Tourism Management 81 (2020) 104132

Contents lists available at ScienceDirect

Tourism Management

journal homepage: http://www.elsevier.com/locate/tourman

Persistent homology in tourism: Unlocking the possibilities

Woon Kian Chong [a], Simon Rudkin [b,*]

[a] International Business School Suzhou, Xi'an Jiaotong-Liverpool University, 8 Chongwen Road, Suzhou, Jiangsu, 215123, China
[b] School of Management, Bay Campus, Skewen, Swansea, SA8 1EN, United Kingdom

ARTICLE INFO

Keywords:
Tourism
Persistent homology
Expenditure
Clustering
Marketing

ABSTRACT

Marketing a destination is costly so efficiency of promotion expenditure is critical; identifying improved techniques to achieve this will be of great value. Clustering techniques have sought to identify target markets but are widely criticised for the biases they induce. Persistent Homology identifies key tourist groupings with similar behaviours without the prejudice of the functional forms inherent in most regression models. It further produces more focused, and therefore easily promoted to, markets. Consequently Persistent Homology can highlight obtainable promotion opportunities that otherwise would be missed. This paper provides an example of its application to identify the highest, and lowest, spenders amongst tourists visiting the United Kingdom. We further provide an intuitive theoretical background highlighting the inherent value of the methodology for tourism research. Potential for impact in applications to other aspects of tourism practice is also great as we signpost therein.

Effective promotions of destinations, attractions and services requires careful consideration of the target markets, likely effectiveness and the costs of gathering data thereon. Within tourism the encouragement of spending by inbound tourists is a key research area (Aguiló, Rossello, & Vila, 2017; Brida, Cortes-Jimenez, & Pulina, 2016; Brida & Scuderi, 2013; Dogru, Sirakaya-Turk, & Crouch, 2017). Increasingly complex methodologies are proposed to capture expenditure drivers (Olya & Mehran, 2017; Shahzad, Shahbaz, Ferrer, & Kumar, 2017). However these techniques necessarily impose relationships onto the data which detract from the ability to understand diversities in tourist behaviour. Persistent homology (henceforth PH) is a topographical data analysis tool that identifies groupings with similar behaviours. PH does so with no regard for artificial constructs. Stemming from mathematics, applications outside of the sciences remain limited. Likewise, practitioners are yet to understand the true potential of PH. In this paper we describe the methodology, provide an example of its use and discuss the opportunities afforded for tourism research and practice.

We make three critical contributions, enabling a discussion of a new set of questions that currently employed systems struggle to address. Methodologically we enhance the literature with a clear, context relevant, exposition of PH and its suitability for tourism research. Ours is the first paper to suggest such. Secondly, we demonstrate the existence of tourist groupings within the United Kingdom International Passenger Survey (IPS) data which other models would not have found. Specifically we identify amongst low-spenders clusters of long-distance travellers who are middle-aged, travelling alone or as couples and usually flying out of the UK; all characteristics for which high spending would be predicted. Finally we discuss the potential for the methodology to be extended and propose a research agenda for the field.

Section 1 provides an introduction to the related literature on expenditure promotion and cluster approaches used within tourism, serving to contextualise the illustration that follows in Section 3. Introducing PH, Section 2 offers a look at applications in other fields that may inspire use in tourism as well as providing an intuitive bivariate exposition of the methodology. Section 3.3 uses data from the IPS to look at the topography of characteristics of the highest, and lowest, 10% of spenders that visit the UK on holiday; through this we demonstrate clusters missed by currently adopted techniques. Reflecting upon the example, Section 4 reinforces the value of PH for expenditure characterisation and targeting of marketing. Building further, Section 5 explores wider scope for future research, theoretical extension and management implementation. Section 6 then concludes.

1. Tourism expenditure and promotion

1.1. Tourism decisions and marketing

Key for value realisation is that tourists do make the choice to travel.

* Corresponding author.
E-mail address: s.t.rudkin@swansea.ac.uk (S. Rudkin).

https://doi.org/10.1016/j.tourman.2020.104132
Received 15 September 2018; Received in revised form 4 February 2020; Accepted 25 April 2020
Available online 16 May 2020

Many contributors to tourist utility are outside the control of the host authorities; such as the weather discussed in Goh (2012), Falk and Lin (2018) and others. For those which may be influenced there is clear research value in knowing to whom promotions should be targeted; sustainability of the holiday on offer is a major concern for Russia for example (Andrades & Dimanche, 2017). Within the destination marketing literature all of these segmentations receive consideration, unified by the aim of ensuring that the first choice of tourists is to travel. In our case this means travel to the UK and hence to contribute to the GDP of the UK through expenditure and the associated multipliers (Frechtling & Horvath, 1999; Yang, Fik, & Altschuler, 2018). Data on those who do not travel is intuitively much harder to obtain relative to surveys of those who have already decided to travel. The IPS illustrated here falls into the latter category. We illustrate the value of PH applied to those who do travel, but the method does not preclude using choice to travel data should such data be available to the user.

Destination marketing is a naturally complex proposition involving the cooperation of agencies at the national, regional and local level; collaboration between governmental, commercial, and non-governmental not-for-profit groups is critical (Gretzel, Fesenmaier, Formica, & OLeary, 2006; Khalilzadeh & Wang, 2018; Line & Wang, 2017; McCamley & Gilmore, 2018; Pike & Page, 2014). A large mass of literature focuses on this very relationship, reviewing the challenge that nations face in trying to "identify and satisfy consumer needs more effectively than their rivals".[1] That PH identifies groupings commonly missed by conventional methods has benefit against the competition goal. However, in its identification of novel groups, PH may generate conflict because current understanding is grounded in the output of these conventional methods. Buy-in to the validity of PH against traditional models is needed to resolve this. Here we make the case that PH enhances understanding and is a trustworthy source of theoretical development.

To profile tourists logistic regression is commonly adopted (Alegre & Pou, 2004; Molera & Albaladejo, 2007) with extensions into discrete-continuous modelling (Ferrer-Rosell, Coenders, & Martínez-Garcia, 2016; Rashidi & Koo, 2016; Wu, Zhang, & Fujiwara, 2013). In each case the aim is to identify the probability of a tourist behaving in a particular way given their characteristics. From each model associated coefficients emerge that inform the user about what to expect when encountering a tourist with specific characteristics. Common results emerging include an increased probability of being a high spender associated with coming from developed nations, longer stays and the respondent being older. In this sense logistic regressions are consistent with other methodologies. We underline the value of our PH analysis by providing a simple logistic regression with the same variable set in supplementary appendix B.

A second well studied approach to profiling involves the use of cluster analysis, Ernst and Dolnicar (2018) present a valuable review of papers using the technique and the challenges their methodologies present. A long-standing trend in tourism research is to first use factor analysis to determine which variables go forward for clustering; this removes information and is widely criticised (Dolnicar, 2003; Dolnicar & Grün, 2008). An advantage of PH is that such a two stage approach is not required as the data cloud can embed any number of variables; limitations remain computer power driven. Fifteen years on from the Dolnicar (2003) critique, and despite the availability of improved computer processing, principal components analysis continues to be employed to reduce the number of factors on which clustering is based. Recent examples of the use of principal component analysis include Brida, Scuderi, and Seijas (2014) and Ramires, Brandão, and Sousa (2018). Robinson, Getz, and Dolnicar (2018) is one of the few papers adopting a full data approach. As well as determining the variables upon which to cluster, studies must also evaluate the distance between points to identify neighbours and explore the method by which the clustering will be conducted. Most common of the approaches are the Euclidean measures and k-means analysis after Hartigan and Wong (1979) .[2] After introducing PH a comparison with the discussed methods is given in Section 2.3.

1.2. Tourism expenditure modelling

Tourism expenditure is an important contributor to GDP globally. Tourism further brings strong multiplier effects for regional and national economies; recent works confirming this include Carmignani and Moyle (2019); Ferrari, Jimenez, and Secondi (2018); Li, Jin, and Shi (2018b) and Yang et al. (2018). Promoting tourism therefore has natural benefits for the host economy. Leveraging that benefit requires consideration of the drivers of expenditure, recognising the potential for targeted destination marketing, and acknowledging the alternative means through which segmentation may be achieved. We demonstrate here how PH may achieve this, building on the understanding gained in contemporary expenditure research.

Length of stay is a primary determinant within the literature, but it is one frought with endogeneity concerns for empirical research. A plethora of work studies the determinants of the length of stay as a function of the same variables used to explain expenditure, recent examples being Alén, Nicolau, Losada, and Domínguez (2014) and Wang, Fong, Law, and Fang (2018b). Aguiló et al. (2017) is amongst a growing set embracing endogeneity through joint determination models, allowing the explanatory variable set to simultaneously determine stay duration and expenditure. PH abstracts from causality as it draws directly from the data cloud to produce associations. Consequently, we employ length of stay free from endogeneity concerns. Further advantaging PH is the ability to cope with categorical data. Stay duration is measured in days and is dominated by just a few values.[3] Boto-García, Baños-Pino, and Álvarez (2019) and others use count data models to address the discontinuity. Like the logit model we use for our comparator results, these count models have their own coefficient interpretation challenges.

Our other covariates are also commonly accepted influence factors for expenditure (Thrane, 2014). Age impacts typically as a proxy for disposable income, those in the latter years of working life being the higher spenders. The Bernini and Cracolici (2015) study of Italian household expenditure finds those over 50 to have highest spending levels, consistent with Thrane (2014) earlier result. For Chinese overseas travellers, Lin, Mao, and Song (2015) identify the over 40s as being the highest spenders; expenditure levels do not fall back to the marginally lower level of 30 year olds until after 70. Gender is always modelled as a key demographic with males usually found to spend more; this is not necessarily a causality and hence PH has the chance to demonstrate such. Air travel is also associated with higher spending as it proxies income as either an indicator of long-haul travel or because it is the more expensive way to reach the UK from neighbouring European countries. Group size impacts significantly on behaviour but the translation of this to expenditure is more vague. One hypothesis for a negative effect is that larger groups can save money by buying together. However, total expenditure would be expected to be higher by sheer weight of numbers. The balance of these counterveiling effects then appears as the expenditure figure reported by respondents. Nationality captures culture, is a loose proxy for income, and is commonly employed in the literature without firm motivation. As a simple targeting variable we use it here free from the assertion that it directly impacts upon expenditure. We

[1] This is to paraphrase Drucker (1984) seminal work on what marketing management should do.

[2] To underline the benefits of PH against these clustering techniques we offer an analysis of our example data as a supplementary appendix.

