.4, August 2019, pp.1723~1730
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i4.11956

Dominated destinations of tourist inside iraq using personal information and frequency of travel

Rula Amjad*[1], Muayad Sadik Croock[2]
[1]Institute of informatics for Postgraduate studies, Iraq
[2]Computer Engineering Department, University of Technology, Iraq
*Corresponding author, e-mail: eng_rula_amjed@uoitc.edu.iq[1], muayadkrook@yahoo.com[2]

Abstract

Tourism today is one of the most important economic and social sectors in the world, which plays a prominent role in the development of countries. This importance has grown as an industry through the social media networks.In this paper, a proposed method has been introduced distinguish the main factors that impact the Frequency of Travel (FoT) among Iraq local tourists. Application (API) graphics and scrapy are utilized to collect information from TripAdvisor social network in a period of (2015-2018). The collected information are reprocessed and coded for the specified nominal data using tied rank. It is important to note that the adopted technique does not lose any data about the attribute and brings different properties beforehand obscure. Data mining ordinal logistic regression is used to extract user's behavior upon local tourism in Iraq. The expected outcome of this work is to discover out the effect of personal information and the type of places on the selection of the local touristic places in Iraq. The collected information was exploited to know the preferred local touristic trends, because there are no statistics on the number of domestic tourists in Iraq. The proposed model was used for analyzing personal information and types of preferred tourism places as a factors affecting frequency of travel in Iraq. The obtained results show the prediction of preferred touristic places by tourists in Iraq.

***Keywords**: FoT, IoT, ordinal logistic regression, scrapy, tied ranks, TripAdvisor*

1. Introduction

The importance of tourism as a catalyst for sustainable development is an important economic requirement to stimulate investment in natural, environmental, religious and heritage tourism [1]. The increasing role of social media in tourism has been gradually increased in research area. Social media is currently acting as an important responsibility in numerous aspects of tourism. It can enable the understanding of the user's interests and presenting user pattern behaviors on a variety of Travel and Tourism services particularly tour attractions and point of interest [2]. This can be considered to support the tourism system by focusing on superlative practices for interacting with consumers. The significant difficulties, adopted by tourism recommendation systems, are used to cover the verifiable connections, existed between each user and related points of interests [3]. Social media turns out to be progressively an approach to express users feeling and their interest, consumers can progressively impact other consumers with their own opinions and feeling. Since social media is minimal effort and inclination free, it act an utility for marketing communications [4].

Numerous Approaches was utilized to clarify movement support from social media mainly shows up from machine learning and data mining fields.Conditional random fields, logistic regression, and integrity models have been utilized to forecast several aspects of liveliness participation. Diverse classification, clustering and regression algorithms have been adopted to classify activity choice sample and cluster users depend on their activity patterns [5]. In the absence of statistics and useful information adopted for tourism marketing in Iraq, there is an urgent need to extract useful information from multiple sources and manage it correctly to guide its use in the labor market. Over time, the use of social networks has become a means of expressing opinions, preferences and tourism experiences.

Our contribution through this research is to adopt these data with their various types after reprocessing and manage them well using the appropriate data mining algorithms. The proposed method using tied ranking with ordinal logistic regression to predict the factors

Received December 3, 2018; Revised March 23, 2019; Accepted April 20, 2019

that can affect the tourism sector in Iraq and what are the factors influencing this field in order to focus on the development of this industry.

The aim of this work is to highlight the role of the information, collected from social networking in tourism field as it is considered as one of the electronic tourism marketing extensions. This information can be directed and managed to find out the places of preference of the domestic tourists in Iraq and what areas are the most focused on. This is to guide the local tourism market in the right direction and develop such market depending on the users' desires. The social networks have been considered as references for obtaining tourist information based on the opinions of friends on the network. In addition, the posts and comments about tourism are considered as well as the frequency of travel that represents how often the user makes check in. frequency of Travel is an important factor in determining the demand for tourism [6].

2. Research Method

Different researchers have been involved in the developing of tourism systems by proposing numerous methods and algorithms. In [7], the authors presented the connection between attributes composes and a person's appearance. The probability of visiting a tourist some kind of touristic places has been adopted in prediction of the prefer places. Furthermore, the impact of gender attributes on tourism behaviors have been considered in the research. In [8], consumer behavior (CB) has been analyzed to exam the related impacts on modern tourism. This analysis probed the CB in the available literature in three main tourism journals from 2000 to 2012. Upcoming research on nine key concepts were examined, including decision-making, inspirations, self-idea and identity, opportunity, attitudes, desires, states of mind, and observations.