[3] For example stays are more likely to be 7 days or 14 days, as these are whole numbers of weeks, than they are to stay for 9 or 11 days say.

direct interested readers to the reviews by Thrane (2014) and Brida and Scuderi (2013) for a full review of variables. Any of the many recent expenditure studies exposited below also add to the review.

Expenditure studies employ increasingly complex methodologies to understand what drives visitors to spend more. Subsequent works have sought to build upon the simple linear models identified in Brida and Scuderi (2013) and Thrane (2014) reviews, by considering the distribution of expenditure (Chen & Chang, 2012; Marrocu, Paci, & Zara, 2015; Almeida & Garrod, 2017; Rudkin and Sharma, 2017). These papers use either traditional quantile regression (Koenker & Bassett, 1978), or in the case of Rudkin and Sharma, 2017 an unconditional quantile regression (UQR) approach after Fortin, Lemieux, and Firpo (2009) .[4] Distributional approaches[5] like these satiate the desire to understand expenditure amongst the highest and lowest spenders, but still rely on the inherent causality. Irrespective of complexity there is still a basic desire to know how a feature of the individual dictates their spending behaviour. Such modelling neglects the information that is held in the relationships between different explanatory variables and how combinations thereof drive the outcome.

Quantile regression has relevance to our proposed analysis as we seek to cluster amongst the highest and lowest expenditure groups. Quantile regression creates an impression that those characteristics with the strongest positive coefficients at higher quantiles are the ones which marketing should encourage. Likewise quantile regression suggests that those with the largest negative coefficients amongst lower spenders should be reviewed more carefully to encourage greater expenditure. Alternatively it could be argued that because the expenditure reduction effect is largest amongst those with already low spending, visits from individuals with that characteristic should therefore be discouraged in order to promote visits from others who would be predicted to be higher spenders. Interpreting quantile regression in this way has application in business, but for tourist boards it is readily argued that the product (here visits to the UK) is non-rival (Rigall-I-Torrent & Fluvià, 2011). Given a non-rival product, all visits should be encouraged and promotions should instead be aimed to extract more from all who do visit. Distributional regression should thus be applied where the interest lies in the relationship between variables as it identifies important causality that ordinary least squares (OLS) misses. However, like all regression techniques, distributional regression approaches, like quantile regression, struggle to effectively identify potential target markets outside the fitted relationships. This is a challenge PH meets.

2. An introduction to persistent homology

Persistent homology is introduced through a twin track perspective. First we look at how the methodology has been employed in the existing literature. We further provide an intuitive bivariate example of the technique.

2.1. Persistent homology in the literature

In the past decades, scholars have found PH an effective computational tool for characterisation of clustering features in large data sets. Edelsbrunner, Letscher, and Zomorodian (2000); Zomorodian and Carlsson (2005) and Zomorodian and Carlsson (2008) are among many examples which do so. In comparison with other traditional clustering techniques, PH has a unique ability to compute data at different topological space and spatial resolutions (Carlsson, 2009; Weinberger, 2011; Zomorodian & Carlsson, 2005). Phrased alternatively, this is the ability to map data clouds on different scales and to cluster using varying distance criteria. Relative to the current tourism literature this means that there is no need to standardise variables before performing the analysis; Ernst and Dolnicar (2018) notes how commonplace this remains. Over a wide range of geometric data analyses, more topological features are detected and assessed for their likelihood as true reflections of features of the underlying space. Features appearing only for certain filtrations may simply be artefacts of noise, sampling processes or parameter specification lacking requisite robustness to nest meaningful conclusions (Carlsson, 2009; Weinberger, 2011; Zomorodian & Carlsson, 2005). PH constructs real-world data in a topological way identifying persistence across multiple scales and providing robust new insights by so doing (Otter, Porter, Tillmann, Grindrod, & Harrington, 2017). Existing homology literature depicts rich applications of PH in neurosciences (Liang & Wang, 2017), sensor networks (De Silva & Ghrist, 2007), signal processing (Huang & Ribeiro, 2018), medical imaging (Chung, Hanson, Ye, Davidson, & Pollak, 2015), and biomedical engineering (Xia & Wei, 2014). Searches of the major academic databases reveal the rapid growth of the methodology. Equally such searches confirm that no attempt has been made to unlock the potential of PH for the benefit of travel research.

Pereira and de Mello (2015) developed a framework for clustering time-series and spatial data based on topological properties, which can precisely identify qualitative aspects of a data set currently missed by traditional clustering techniques. The authors argue that computational topology may be of great help to detect similarities in recurrent behaviour for time series data sets and spatial structures in spatial data sets. Obvious extensions to patterns in visitor numbers, and seasonality in demand for travel related products, immediately suggest themselves as applications in tourism management. An ability to exploit patterns beyond those captured by the standard modelling techniques makes a large contribution to understanding relationships with internet social behaviours, the effects of climate change, pollution and other less commonly studied items. Results on the former (Li, Chen, Wang, & Ming, 2018c, 2017b; Siliverstovs & Wochner, 2018), and latter (Li, Song, & Li, 2017a; Wang, Fang, & Law, 2018a), highlight the state of research for these topics and the variables currently candidates for the homology.

Xia and Wei (2014) explore the utility of PH for protein structure characterisation, protein flexibility quantification, and protein folding stability forecasting. PH is employed to extract the molecular topological fingerprint (as a unique grouping function) in order to understand the relationship within the protein structure. The topological structures are further used as multivariate features, which allows a better tracing of geometric origins contributing to topological invariants. Xia, Zhao, and Wei (2015) extend this work and propose a multi-resolution approach to calculate large molecular datasets which would be missed by normal point cloud methodologies. These analyses conclude that the PH-based model delivers an excellent quantitative prediction with simplification of complex data. Intuitively such gains map into large tourism data sets. These biological applications then map into tracing tourists expenditure patterns and identify tourism participation behaviour. Relationships within protein structures can be likened to the interactions between the type of tourist that will travel (Glover & Prideaux, 2009), where tourists travel from (Asero, Gozzo, & Tomaselli, 2016), the length of stay (Wang et al., 2018b), how much they spend (Bernini & Cracolici, 2015) and so on. Entrepreneurially, in this study, we embed computational geometry and topology in clustering technique that will enhance the visibility of information within the tourism data.

Liang and Wang (2017) adopt PH methodologies to examine the relationship between brain structure and the function of network neuroscience adopting PH methodologies. They formulate the problem as a topological clustering of structure-function connectivity via matrix functions, and find a stable solution, exploiting a regularisation procedure to cope with large matrices. Liang and Wang (2017) revealed an innovative measure of network similarity based on PH for assessing the

[4] See Borah and Basu (2013) for an intuitive discussion of the differences between the Fortin et al. (2009) and Koenker and Bassett (1978) techniques and why these differences are of interest to practitioners.

[5] We can use the terms “distributional”and “quantile”regression interchangably as both QR and UQR are distributional techniques.

quality of the network mapping. So doing enables a detailed comparison of network topological changes across all possible thresholds, rather than just at a single, arbitrary threshold that may not be optimal. This approach supported by Carstens and Horadam (2013) and Petri et al. (2014)'s applications of PH in social network structure demonstrates that PH is able to distinguish between networks that are considered similar in form by other approaches. Social network analysis is growing in tourism (Luo & Zhong, 2015; Wang, Li, & Lai, 2018c); evaluating the networks through PH, drawing on the approach in neuroscience, has further potential. Obvious networks within tourism include travel flows, transportation networks and the social networks upon which reviews are posted.

The literature above employs PH to construct topology-based multiscale models. The fundamental idea is to use geometry theory, with clustering abilities, to develop topology-function relationships/networks. However the applications of PH to management fields is limited, and there is no literature about the use of this method as a clustering tool in tourism management. Derived from the literature, we use PH to analyse and visualise the tourism data and identify groupings with similar tourists expenditure patterns to understand the spending behaviours of the highest and lowest spenders in the UK. PH has strong potential wherever there are multiple ordinal characteristic variables; such data is highly prevalent in management. Relevance for PH over other approaches begins where there is either no known relationship between the characteristic and the output, or the relationship is known to be too complex for standard modelling. However the true strength for PH is harnessed where the links between input characteristics and the outcome vary according to the values of other characteristics. A simple tourism example would be if increasing the length of stay was linked to higher expenditure for females who travel on business and leave the UK by air, but lower expenditure for males who also travel on business but leave the UK by non-air means. Whilst these are the conditions under which PH offers most, PH speaks to all datasets if only to confirm that there is no more useful insight beyond that from a simple model. In confirming understanding, or delivering new insight, PH delivers deep value. We return to potential applications in Section 5.

2.2. A bivariate illustration of persistent homology

The key feature of PH is that it can connect and measure topological data to find "holes"with clustering features within finite metric spaces (Carlsson, 2009). Data of the type collected in many empirical tourism applications fits this requirement as stay duration, travel party size, nationality, travel mode etc. are all finite in their range of possible values. Understanding of those holes is best derived diagrammatically beginning with a simple plot in two-dimensional space. For an algebraic discussion of the methodology and its theoretical underpinnings please see Carlsson (2009), Edelsbrunner and Morozov (2013) and others.

PH begins with an exploration of the data points available to the analyst, viewing these as realisations of a bigger surface upon which actualities reside. In statistical analysis this sees points as draws from the underlying population distributions to which the data refer, constructing this inference being the very aim of surveys like the IPS. Mapping, common in tourism research, is essentially a process for joining information points to construct the geographical topology in defined space; data topography does likewise on dimensions defined by the variables of interest. PH identifies those regions for which insight is lacking, viewing them as features of interest which warrant further understanding.

We illustrate this here using a two dimensional example of length of stay and travel group size. Both can only take integer values, but do display variation within the data to make a scatter plot of interest. Ex-ante there is no particular reason to expect shapes within this data, although the low instances of long stay and large groups mean the combination of high stay durations for many co-travellers will not appear often. Fig. 1 illustrates this clearly with many points being clustered into the bottom left hand, short-stay small group, section. Meaningful inference is drawn in those regions where there are data points, but where sparsity exists this becomes harder such that holes may appear in our knowledge.

Filtration focused on three sample points only for clarity. All other points are connected as demonstrated by panel (a). Filtration circles are for illustration only and do not represent specific values of the filtration parameter ε. A colour version of this figure is available in the online edition.