In [9], the methodology was present for extracting the city's tourism area in detail. Methodology consists of gathering social media data of meditation region, point of intrest data, users' profile data, and geo-data (with longitude-latitude coordinate) and separating the city space into grids. Finally applying a community detection algorithm to find powerfully combined grids; and using pattern analysis methods to extract and interpret the tourism districts of the city

In [10], Based on huge information in tourism examine, the tourism-related huge information are classified into three essential classes. The first one was information (created by clients) that included the web literary information and online photograph information. The second one was the gadget information (by gadgets) that involved GPS information, versatile meandering information, and Bluetooth information. The third class was the exchange information (by activities that has a web sought information, page visiting information, and web based booking information.

The paper of [11] utilized the model of auxiliary similitude coefficients to break down the transportation decision component of travelers, afterward take air travelers for instance, and set up the variables influencing model of travelers' movement decision in view of ordinal-logit method. The examination demonstrated that the value, time, solace and flexibility were the four fundamental components for traveler in travel options.Age and sexual orientation were traveler's physical characteristics that adopted in this work; however their effect on traveler travel decision ability was not noteworthy. Education and wage level had huge emphatically affect, particularly the salary levels.

A multi-staged social media expository structure was developed in [12] .data crawling was evaluate the arrangement of information interactions between the members of a tourism organization’s social network community and recognized important actors and information content within the social network. The smart tourism industry environment is related with administrative use of Facebook for enacting local travel industry in Korea [13, 14].

Feacebook API was employed in [15] to retrieves information such as age, sex, work, location, education, relational status from user’s TripAdvisor account. The research proposed theoretical frameworks and application for designing Adaptive tourism system for recommendation. Another study was developed the sense of the typical importance of travel to hold social respond provided by peer groups on social media platforms. Predictive validity was assessed by examining the relationship between social respond and plan to visit Cuba over three period horizons (15 and 10 years) and by examining the relationship between the social return scale in association with theory of planned behavior form (e.g., attitudes, perceived behavioral control, and objective model) [16].

In [17] the advantages and disadvantages of utilizing social media information in travel requisition pattern was investigated. An essential possibilities for utilizing web-based social networking information to develop structure for assessing travel request, overseeing activity and long haul arranging purposes. The main challenges issue of utilizing social media information identify with the intricacy of removing profitable from the enormous data. For this reasons work of expert content and information mining procedures was required.

3. Hypothesis

Different components related to tourist have been considered in the proposed method as hypothesis as follows:

- Educational achievement: People who are in high educational achievements have been classified with a high level of recurrence in visiting tourism places. Therefore the FoT factor is notably high for them.
- Gender: Men are more reluctant than women on tourist sites due to the nature of conservative Arab societies. In addition, the participation of men for public information more than women.
- Work: type of work can affect the person's income and thus the nature of the touristic sites due to the economic reasons.
- Age: People with younger ages are the most frequently users of social networking sites in comparison with older people. This can tend to spread their activities including tourism more than older people.
- Type of interest (type of preferred places): Statistics of the Iraqi Tourism Authority in the last three years indicate that this country depends currently on religious tourism and environmental tourism, especially in Iraqi Kurdistan. This can affect the number of domestic and foreign tourists.

4. Conceptual Model

The conceptual model of Figure 1 has been adopted in the proposed method. This model deals with the effect of personal information of tourist and type of tourism places, interested by tourist to control variables on the frequency of travel (FoT).

5. Proposed Model

This section includes the research concepts followed by the proposed method to perform the required model.

5.1. Description of the Data

The data set was collected from the public posts of 524 TripAdvisor user’s profiles using scrapy [18]. Scrapy is an open source and community oriented system for extracting the information required from sites. The Graph API [19] is also used to perform the required aim. Web Crawler, also known as Web Spider or Web Robot or an ant or automatic indexer [20], is one of the core concepts of "Internet of Things" [21]. The Data was collected within the time period from 2015-2018 .Time series from (2015-2018) has been chosen for two reasons, the first of which is the relative improvement in the security situation in Iraq within this period. The second reason is the increase of using of social networking in the last seven years in Iraq [22]. The collected information from user profiles represents independent variables (age, gender, academic achievement, and work). Point of Interest (POI) represent touristic sites have been classified as (religious, environmental, culture and adventure).Depending on the number of trips the user has shared on Trip Advisor, it is possible to guess whether the user prefers tourism in the local areas or not depending on the following values (Very Likely, Likely, Somewhat Like, and Unlike). Table1 represents the independent variables in data set and their values.

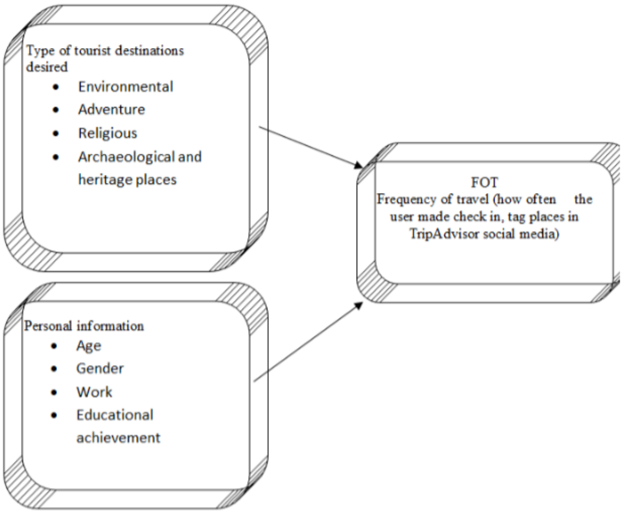

Type of tourist destinations desired
Environmental
Adventure
Religious
Archaeological and heritage places
Personal information
Age
Gender
Work
Educational achievement
FOT
Frequency of travel (how often the user made check in, tag places in TripAdvisor social media)

Figure 1. The conceptual model

5.2. FOT and Analytical Technique

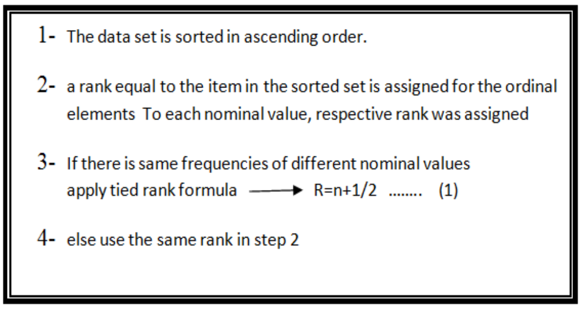
In this work, users' (check in) was adopted to identify the FoT. How often they travelled to tourist places in order to quantify frequency of travel. This in turn can help in studying the factors that influence the decision of travel and tourism inside Iraq.In order to extract and utilize the hidden useful information; preprocessing and data mining algorithm was adopted.

Table1. Independent Variables in the Dataset

Parameters	Value	Parameters	Value
Work	**N**	**Prefer Environmental _place**	**N**
employee	312	YES	343
Un employed	7	NO	181
entrepreneurs	33	**Prefer Religious places**	**N**
house wife	33	YES	63
retired	7	NO	461
Gender	**N**	**Prefer Adventure places**	**N**
female	215	YES	205
male	309	NO	319
Age	**N**	**Prefer Culture places**	**N**
15_19	40	YES	115
20_24	92	NO	409
25_29	92		
30_34	101		
35_39	76		
40_44	84		
45_50	19		
>50	20		
Educational achievement	**N**		
B.Sc.	298		
diploma	17		
high school	65		
Msc	97		
Phd	47		

5.2.1. Data Preprocessing (Tied ranks)

For data preprocessing an approach for coding nominal data was anticipated [23]. The numerical outcomes of interpretation are changed by their ranks. Nominal data cannot be sorted according to their values, but it can be grouped according to identical values. In the n–element subset consisting of identical elements, these elements maybe numbered from 1 to n. For each of them can be assigned a rank that is equal to the average value of these numbers. More numerous elements will have higher rank than less numerous elements. The numerical results of observations are replaced by their ranks. On the other hand, it may happen that in the sorted set there are different elements with the same values. In this case, the ranks assigned to identical values should be the same. Such elements receive rank that is equal to their average position in the sorted set. These are so–called tied rank. Figure 2 represents the steps of ranking procedure used in this work:



1- The data set is sorted in ascending order.