For these data points we have complete information, we can identify respondents and know their full characteristics, but we do not know anything away from the exact measurement. To understand the broader shape of the data a process of filtration is employed, the level of filtration being the radius around each data point over which a search for neighbouring data points is undertaken. Once two circles join we consider that a link or "edge" has been born. In Fig. 2 this is represented by panel (a), where, zooming in on a region of Fig. 1, gaps are clearly seen. Three data points are singled out for further analysis as they are unconnected by the initial filtration; we represent these points as squares. Most points do have overlap in their circles and so are connected by edges, and though these black squares connect into other points they do not connect with each other. Hence, we view these black squares as being features in their own right and, in PH terminology, say that there are three features in dimension 0. Formally the betti number in dimension 0 for the reduced dataset is 3, $.b_0 = 3$[6]

As we increase the filtration level we arrive at panel (b) and the circles of the right hand two points meet. Here a new feature, an edge, forms, but the individual data points that are now connected no longer exist as features in their own right. Now we have $b_0 = 2$. Increasing the filtration further sees the remaining square connected and reduces us to a single feature. The resulting feature, a triangle shaded light green, is interesting because within it lies an area, shaded dark green, upon which the filtration does not provide information; the light green triangle is a "hole". Note here that the hole is defined as the area enclosed by the edges that connect the points on its outside, the existence of the uncovered area, dark green shading in this case, within that shape simply

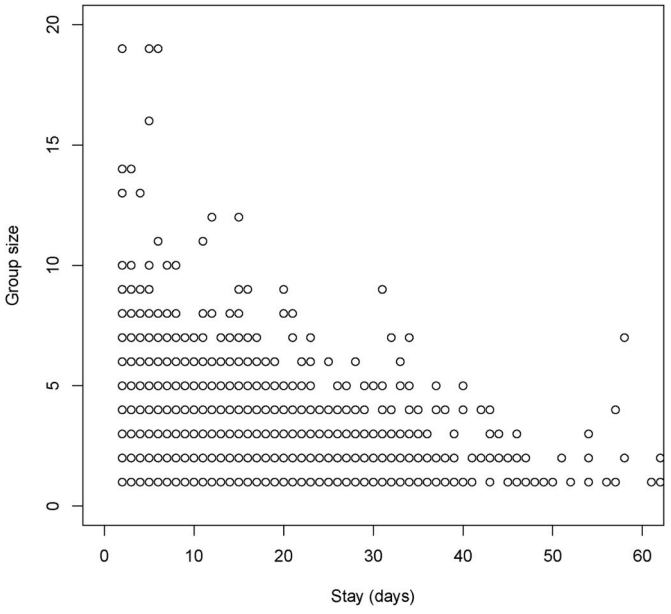


Fig. 1. Bivariate plot of stay duration and group size.

[6] As in classic statistics we are using *b* here to denote an estimate of the true population betti number β.

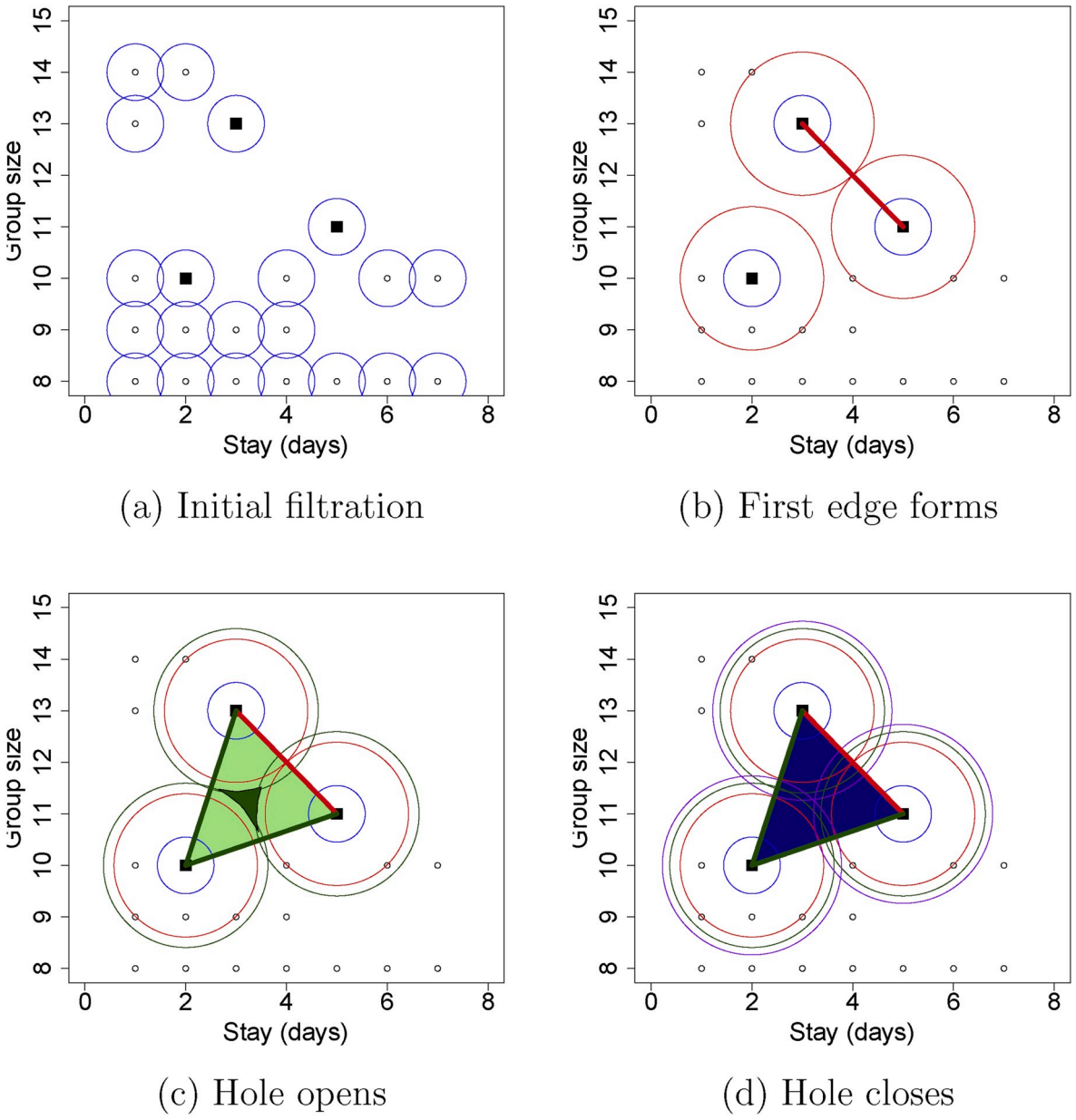
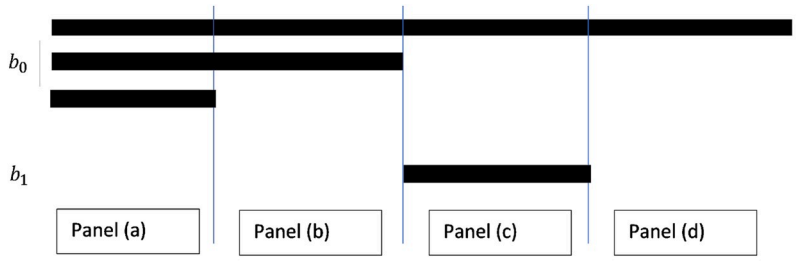


Fig. 2. Persistent Homology with two tourism variables.



Notes: Horizontal axis represents filtration used in construction of the PH cover. Panel labels correspond to those of Figure 2. In the bivariate case this would be the radius of the circles drawn around points. In multiple dimensions ϵ is the radius of the ball. This figure is not drawn to scale. Such plots would readily extend to higher dimensions of the topology by adding betti numbers for dimensions 2 (b_2), 3 (b_3), etc.

Fig. 3. Barcode plot for bivariate tourism example. Notes: Horizontal axis represents filtration used in construction of the PH cover. Panel labels correspond to those of Fig. 2. In the bivariate case this would be the radius of the circles drawn around points. In multiple dimensions ε is the radius of the ball. This figure is not drawn to scale. Such plots would readily extend to higher dimensions of the topology by adding betti numbers for dimensions 2 (b_2), 3 (b_3), etc.

confirms that it is a hole. Holes are examples of features in dimension 1, and so our first dimension 1 feature exists and $b_1 = 1$. Continuing to increase the filtration will see the circles completely overlap, the hole disappears; Fig. 2 panel (d) shows the result. We shade the triangle dark blue to show that it has been filled. Following the disappearance of the hole $b_1 = 0$ once more. Through this all there remains a single feature in dimension 0, the set of three connected points, and $b_0 = 1$.

This process is represented using a barcode plot, for which Fig. 3 is the example derived from the case in Fig. 2. Moving from left to right we have the level of filtration increasing so the left hand side has three features in dimension 0 and none in dimension 1. As the edges form so the bars in b_0 die, before the formation of the hole creates the dimension 1 feature. This then dies as the hole closes and we are left with a single feature in dimension 0 on the right hand side. With two variables these patterns are trivial to see, even with the added complexity of studying the full scatter diagram in Fig. 1. However, tourism surveys offer multiple variables and, unlike mapping, do not define a neat two dimensional space upon which to base studies. Hence, the irrelevance of variable numbers to PH is a valuable facet. These barcodes are representations of the radius of the balls around each point that create, or destroy, the feature and are hence of identical form for any number of variables. In the literature an advantage of PH is the robustness to noise within the data, by removing those features that have a too short life it is possible to segregate data noise from meaningful content. Barcode plots identify such features neatly, but their existence is readily confirmed from the difference between birth and death filtrations.

From an analytical perspective the hole that was in the dimension 1 remains a space in which there is no data to form connection irrespective of the number of variables; interpreting that hole is application dependent. An analyst may evaluate the contents of the hole, or consider those points surrounding its edge; these points organically identify themselves as being in the dataset with the interest then being in what unites these observations.

Apparent from the bar chart of Fig. 3 is that there are important questions to answer on the choice of filtration level. If the level is too low then no circles will overlap and there will be no features in dimension 1; such a position would be obviously disadvantageous to any topological data analysis. Should the filtration be set too high then all holes will fill and only a very limited number of features will be identified. Between the extremes the number of holes will vary as small holes form and fill and bigger holes are created. In our empirical example we use a larger filtration to maintain a manageable number of holes; the actual choice of ε is left to the individual investigator.

Having explored filtration, and established firmly the variables over which the homology is performed, it is possible to produce a first set of results. These results will provide a clear picture of the way in which the sample is representing the population. We could go further and compare the topography of datasets, for example between types of traveller or nationality, through bar code comparisons. This is an emerging field in PH research, but here we limit ourselves to the workings of the technique and simple applications as a route into its use. We return to implications and potential uses of PH after our illustration.

2.3. *Persistent homology or clustering?*

As presented PH offers a means to identify groupings within a dataset that have potential interest to the data analyst. For this paper that means sets of individuals who have similar expenditure when in the UK. Table 1 offers a summary of four approaches to clustering that are commonly used in the tourism literature; these are exposited for the IPS data in the supplementary material. For each of the suggested alternatives to PH examples of their use are given, the approach they use to generate clustering discussed, insights from their employ on the IPS data drawn, and evaluation of their advantages and disadvantages summarised. First of the four considered approaches, k-means was used in Dolnicar (2003) and is noted by Dolnicar and Grün (2008) to be present in almost 90% of all studies in the Tourism literature. Trimmed clustering is less common in tourism but is used here to show how any questions of extreme values might be considered. Hierarchical clustering is more common, recent examples including Dimitrovski and Todorović (2015) and Veisten, Haukeland, Baardsen, Degnes-Ødemark, and Grue (2015). Chirieleison and Scrucca (2017) employ model based clustering in destination marketing, representing one of very few studies to have adopted the approach to date. Despite clear advantages of other approaches, k-means remains the overwhelming majority of work in the field. Contrasting the four with PH reveals, outwith the potential for newer approaches to outperform k-means, that the critical differences that give PH such potential as a means of understanding data in the tourism sector and beyond stand against all. Any feature found by PH must be representative of the data precisely because all features come directly from the data being studied without any assumption on the grouping construction. Further, the notion that similar individuals may behave similarly, but that small changes in characteristics may alter expenditure behaviour, is intuitive but clustering algorithms do not embed this in the way that the locality of a hole in PH does.

Rodriguez et al. (2019) concludes that there is no one clustering methodology that may be optimal for all situations and purposes. Table 1 builds upon the insights from that paper to highlight the decision process through which PH is recommended here. In the IPS dataset there are a number of characteristics which are derived from likert-scale questions that do not have the normal distribution needed for model based clustering. Whilst there is merit in the ease with which k-means

Table 1
Comparison of clustering methods and persistent homology.

	Clustering Methdodologies:				Persistent Homology
	K-means	Trimmed	Hierarchical	Model Based	
Reference	Hartigan and Wong (1979)	Fritz, Garcia-Escudero, and Mayo-Iscar (2012)	Johnson (1967)	Fraley (1998)	Carlsson (2009)
Cluster Identification Process	Partition observations around centroids to minimise total distance between points and centroids	Remove proportion of outliers suggested by researcher then cluster using partitioning	Linkage approach which cuts characteristic tree diagram at appropriate cluster number	Model based such that selected clusters are optimised for user specified model fit	Organically identified from dataset using filtration of point cloud
Number of Clusters	Algorithms developed, including elbow plot	Set by user	Set by user	Set by user	Determined from dataset via radius of filtration
Size of Clusters	Set by data density	Set by data density	Set by data density	More evenly distributed	Smaller focused units
Intra-cluster Characteristic Variability	Large	Large	Large but with focus on some variables	Focuses to reduce variation in many variables	High similarity enforced by filtration
Distance Matrix	Full needed to optimise centroids	Full for trimmed data	Builds upward	Full	Localised by radius of filtration
Advantages	Well understood Efficient with large data	Removes outliers common in tourism data Efficient with large data	Easily motivated from top-down and bottom-up	Suitable with incomplete datasets Best with normal distributed characteristics	Fully representative as from data shape Recognises outcome changes in small regions of the characteristic space
Disadvantages	Not useful if clusters very different sizes	Lose information from trimmed	Cut point arbitrary	May fail to find small clusters	Not all data points are on edge of holes

Notes: Comparisons of methods for general datasets. For fuller evaluation of methods consult the respective references. Size of clusters denotes number of observations within each group. Characteristics refers to spread of axis variables within clusters. Distance matrix refers to the matrix of distances between data points within the overall characteristic space. Advantages are based on authors experience and Rodriguez et al. (2019). Clustering for the UK IPS data is provided in the supplementary appendix.

and the related trimmed clustering approaches are operationalised, they are challenged by an inability to recognise small clusters, and the loss of information from the trimmed observations. In marketing and destination management it is helpful to have all of the information retained and to obtain the maximum possible insight into even the smallest of clusters. Targeted marketing in-particular demands the focus of small clusters, which k-means cannot easily deliver. Finally, whilst hierarchical clustering is intuitive when considered as observations joining together, or a grand coalition of all observations splitting, the arbitrary nature of the cutting process begs questions. Hierarchical and model-based clustering are advantaged by not needing the full distance matrix between points in the data, like PH, but both have critical assumptions that are not readily overcome. Hierarchical clustering uses distances between points and is therefore most similar to PH intuitively. However where a given filtration creates connections between data points for PH, the same cut for hierarchical clustering would produce a large number of small clusters. Selecting a subset of these for marketing would then be the closest to studying the PH holes.[7] Any hole identified by PH would touch on a number of these small clusters, identifying a means through which spending encouragement, or destination advertising, could be honed for the full set of touched clusters. Increasing the filtration in the hierarchical clustering would reduce the number of clusters problem, but would lose the detail that the PH holes offer. PH thus stands as the best way to identify smaller clusters and the true hierarchy of the data, and to do so without any requirement for model assumption.

Cluster comparisons thus reveal that no one method can always produce the best outcome. Indeed, left to automatic parameterisation, most methods underperform when faced with artificial datasets of known properties (Rodriguez et al., 2019). PH, as developed here, is different because it is not using any parameters beyond the filtration radius, and from there barcodes ensure that the researcher is aware of the robustness of the results that emerge. PH targets identifying groups within the overall data cloud that can be readily understood by researchers, rather than the clumsy global nature of clustering approaches.

3. Persistent homology for tourism expenditure

PH delivers many potential benefits for analysis informing strategy across tourism. Here we take data from the UK IPS to demonstrate the ways in which PH can be used to cluster amongst high, and low, spending visitors. For the purposes of this paper deciles are used to segregate the market, given the large dataset having a strong sense of the top spenders is important. Using a coarser grading would increase computation times and make it less clear exactly how low (high) spending a cluster within the larger sub-sample would be. We introduce the dataset in Section 3.1 and discuss the challenges of visualising patterns across multiple dimensions. After briefly discussing the implementation of PH in Section 3.2 we present the resulting homology in Section 3.3. Returning to the representation challenge we illustrate our clusters within the data space, and present example clusters to show how PH often brings out surprising results. From our example these groups of interest are exposited in Section 3.5.

3.1. Data

Our specific interest is in the flows of passengers from overseas who are leaving the UK after a holiday. This could be by sea or air, but we do restrict it to holiday makers and so exclude all business travellers and those who are visiting friends or family in the UK. Within the survey there are a large number of questions but many do not apply to many of the respondents and as such we limit our focus to the key variables mentioned in Table 2.

Whenever seeking to identify groups of interest a trade off exists between including detail and computational attainability; consequently we categorise the length of stay into eight groups rather than focusing on the number of days individually. These groupings represent very short stays of less than three days, short stays, approximately one week, a week and a half, two weeks, three weeks, four weeks and a month or longer. As can be seen in Table 2 there are large numbers in these lower categories, including 34 whose spending ranks them in the top 10% of

[7] At an extreme in practice it may be that the marketing agency is able to produce promotions aimed at every cluster, but this is to most intents and purposes impractical.

Table 2
Summary statistics.

Variable	Levels	Top 10	Bottom 10	All
Length of Stay	1 or 2 Days	34	932	3027
	Between 3 and 5 days	243	377	6319
	Between 6 and 8 days	307	105	2597
	Between 9 and 12 days	284	39	1203
	Between 13 and 18 days	372	38	1131
	Between 19 and 26 days	103	19	310
	Between 27 and 29 days	56	7	107
	More than 30 days	94	17	209
Age:	Under16 years (individual)	30	105	632
	Under 16 years (party)	38	218	1186
	17–24 years (individual)	134	267	2029
	17–24 years (party)	14	44	278
	25–34 years	286	275	3102
	35–44 years	286	260	2921
	45–54 years	332	253	2917
	55–64 years	251	130	1668
	65 years +	165	131	1077
Gender:	Female	696	797	7572
	Male	797	737	7331
Transport:	Sea	128	822	3707
	Air	1365	712	11196
Group size:	1 person	605	592	4993
	2 people	601	482	5891
	3 people	114	145	1550
	4 people	91	204	1677
	5 people	44	56	448
	6 people	26	26	202
	7 or more people	12	29	142
Nationality:	United States (1)	339	125	1894
	Germany (2)	52	131	1593
	France (3)	12	303	1415
	Italy (4)	15	57	842
	Netherlands (5)	19	148	823
	Spain (6)	15	44	704
	Ireland (7)	8	112	604
	United Kingdom (8)	63	112	539
	Australia (9)	139	38	529
	Canada (10)	68	30	390
	India (16)	44	37	221
	China (18)	66	6	195

Notes: For brevity only the 10 nations providing the highest total visitor numbers, and other select major nationalities, are reported. Ranks in terms of overall visitor numbers to the United Kingdom are reported in parentheses after the nation. Figures are based on the IPS sample of tourists and vary from total numbers when other visiting purposes are considered. Individual in the younger age groups refers to travelling alone, whilst party indicates that the respondent was part of an organised trip with others. All calculations on ONS (2017).

overall spenders. Categories within age are dictated by the IPS, whilst group sizes are cut at 7 or more based on the overall numbers being less than 1%. Most respondents in the survey were aged between 25 and 54, with group sizes typically just being one or two. A slight dominance of females is seen for the overall sample, and in the bottom 10% of spenders but at the top end there are noticeably more males in the group. Befitting of the long distances many respondents travel, air transportation is the most common departure mode; the bottom 10% shows very close figures for air and sea however.

Fig. 4 demonstrates the challenges when trying to form relationships amongst variables in the IPS via inspection. Whilst it is clear that the top 10% of spenders typically stay longer than the lower 10%, differences in group size and age are harder to pick up from scatterplots. There is also no strong association between any of the three variables plotted which makes for a large cloud of points. Adding in further dimensions for nationality, gender and mode of transport will further complicate this picture. PH enables us to identify patterns from within these clouds of data drawing topographical inference when even simplified data is not mapped by the mind.