2- a rank equal to the item in the sorted set is assigned for the ordinal elements To each nominal value, respective rank was assigned

3- If there is same frequencies of different nominal values apply tied rank formula ⟶ R=n+1/2 .... (1)

4- else use the same rank in step 2

Figure 2. Steps of tied ranks procedure

according to the formula (1), ranking results are shown in Table 2.

Table2. Tied Ranks of Numerical Data

Parameters	Value		Parameters	Value	
Work	**N**	**Rank**	**Prefer Environmental _place**	**N**	**Rank**
unemployed	7	1.5	YES	181	0
retired	7	1.5	NO	343	1
entrepreneurs work	33	3.5	**Prefer Religious places**	**N**	**Rank**
house wife	33	3.5	YES	205	0
student	132	5	NO	319	1
employee	312	6			
Gender	**N**	**Rank**	**Prefer Culture places**	**N**	**Rank**
female	215	1	YES	115	0
male	309	2	NO	409	1
Age	**N**	**Rank**			
>50	20	1			
15_19	40	2			
20_24	92	3.5			
25_29	92	3.5			
30_34	101	5			
35_39	76	6			
40_44	84	7			
45_50	19	8			
	524				

5.2.2. Data Mining Algorithm (Ordinal Logistic Regression)

Ordinal logistic regression is a standout amongst the most prevalent strategies for investigating ordinal result factors. With ordinal data, it is natural to consider probabilities of cumulative events, like specific score or worse [24].Ordinal logistic regression is used as solution to classification problems that use a linear combination of the observed features and some problem-specific parameters. This is to estimate the probability of each particular value of the dependent variable. Figure 3 shows the progression steps of model workflow.

6. Empirical Model

The following empirical model, represented as an equation, is used to test the proposed hypothesis [25]:

$$FoT = \beta 1 + \beta 2Age + \beta 3Gen + \beta 4Edu + \beta 5Work + \beta 6Hertige + \beta 7Env + \beta 8Relg + \beta 8Adv + U \quad \ldots\ldots \quad (2)$$

where,
- FoT = Frequency of travel

- Age = Age of the users
- Gen = Gender
- Edu = education
- Work = type of work
- Env = preferred environmental places
- Heritage = preferred heritage places
- Relg = preferred religious places
- Adv = preferred Adventure places
- U = Error term

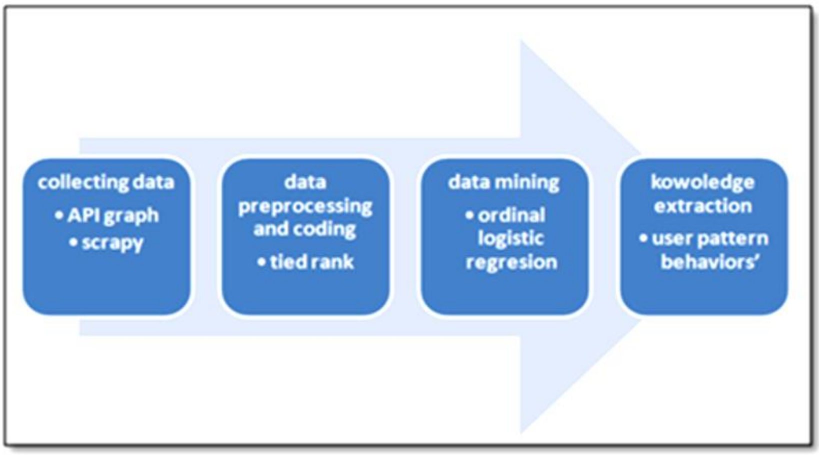

collecting data
• API graph
• scrapy
data preprocessing and coding
• tied rank
data mining
• ordinal logistic regresion
kowoledge extraction
• user pattern behaviors'

Figure 3. Model workflow

7. Results and Discussions

For predicting an ordinal variable, ordinal logistic regression was used as we mentioned earlier. Ordinal logistic regression rates exist on subjective scale where only the comparative ordering between diverse values is considered.in this work the values of class are (very likely, likely, somewhat like, and unlike). The results of ordinal logistic regression are given in Table 3 as an odd ratio [26] and the accuracy when using tied ranks and simple ranks.