3.2. Implementation

PH may be computed using a series of code libraries implementable in regularly employed statistical packages that are commonly utilised in the tourism management literature. We use the Javaplex analysis library of Tausz, Vejdemo-Johansson, and Adams (2014), which has a convenient interface with MATLAB (MATLAB Optimization Toolbox, 2017). Many others use the PHAT library of Bauer, Kerber, Reininghaus, and Wagner (2017) for Python (van Rossum, 1995), or the TDA package (Fasy, Kim, Lecci, & Maria, 2018) for R (R Core Team, 2018). A strong advantage of Python and R are their open source nature and the fact that the respective PHAT and TDA packages are free to use. Those who are interested are directed to Otter et al. (2017) and the respective software articles mentioned herein. Implementation in this paper uses an Intel Core i5 with 8 GB of RAM, and a 2-core 2.50 GHz CPU. The machine runs a 64-bit ×64 based processor operating system. This can be considered typical of the computers used by most researchers and practitioners.

Whichever software package is employed, the dataset must be provided in a way which permits construction of a point cloud. In our case we have a serial number to identify individual respondents but we can not provide that to the homology as it is not a real variable associated with the individual; this must be removed prior to running the code. That results depend solely on the maximum filtration level, ε_{max} adds a robustness over techniques that require multiple selections to be made by the researcher. Indeed the computation of the homology runs from zero up to the ε_{max} requiring solely that the upper limit has meaning for the barcodes. In the Tausz et al. (2014) implementation cluster membership is only available for the maximum ε and hence the decision here is what facilitates the reported clusters; we can easily run the homology again to identify clusters at an alternative ε level.

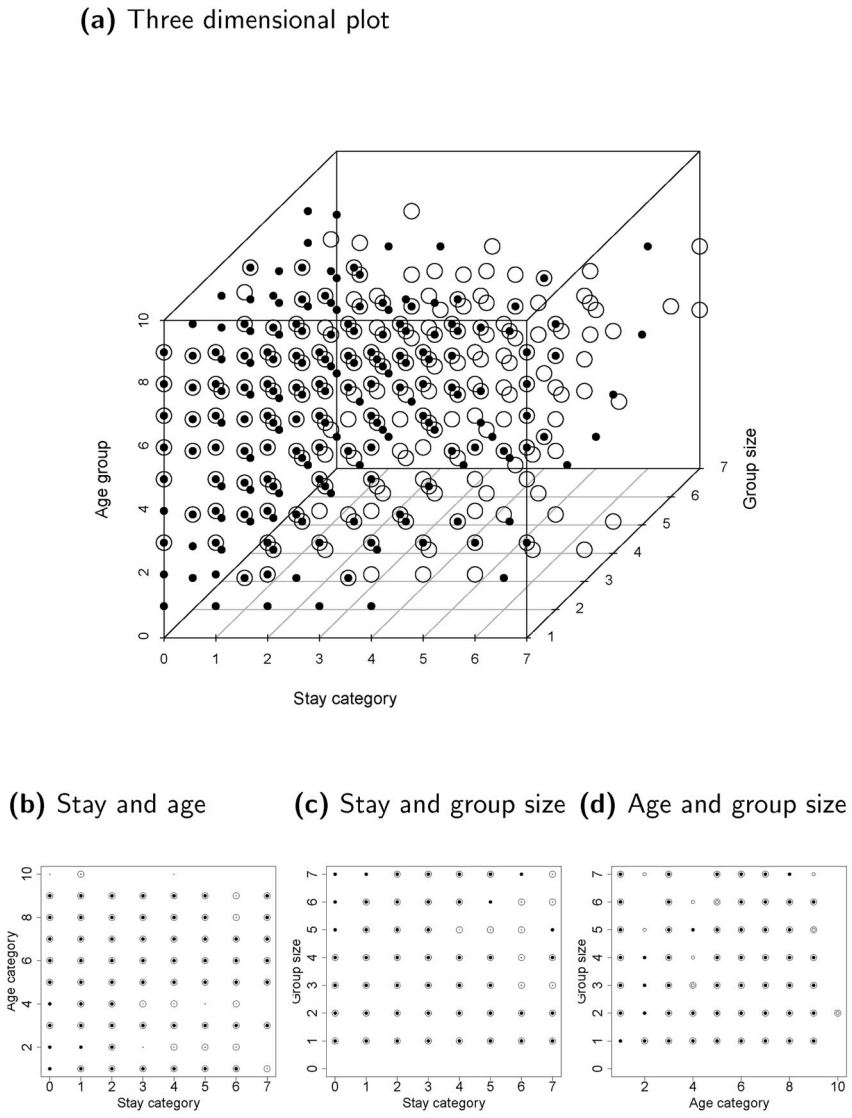
4. Results

Section 2.2 discussed a two variable example, showing with Fig. 3 how barcode plots can be used to illustrate features in the data as filtration is increased. Fig. 5 provides the bar codes from our multivariate homology on the IPS data. We split the sample into top 10% and bottom 10% and so we have 4 barcodes in total. As points connect and edges form, so the number of features in dimension 0 falls, we see the death of a large number as the filtration passes 0.5 as the integer value nature of many variables makes neighbours a fixed 1.0 apart. We use a filtration of 1.5, which allows observations 3 apart to be connected; this has the effect of controlling the number of clusters which appear but also allows wider differentials amongst the observations within the clusters. Optimising this is left to the investigator as there is a natural trade off between cluster numbers and cluster usability.

Diagrams like these inform solely of the persistence of the homology; to extract value we need to explore the characteristics of the holes that the dimension 1 bars represent. Table 3 provides a full set of summary statistics for the 1.5 filtration level.[8] In every case the average value for the cluster is stated and the standard deviation, reported in parentheses, gives indication of the within cluster variability. Also listed are the number of nationalities within each cluster, the betti number at which the cluster is born, and finally the number of members whose points in the data cloud join to form the hole. We see 18 holes exist at $\varepsilon = 1.5$ in the top 10% and 15 in the bottom 10%.

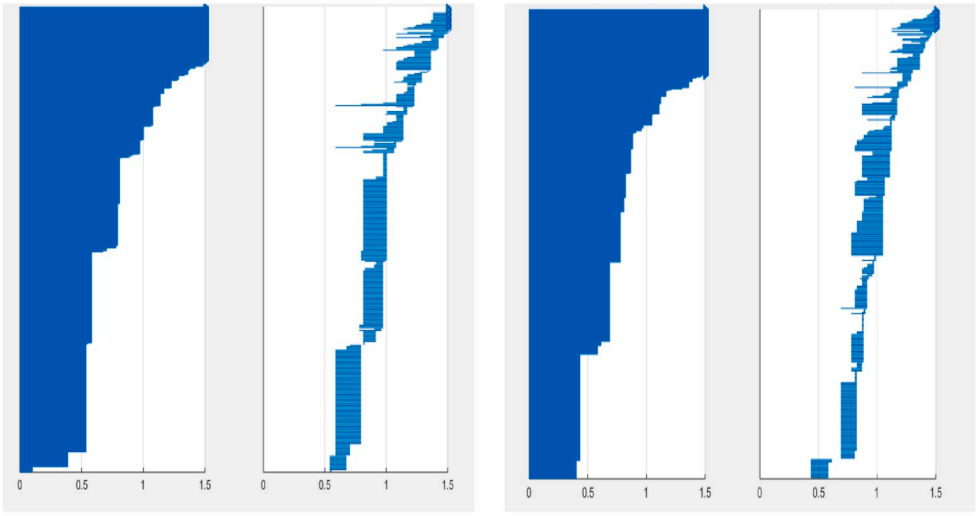
Figures report the average value for each variable within the cluster, the figure in parentheses being the standard deviation for that variable and cluster. For nationality we report the number of nationalities represented within the cluster. Betti reports the betti number at which the particular feature appears. *N* reports the number of observations within the cluster. Stay and age are categorical and as described in Table 2.

[8] We provide full details of the cluster members in the supplementary material section C.



Notes: Unfilled circles denote datapoints in the top 10%, solid circles the bottom 10% of spenders. Top 10% drawn larger to illustrate overlap. Three dimensional plot generated using Ligges and Mächler (2003). All data from ONS (2017)

Fig. 4. Visualising patterns within the IPS data. Notes: Unfilled circles denote datapoints in the top 10%, solid circles the bottom 10% of spenders. Top 10% drawn larger to illustrate overlap. Three dimensional plot generated using Ligges and Mächler (2003). All data from ONS (2017).



Notes: Horizontal axis shows the degree of filtration ϵ and ranges from 0 to the 1.5 used in the homology that follows. All data from ONS (2017).

Fig. 5. Persistent homology barcodes. Notes: Horizontal axis shows the degree of filtration ε and ranges from 0 to the 1.5 used in the homology that follows. All data from ONS (2017).

Viewing the 18 top 10% and 15 bottom 10% clusters in this way shows that they are all organised around a single departure transportation mode, and single gender of respondent. From a marketing perspective this means that targets are identified as "female aeroplane users"for example. Across the three variables illustrated in Fig. 4 we can see more variation. There are some clusters of solo-travellers that show no variation in group size.[9] Likewise the bottom 10% contains four clusters of travellers who do not stay overnight. There are no single age category groupings. For nationality it is meaningless to compute an average for a cluster, but we do see several clusters that are only single nationality; one such cluster, number 5 in the bottom 10%, is an example in Section 3.5.

As we saw in the bar codes, much of the homology present at $\varepsilon = 1.5$ forms above $\varepsilon = 1$; all of the minimum betti numbers are in this range. Sizes range from just 5 to 25 members, with most of the higher spending groups having more than 10. In this paper 1% of the total sample would be around 16 members and so total cluster membership is approximately 10% of those in each sample. For a large dataset we would expect the majority of respondents to behave similarly and for there to be strong connections between observations. That the remaining 90% do not form holes through the homology is certainly not unexpected.

4.1. Bivariate and trivariate visualisations

Visualising these clusters remains challenging from their multi-dimensional nature. Fig. 6 shows that within our three multiple category variables there is evidence of grouping for most identified features but that there are fewer differences between the top and bottom sets than might be expected. Only in the length of stay dimension does any segregation between the highest and lowest spenders appear, the outline circles of the top 10% being to the right of the dots that represent the lowest 10%. In the lower part of Fig. 6 the bivariate analysis shows both age and group size have significant overlap.

A useful way to understand the clusters is to plot a histogram of the average values from Table 3. Fig. 7 does this, demonstrating clearly the overlap in both age categories and group sizes. Amongst the highest spenders group sizes are slightly smaller, with averages closer to 1 being much more common in the top 10%. Maybe less surprising is that there is higher spending from older respondents; the distribution in the top row is notably more centered on 45–54 year olds compared to the broader spectrum in the lower row. These age results are consistent with Bernini and Cracolici (2015) analysis of Italian households and broadly with Lin et al. (2015) work on China.

4.2. Example clusters

For brevity the full set of clusters is reported in Section C of the supplementary material and here we present just two from each of the two expenditure levels. Through these we can see just how PH can connect across characteristics and identify sets that would not be picked up by standard regression techniques. In discussing the groupings we refer to logistic regressions for the probability of an individual with a given characteristic being in the spending group analysed; commonplace in the literature on tourism market segmentation a full coverage of these regressions is available in the supplementary material.

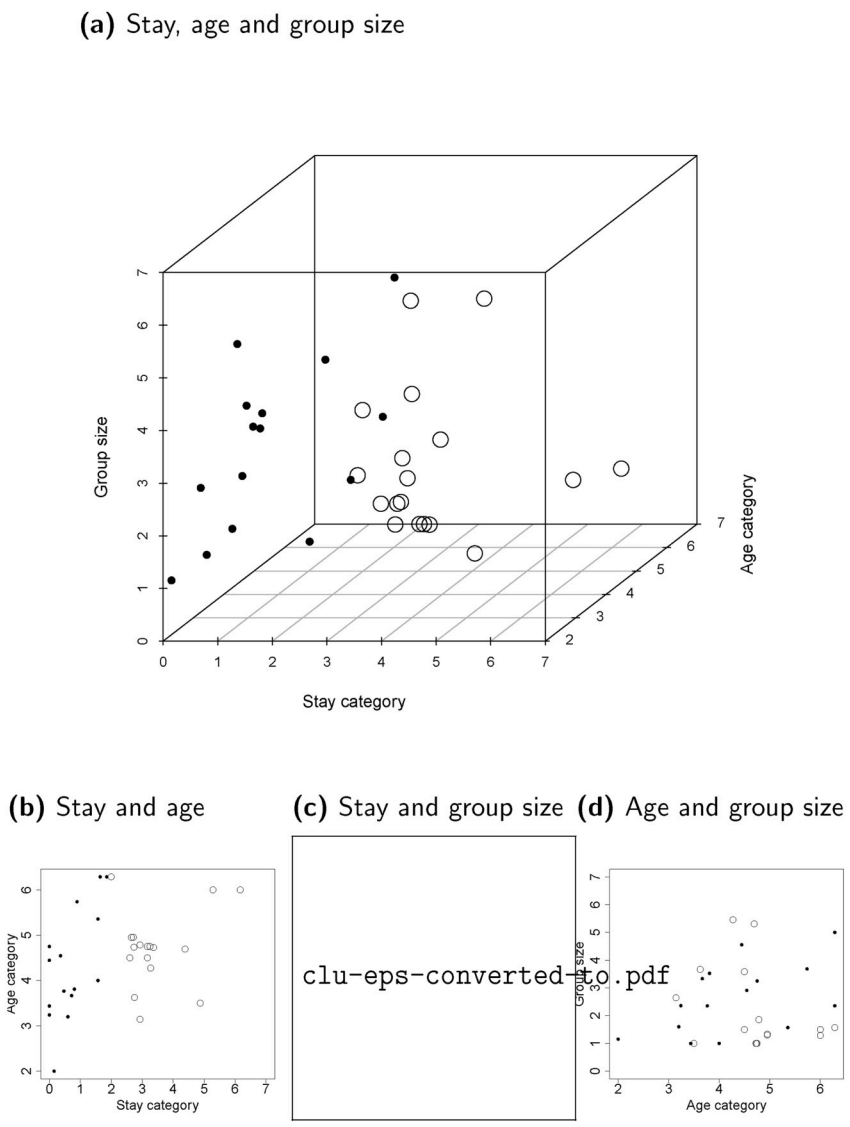
Two clusters selected from the bottom 10% are shown in Table 4; each has very different characteristics. Cluster 4 comprises American females who depart by air and travel either alone or in small groups, whilst cluster 1 is made up of European males who depart by sea and travel in larger groups. Both groups show large variations in stay duration, and also feature ages from 35 upwards; variability like this is to be expected from PH.

Logistic regressions tell us that departure by sea is more likely to produce lower 10%, whilst there is no significant gender effect. Older respondents are, ceteris paribus, less likely to be in the bottom 10% so this is surprising, as is the inclusion of longer stayers. We do find that larger groups are more likely to produce low spenders, so the second cluster fits that observation, but cluster 4 is at odds with the regression results. Likewise, although France has a significant association with lower spending, being from the USA or Germany would normally only be linked to increased probabilities of being in the top 10% and have no significant relationship with the bottom 10%. Consequently much of

[9] Clusters 2,5,7,13 and 16 from the top 10% and clusters 5 and 8 from the bottom 10% are such.

Table 3
Summary of individual clusters.

Group	Cluster	Stay	Age	Male	Flow	Persons	Nations	Betti	N
Top 10%	1	2.929	3.143	0	1	2.643	3	1.245	14
		(0.616)	(1.791)	(0)	(0)	(1.082)			
	2	2.733	4.733	1	1	1	4	1.38	15
		(1.335)	(0.799)	(0)	(0)	(0)			
	3	6.167	6	1	1	1.5	2	1.485	6
		(0.983)	(0.632)	(0)	(0)	(0.548)			
	4	4.385	4.692	0	1	5.308	2	1.365	13
		(1.758)	(0.63)	(0)	(0)	(1.251)			
	5	4.875	3.5	0	1	1	1	1.17	8
		(1.356)	(1.195)	(0)	(0)	(0)			
	6	2.714	4.952	1	1	1.333	4	1.38	21
		(1.146)	(0.865)	(0)	(0)	(0.483)			
	7	3.364	4.727	1	1	1	4	1.485	11
		(1.12)	(0.905)	(0)	(0)	(0)			
	8	3.273	4.273	0	1	5.455	3	1.365	11
		(0.786)	(1.191)	(0)	(0)	(1.128)			
	9	3.167	4.5	1	1	3.583	1	1.485	12
		(0.937)	(1.446)	(0)	(0)	(1.443)			
	10	2.6	4.5	0	1	1.5	4	1.38	10
		(0.516)	(0.527)	(0)	(0)	(0.527)			
	11	2.455	6.727	0	1	1.727	1	1.23	11
		(1.508)	(1.348)	(0)	(0)	(0.467)			
	12	2.65	4.95	1	1	1.3	4	1.38	20
		(1.182)	(0.887)	(0)	(0)	(0.47)			
	13	3.167	4.75	1	1	1	4	1.38	12
		(1.115)	(0.866)	(0)	(0)	(0)			
	14	2.75	3.625	1	1	3.667	3	1.365	24
		(1.032)	(2.223)	(0)	(0)	(1.949)			
	15	5.286	6	0	1	1.286	1	1.29	7
		(1.38)	(1.155)	(0)	(0)	(0.488)			
	16	3.25	4.75	1	1	1	4	1.38	12
		(1.055)	(0.866)	(0)	(0)	(0)			
	17	2	6.286	0	0	1.571	1	1.14	7
		(0.816)	(1.38)	(0)	(0)	(0.535)			
	18	2.929	4.786	0	1	1.857	4	1.485	14
		(0.73)	(0.699)	(0)	(0)	(0.77)			
Bottom 10%	1	1.857	6.286	1	0	5	2	1.2	7
		(0.9)	(0.951)	(0)	(0)	(0.816)			
	2	0.895	5.737	1	0	3.684	2	1.395	19
		(1.15)	(0.933)	(0)	(0)	(1.529)			
	3	0.154	2	0	0	1.154	4	1.29	13
		(0.376)	(1.414)	(0)	(0)	(0.376)			
	4	1.643	6.286	0	1	2.357	1	1.125	14
		(1.55)	(0.914)	(0)	(0)	(1.277)			
	5	0	3.438	0	0	1	6	1.26	16
		(0)	(1.931)	(0)	(0)	(0)			
	6	0.81	3.81	1	1	3.524	4	1.2	21
		(0.68)	(1.914)	(0)	(0)	(1.167)			
	7	0.364	4.545	1	1	2.909	2	1.455	11
		(0.505)	(1.44)	(0)	(0)	(0.944)			
	8	1.571	4	0	1	1	3	1.44	7
		(0.787)	(1.528)	(0)	(0)	(0)			
	9	0	4.444	0	0	4.556	1	1.47	9
		(0)	(1.333)	(0)	(0)	(1.59)			
	10	0.6	3.2	0	0	1.6	1	1.305	5
		(0.548)	(1.643)	(0)	(0)	(0.548)			
	11	0.471	3.765	0	1	2.353	2	1.2	17
		(0.514)	(1.562)	(0)	(0)	(1.455)			
	12	0	3.24	0	0	2.36	5	1.17	25
		(0)	(1.943)	(0)	(0)	(1.578)			
	13	0.722	3.667	1	1	3.333	3	1.425	18
		(0.669)	(2.029)	(0)	(0)	(1.138)			
	14	0	4.75	1	1	3.25	1	1.47	8
		(0)	(1.581)	(0)	(0)	(1.282)			
	15	1.571	5.357	0	1	1.571	3	1.425	14
		(0.756)	(2.24)	(0)	(0)	(0.756)			



Notes: Smaller solid circles are used to denote the bottom 10%. Larger outline circles provide the locations of the means from the top 10% clusters. Three dimensional plot generated using Ligges and Mächler (2003). All data from ONS (2017)

Fig. 6. Average values within clusters. Notes: Smaller solid circles are used to denote the bottom 10%. Larger outline circles provide the locations of the means from the top 10% clusters. Three dimensional plot generated using Ligges and Mächler (2003). All data from ONS (2017).

what PH identifies runs counter to the results of market segmentation regressions.

Our top 10% spending clusters reported in Table 5 also feature both genders and both departure mode options. All are either solo travellers or have at most one companion, there is great consistency within these characteristics. Cluster 2 covers the age range 25–54, whilst 17 comprises slighty older members being 35 upwards. The first cluster is made from visitors from India and the Middle East, whilst the second comes from the United States of America. Strong similarities between cluster 17 here and cluster 4 of the bottom 10% demonstrate that clusters from opposite ends of the expenditure range can in fact share many similar characteristics. Specifically, both clusters 17 and 4 capture a range of stay durations, comprise female respondents who are over 35 years of age primarily travelling in small groups, and are USA nationals. There are differences in that cluster 4 has some lower group sizes and is comprised of air travellers. At this point caution is urged against concluding the differences are what explains the expenditure. In such cases any conclusion would only apply to the specific data being studied in the clusters. Once again there is variation amongst stay duration and age, with clustering focusing on group size in particular.