Odd ratio supposes that tourists, who prefer "Environmental tourism places", are 1.22 times better estimation to have a high FoT.While, tourists, who prefer "Adventure tourism places", are 1.026 times and religious places are 1.0245 times better estimation to have a high FoT. In terms of academic achievement, the tourists with the bachelor's and master's degrees have a "somewhat like" proportion of FoT more than the holders of the doctorate and students inside Iraq. For work type of employees and entrepreneurs is also accounted to have higher FoT than others. As for gender, the hypothesis that supposes males go to tourist places more than females. Travel frequency is higher in users within the ages of range from (15-29) years according to (P value and odd ratio).

Table 3. Odd ratio

Odds Ratios	Class(preferring local tourism)			
Variable	Very Likely	Likely	Somewhat Like	Unlike
Age	1.083	1.0856	1.0684	1.0827
Gender	1.0461	1.0404	1.0605	1.0333
Work	1.0078	1	1.0018	0.9957
Educational achievement	1.0088	1.0051	1.0151	1.0085
Environment	1.2263	1.2864	1.0023	0.9788
Adventure	1.0267	1.0196	0.9965	0.9849
Religious	1.0245	1.0195	1.0194	1.0138
Culture	0.0066	0.9956	0.995	0.9944
Accuracy with Tied ranks coding using ordinal logistic regression				84%
Accuracy with simple coding using ordinal logistic regression				78 %

The hypothesis of this test was done using SPSS (Statistical Package for the Social Sciences), a p-value [27] helps to determine the significance of our results. The small p-value (typically≤0.05) indicates strong evidence against the null hypothesis. All null hypotheses are rejected. The results of p-value are given in Table 4.Shaded cells in the table represent the factors that must be considered to have effects on frequency of travel in Iraq. For example the p value of work >.05 so the hypothesis of work is rejected ,even retired people, housewife, and student preferred culture places but they cannot be determined because the p value is not trivial at level of significance.

Table 4. P-Value

P value(sig)			
Independent variables	Sig	Independent variables	Sig
[age=45_50]	0.997	[work=house wife]	0.100
[age=>50]	0.996	[work=free work]	0.100
[age=35_39]	0.255	[work=student]	0.631
[age=40_44]	0.484	[work=employee]	0.023
[age=30_34]	0.155	[Educational=PhD]	0.033
[age=20_24]	0.039	[Educational=high school]	0.067
[age=25_29]	0.039	[Educational= [M.Sc.]	0.004
[age=15_19]	0.042	[Educational=B.Sc.]	0.023
[gender=female]	0.041	[environment=yes]	0.004
[gender=male]	0.039	[Adventure=yes]	0.0159
[work=retired]	0.997	[religious=yes]	0.043
[work= un_emp]	0.997	[culture=yes]	0.806

8. Conclusions

This research presented the importance of managing data and extracting knowledge from collected data over social networks. The collected information and data included tourists and touristic places in order to know tourists' behaviors and their orientations. This was done based on personal information and FoT in order to support Iraqi local tourism, which in turn can significantly support economy of Iraq. The proposed method adopted TripAdvisor social media as a source of information .the data was collected using (scrapy and API) tools to extract the beneficial factors in prediction of tourists' preference of touristic places inside Iraq. The proposed work used tied rank and ordinal regression showed that Using of Tied ranks of numerical data, added more accuracy to our ordinal regression. In addition, the personal information of age, gender, work, educational, and types of preferred tourism places had an effect on the FoT of tourists. Furthermore, the control factors assumed a vital part in choosing the recurrence of movement.

References

[1] World Tourism Organization and Organization of American States. Tourism and the Sustainable Development Goals–Good Practices in the Americas, UNWTO, Madrid. 2018. DOI: https://doi.org/10.18111/9789284419685.

[2] R Živković, J Gajić, I Brdar. *The Impact of Social Media on Tourism*. Sinteza 2014-Impact of the Internet on Business Activities in Serbia and Worldwide, Belgrade, Singidunum University, Serbia, 2014: 758-761. DOI: 10.15308/sinteza-2014-758-76.

[3] Zeng, Benxiang, Rolf Gerritsen. What do we know about social media in tourism? A review. *Tourism Management Perspectives.* 2014; 10: 27-36.

[4] Kotler P, Kartajaya H, Setiawan I. Marketing 3.0. John Wiley & Sons, Inc, New Jersey. 2010: 8-9.