Unsurprisingly these nationalities are associated with higher spending by the logistic regressions, as are stays between 3 and 18 days. For age there is little significant prediction for being in the top 10% and hence little can be said of the consistency of the clusters in this dimension. Males are significantly more likely to be in the higher spending group, as are air departures, so to this extent cluster 17 goes against what regression models would predict. All other group sizes are less likely to be in the top 10% relative to solo-travellers and hence there is greater consistency with these two clusters.

5. Summary

Using data from the 2016 UK IPS we have shown how PH can be used to identify clusters within expenditure, highlighting how many of these clusters differ significantly from the suggestions that standard market segmentation methods would produce. Visualising patterns in data to

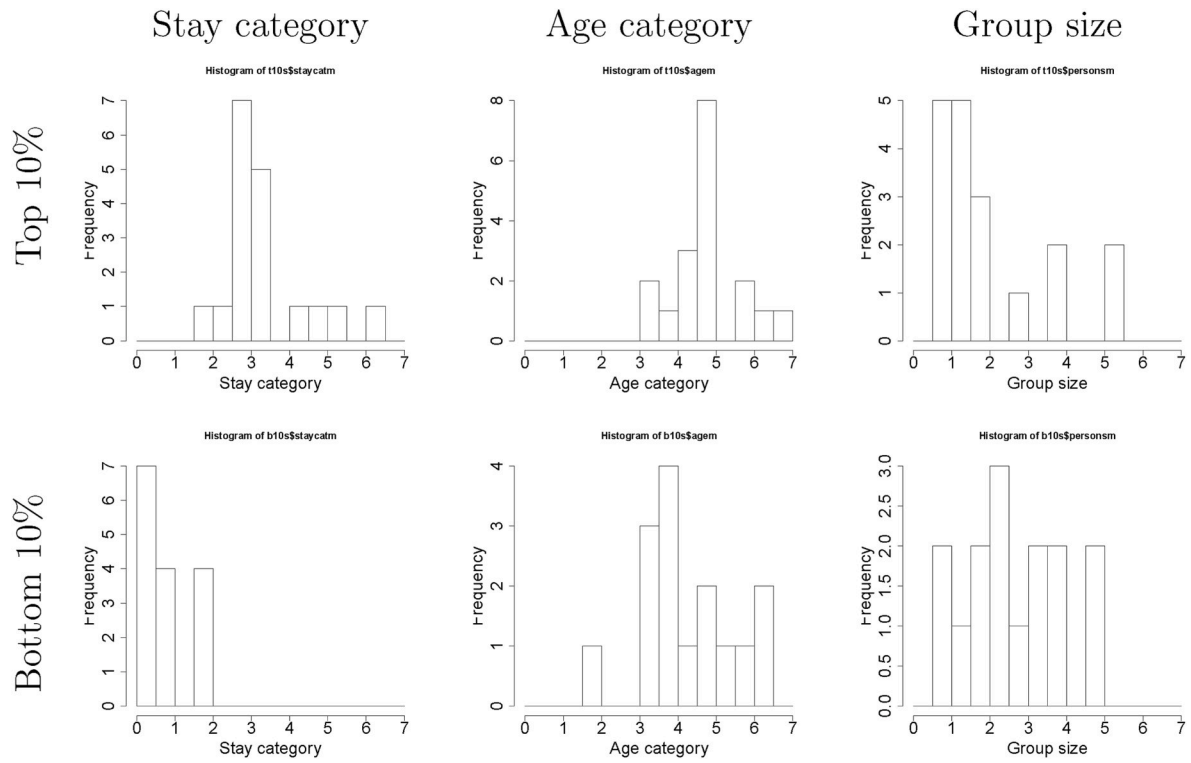


Fig. 7. Histograms of cluster means.

Table 4
Example clusters from lowest 10%.

Cluster	Stay	Age	Male	Air Departure	People	Nationality	Min ε
4	13–18 days	65 years +	0	1	2	USA	1.125
	13–18 days	45–54 years	0	1	3	USA	
	3–5 days	55–64 years	0	1	2	USA	
	3–5 days	45–54 years	0	1	3	USA	
	3–5 days	35–44 years	0	1	1	USA	
	3–5 days	45–54 years	0	1	1	USA	
	3–5 days	45–54 years	0	1	4	USA	
	6–8 days	45–54 years	0	1	4	USA	
	6–8 days	35–44 years	0	1	1	USA	
	9–12 days	45–54 years	0	1	4	USA	
	9–12 days	55–64 years	0	1	4	USA	
	9–12 days	45–54 years	0	1	1	USA	
	9–12 days	65 years +	0	1	2	USA	
	6–8 days	45–54 years	0	1	1	USA	
1	3–5 days	55–64 years	1	0	5	France	1.2
	3–5 days	45–54 years	1	0	5	France	
	9–12 days	45–54 years	1	0	5	Germany	
	6–8 days	35–44 years	1	0	6+	Germany	
	9–12 days	45–54 years	1	0	4	Germany	
	6–8 days	65 years +	1	0	4	Germany	
	3–5 days	45–54 years	1	0	6+	Germany	

uncover these results is challenging and hence the use of PH uncovers information that would be otherwise hidden. As we subsequently discuss, using this information effectively is a second challenge. As presented the value of PH lies in its ability to deliver deeper understanding of the full data cloud.

6. Analysis and opportunities

Our example on inbound tourist expenditure in the United Kingdom highlights the ability of PH to identify characteristic groupings which would be missed by models such as OLS or quantile regression, typically used in the literature. Amongst lower spending visitors we highlighted groups of Americans who were in the UK for short stays and who were either travelling alone or in pairs. Many of these characteristics would be associated with higher spending and therefore the groups would not be the targets of promotions that sought to sell the value of visits to the UK. Instead it would be expected these groups would receive information that talked of luxury products and the things higher spenders would normally be buying. Amongst the top spenders we saw fewer clear patterns; groupings were clustered more by characteristics such as group size and gender rather than easily marketable factors like nationality. Our PH analysis thus urges more careful consideration of promotional activity, suggesting a medium through which that may be better targeted.

PH is computationally expensive; the construction of simplices rapidly eats memory and necessitates larger amounts of processor time.

Table 5
Example clusters from highest spending 10%.

Cluster	Stay	Age	Male	Air Departure	People	Nationality	Min ε
2	9–12 days	25–34 years	1	1	1	Kuwait	1.38
	6–8 days	25–34 years	1	1	1	Kuwait	
	3–5 days	35–44 years	1	1	1	India	
	3–5 days	35–44 years	1	1	1	Saudi	
	13–18 days	35–44 years	1	1	1	India	
	19–26 days	45–54 years	1	1	1	India	
	19–26 days	45–54 years	1	1	1	Saudi	
	13–18 days	45–54 years	1	1	1	Kuwait	
	6–8 days	25–34 years	1	1	1	Saudi	
	3–5 days	25–34 years	1	1	1	Saudi	
	6–8 days	25–34 years	1	1	1	U.A.E.	
	9–12 days	35–44 years	1	1	1	Kuwait	
	9–12 days	25–34 years	1	1	1	India	
	6–8 days	25–34 years	1	1	1	India	
	9–12 days	35–44 years	1	1	1	India	
17	9–12 days	45–54 years	0	0	2	USA	1.14
	6–8 days	35–44 years	0	0	2	USA	
	6–8 days	35–44 years	0	0	1	USA	
	3–5 days	35–44 years	0	0	1	USA	
	3–5 days	55–64 years	0	0	1	USA	
	6–8 days	65 years +	0	0	2	USA	
	9–12 days	65 years +	0	0	2	USA	

Notes: Group sizes are capped at 6 or more so a + is added to the number 6. Saudi is used as a short form of Saudi Arabia. ε is the level of filtration within the persistent homology and hence the minimum ε in the final column reports the birth point of the cluster. Cluster numbering is provided by the software and corresponds to the full set of clusters listed in the supplementary material.

However new algorithms have reduced this to a local optimisation process that needs little more than the construction of the distance matrix required for other clustering approaches. Whilst computation limitations need not apply to larger organisations seeking to get a greater understanding of their data, advances in processing bring PH within reach of all. Research institutions likewise having more supercomputers capable of processing the data mean that the physical sciences academic world has already gained greatly from the scope of PH. This paper is a positioning paper for the approach to recognise the new potential and has therefore been constructed on a limited number of machines, with the explanatory variables kept low in number and broad in categories. It should be readily apparent therefore that fine graining the categories and widening the number of variables is a ready extension of what has been done from our bivariate case to the six explanatory factors discussed in Section 3.3.

Our empirical work is inevitably limited by the dataset; there are a number of other fields which would be of interest, such as income, which are not available in passenger surveys. Filling these gaps is an area for further research but will not diminish the value of PH. Income for example can be expected to be a useful clustering variable. High income leads to high spending, but again there is the potential for PH to identify frugal high income individuals who then appear in the lowest 10%; such a grouping would be an ideal target for promotions to encourage greater spending. We therefore open the challenge to the research community to develop more informative datasets offering a tool that can set free the full informative capabilities thereof.

IPS data is tasked with understanding travel to, and from, the UK meaning many questions which could help understand motivations could be exploited. There is also an opportunity to extend the analysis into regional and accommodation variables which are available for a subset of respondents. We focused on holidaymakers for the example but the work could be extended easily to cover business travellers and those who are visiting friends and relatives. Such purposes are typically assumed to lead to higher and lower spending respectively but again we might expect PH to identify clusters of low spending business travellers or high spending relative visitors.

Using PH has advantages when looking for features within a particular sub-sample. In our example we considered the top 10% and bottom 10% of spenders amongst inbound tourists to the UK. However there are many other interesting questions within tourism that equally suggest subsamples. Within the IPS we could consider the characteristics of air travellers versus sea, budget airlines versus traditional full-service carriers, use of package holidays etc. Outside the IPS data there are applications in understanding demand drivers for specific visitor attractions, hotels, destinations, etc.