[5] Hasan, Samiul, Satish V Ukkusuri. Location contexts of user check-ins to model urban geo life-style patterns. *PloS one 10.5 (2015): e0124819*. 2015.

[6] Park Sangwon, et al. Travel personae of American pleasure travelers: a network analysis. *Journal of Travel & Tourism Marketing.* 2010; 27(8): 797-811.

[7] Frew E, Shaw R. The relationship between personality, gender, and tourism behavior. *Journal of Tourism Management.* 1999; 20: 193–202.

[8] Scott A, Girish Prayag, Miguel Moital. Consumer behavior in tourism: Concepts, influences and opportunities. *Current Issues in Tourism.* 2014: 17(10): 872-909.

[9] Shao H, Zhang Y, Li W. Extraction and analysis of city's tourism districts based on social media data. *Computers, Environment and Urban Systems*. 2017; 65(2017): 66–78. doi:10.1016/j.compenvurbsys.2017.04.010

[10] Jingjing Lia, Lizhi Xubc, Ling Tanga, Shouyang Wangde, Ling Lia. Big data in tourism research: A literature review. *Tourism Management.* 2018; 68: 301-323.

[11] Y Mei, H Jinchuan, Z Meifeng. *Mechanism Analysis and Modelling of Passengers' Travel Choice.* Sixth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2014: 406-409.

[12] Park D, Kim WG, Choi S. Application of social media analytics in tourism crisis communication. *Current Issues in Tourism.* 2018: 1–15. doi:10.1080/13683500.2018.1504900.

[13] Lee W. A Study on the tourism information supply service through social media-based information and communication technology in local government: focused on podcast. *Korea Journal of Tourism and Hospitality Research.* 2012; 26(6): 55–77.

[14] Park JH, Lee C, Yoo C, Nam Y. An analysis of the utilization of Facebook by local Korean governments for tourism development and the network of smart tourism ecosystem. *International Journal of Information Management.* 2016; 36(6): 1320–1327. doi:10.1016/j.ijinfomgt.2016.05.027.

[15] Etaati, Leila, David Sundaram. Adaptive tourist recommendation system: conceptual frameworks and implementations. *Vietnam Journal of Computer Science*. 2015; 2(2): 95-107.

[16] Boley BB, Jordan EJ, Kline C, Knollenberg W. Social return and intent to travel. *Tourism Management.* 2018; 64: 119–128. doi:10.1016/j.tourman.2017.08.008.

[17] Rashidi TH, Abbasi A, Maghrebi M, Hasan S, Waller TS. Exploring the capacity of social media data for modelling travel behaviour: *Opportunities and challenges. Transportation Research Part C: Emerging Technologies*. 2017; 75: 197–211. doi: 10.1016/j.trc.2016.12.008.

[18] A Fast and Powerful Scraping and Web Crawling framework. Online: https://scrapy.org. 2018.

[19] Developers of TripAdvisor. Content API. 2019. Online: https://developer-tripadvisor.com/content-api/2019/.

[20] Bahrami M, Singhal M, Zhuang Z. A cloud-based web crawler architecture. In2015 18th International Conference on Intelligence in Next Generation Networks 2015 Feb 17 (pp. 216-223). IEEE. L Atzori, A Iera, G Morabito.The internet of things: A survey. Computer Networks. 2010; 54(15): 2787-2805.

[21] Social Media Statistics & Facts. 2018. Online: https://www.statista.com/topics/1164/social-networks.

[22] Gniazdowski Z, Grabowski M. Numerical Coding of Nominal Data. *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki.* 2015; 9(12): 53-61.

[23] Richard Williams. Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology.* 2016; 40(1): 7-20.

[24] S Unni, S Gunasekar, D Gupta. *Impact of self concepts & person concepts on the travel frequency of an Indian tourist*. 2016 International Conference on Communication and Signal Processing (ICCSP). 2016: 2047-2050.

[25] Persoskie, Alexander, Rebecca A Ferrer. A Most Odd Ratio: Interpreting and Describing Odds Ratios. *American Journal of Preventive Medicine.* 2017; 52(2): 224-228.

[26] Amrhein, Valentin, Fränzi Korner-Nievergelt, Tobias Roth. The earth is flat (p>0.05): significance thresholds and the crisis of unreplicable research. *PeerJ*. 2017; 5: e3544.