7. Implications

7.1. Theoretical implications

Travel behaviour has received much attention in the literature. It is widely acknowledged that demographic profiles are important; they are often utilised to help academics to have a better understanding about tourism expenditure (Wu et al., 2013). The biological characteristics of men and women, together with their age are well documented influences in travel behaviour (Bernini & Cracolici, 2015; Wilkes, 1995; Wong, Fong, & Law, 2016; Wu et al., 2013). These characteristics are especially studied for their effects on decision making, and risk taking propensities. Such studies provide theoretical background for the effects of gender and age on tourism expenditure behaviour. Although there are many studies collating evidence from a broad perspective, existing results are inconsistent, and only offer partial understanding of travel behaviour. Furthermore, existing studies miss important characteristics in cluster analysis because they commonly used methods which are often corrupted by noise. Inductive theory has grown from these noisy empirical results, and hence there is a real risk that the polices are equally flawed. PH robustness to noise addresses these challenges to provide more robust theoretical underpinning.

It is the ability of topographical data analysis techniques like PH to break down the complexity of data structure without being reliant on imposed relationships which offers most to developing theory in tourism management. PH offers a laboratory in which data changes may be trialled and evaluated by perturbations to the cloud. Such a test bed is a valuable tool for testing policy or business strategy constructs. Unique filtration processes ensure that any correspondences identified have the requisite robustness for inductive research. Our example uses simplified expenditure data to show patterns therein that indicate a lack of adherence to established simple relationships.

Rather than the identified clusters per se, the theoretical advancement from our work is in marketing, where identifying individuals and

reshaping the dataset by changing characteristics for particular observations has strong potential. For the lowest 10% we could imagine successfully creating a campaign that removes one of the clusters from the bottom decile of expenditure. Such a move would then cause the homology to identify new individuals as the hole inevitably expands. Simultaneously new observations would become a part of the 10%, potentially further adapting the homology and creating new features. As a loop this process has the ability to raise the threshold of the lowest 10% as well as signposting the effectiveness of marketing. We further regard the trajectory of characteristics targeted through the alterations of the homology as a means to evaluate marketing effectiveness. Such analysis is un-trialled, with PH itself yet to make inroads, but it is introduced here as a theoretical advancement that derives from our work.

Our review of the PH literature also highlighted areas such as time series analysis and understanding social networks where active tourism research fields could readily employ topological data analysis. As the wealth of data on individual behaviour patterns grows, so the opportunity to use that data productively within research and practice presents itself. PH offers scope to understand these new data sources without recourse to imposing formal relationships. Theory often struggles to keep pace with data possibilities, since growing complexities in the underpinning of both, like the need to embed advanced statistics to use distributional regression, detach theorists from the empiricists. Methods like PH that can effectively inform inductive research thus have obvious value. We posit that PH speaks to these topics not only from a clustering perspective, but also as a forecasting tool, an exploratory technique for new data and as a way to usefully identify opportunities to advance theory.

7.2. Managerial implications

Marketing practitioners are desperately seeking for new data science methodologies[10] to embrace the big data era of tourism management research (Alaei, Becken, & Stantic, 2019; Li, Xu, Tang, Wang, & Li, 2018a; Mariani, Baggio, Fuchs, & Höepken, 2018). To the best of our knowledge, this study is the first of its' kind to analyse tourism data in a high dimensional, topographically faithful way. In so doing key details from within the global picture are unearthed. The ideas presented in this study offer a complementary perspective to many existing theories advocated by industry leaders. Flaws in clustering techniques are well studied and are evidenced here again. Through our contrasting of PH with market segmentation techniques we have shown how PH may address many of these flaws in the presently adopted approaches in a clear, robust, and transparent way. Practice thus makes a call to research for something robust to noise, and free from imposed relationships, constructed to optimally identify targets for effective promotion. This study answers that call to deliver marketing solutions that more accurately pertain to the travel behaviours of inbound tourists.

A seeming disregard for small detail in targeted marketing's important clustering phase dictates strategies born of the results are empirically limited in their directional capabilities. PH exploits fundamental geometrical topology to equip managers to make better decisions when allocating their marketing budgets. One of the key features of PH is that it may exploit the geometric data embedded in a dataset to generate a better quality of quantitative modelling. We propose that the topological connectivity of the elements of tourist behaviour can provide a natural segmentation for tourist spending behaviour. To this end, we propose a PH-based model to characterise spending behaviour evolution that can provide excellent marketing value for managers. Fine tuning the categories and expanding the dataset can only extenuate these advantages.

Consumer-centric approaches in tourism management are well studied, with a particular focus on the tourist experience (Mathies, Gudergan, & Wang, 2013; Vogt, 2011). Marketing the right content to the right people at the right time is targeted but too often missed. Hence, the quality of the marketing activities of travel service providers, destinations and policy-makers. Systematic literature assessments on big data in tourism research have been conducted recently by Li et al. (2018a) and Mariani et al. (2018). Both reviews find that there is a need to explore new analytic methods to effectively appreciate tourists' preferences. According to Jang and Ham (2009) and Lin et al. (2015), the biggest obstacle to effective marketing for the tourism and hospitality sector is an inability to obtain actual information about "real" consumers. This is extenuated by a need to know how much these consumers are willing to spend. Overcoming such limitations paves the way for the consumer-centric marketing approaches to function.

Furnished with grounded theories of geometry and computational topology, PH allows us to aggregate and integrate consumer data. PH provides concise summaries of information for managers to draw actual marketing insights. In this study a simple dataset from the UK IPS is used to formulate the expenditure patterns of similar individuals who surround a data hole. It was demonstrated that these tight similarities within the data cloud provided a detailed picture despite the coarseness of the data. Taking this to a better firm focused marketing database can only reinforce the benefits that the IPS provides. Clusters that emerge are clearly identified and provide precisely the concise summary that managers crave and enable the marketing of the right content to the right people at the right time. PH is the first stage analysis the consumer-centric approach demands.

The analysis of tourist behaviour can help practitioners/government implement better marketing activities/policies which can effectively eliminate factors that could hinder tourist activities, and which can encourage tourism expenditure. In addition, the more sophisticated representation of spending groupings can provide a more accurate assessment of the tourist market. This assessment can assist in the design and implementation of marketing strategies for both practitioners and government. Our results exploit to date uncharted, intrinsic topological features about tourist spending patterns helping to formulate better marketing strategies.

Practical value in the application of PH from previous studies across the physical sciences has demonstrated how sensor data, multi-dimensional characteristics and images can be employed by practitioners to understand the true information contained within their datasets. In the current literature most of the successful PH applications have been limited to analysis and barely bring the benefits to the next stage of the business decision making process. This paper initiates the applicability of PH in tourism research, and more widely within marketing research, founding such in previous work employing PH in other areas. Through the work emerges a clustering tool that is more precise, concise and easier to leverage. Hence whilesoever the data may be gathered there are benefits from a topological understanding of data that can be employed; PH offers a lens through which those benefits are identified.

Forecasting abilities give obvious benefit to managers; developing understanding in this direction, PH can deliver clear value. For example work using search engine behaviour often seeks to assign causality to visitor numbers. Using PH managers might identify stable patterns in the data that are seen as precursors to high demand periods. Subsequently, preparations and pricing can be naturally adjusted. A similar tale could be told for periods of downturn and low demand. Social networks and social media are key elements of successful destination marketing; better characterising behaviour in these areas has managerial payoff. Work which draws together extant tourism research and the development of PH in network analysis has direct merit. The suggested extension into systematic data point removal gives scope for managerial analysis as well as theoretical development. Future work will expand on all of these.

[10] See Sousa and Rocha (2019) for a discussion of the wider business response.

8. Summary

Persistent Homology (PH) is an example of topographical analysis which identifies patterns within data that statistical methodologies typically used in tourism would not. By not being limited by pre-conceived ideas of relationships, it offers tremendous potential across the field. This paper is a first step in unlocking those opportunities. Through an example of inbound tourist expenditure in the United Kingdom we have empirically demonstrated these benefits, showing cases where predicted high spenders appear in the lowest 10% and exposit the difficulty of targeting top 10% promotions through nation specific advertising. Extending the discussion to the wider field, we identified a number of future uses. In each case the insight offered by PH could increase the efficiency of promotion spending by advertisers and tourist authorities alike.

Notwithstanding the criticisms of the dataset, which have been raised against all analyses of the IPS dataset, PH will continue to bring the benefits of understanding that we have shown here. Requirements for more advanced computational power are a limitation, but neither practitioners nor academics are expected to be constrained by such in the longer term. Consideration of the methodologies that exploit new capabilities is timely. This paper is necessarily an introduction to the methodology and many further directions exist from both empirical and theoretical standpoints. PH may simply reaffirm that there are no useful cases that do not comply with understood relationships, but such has value in focusing attention. In most cases where it has been applied PH has discovered important data-driven insight that then adds clear value. For tourism, its practitioners, policy-makers and researchers there are many exciting possibilities unlocked, both those discussed and many more.

Author contribution

Dr Woonkian Chong was responsible for the inception of the persistent homology idea, data analysis and the literature review. Dr Simon Rudkin prepared the data, conducted robustness tests, reviewed the statistical literature and constructed the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tourman.2020.104132.

References

Aguiló, E., Rossello, J., & Vila, M. (2017). Length of stay and daily tourist expenditure: A joint analysis. *Tourism Management Perspectives, 21*, 10–17.

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research, 58*(2), 175–191.

Alegre, J., & Pou, L. (2004). Micro-economic determinants of the probability of tourism consumption. *Tourism Economics, 10*(2), 125–144.

Alén, E., Nicolau, J. L., Losada, N., & Domínguez, T. (2014). Determinant factors of senior tourists length of stay. *Annals of Tourism Research, 49*, 19–32.

Almeida, A., & Garrod, B. (2017). Insights from analysing tourist expenditure using quantile regression. *Tourism Economics, 23*(5), 1138–1145.

Andrades, L., & Dimanche, F. (2017). Destination competitiveness and tourism development in Russia: Issues and challenges. *Tourism Management, 62*, 360–376.

Asero, V., Gozzo, S., & Tomaselli, V. (2016). Building tourism networks through tourist mobility. *Journal of Travel Research, 55*(6), 751–763.

Bauer, U., Kerber, M., Reininghaus, J., & Wagner, H. (2017). Phat–persistent homology algorithms toolbox. *Journal of Symbolic Computation, 78*, 76–90.

Bernini, C., & Cracolici, M. F. (2015). Demographic change, tourism expenditure and life cycle behaviour. *Tourism Management, 47*, 191–205.

Borah, B. J., & Basu, A. (2013). Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. *Health Economics, 22*(9), 1052–1070.

Boto-García, D., Baños-Pino, J. F., & Álvarez, A. (2019). Determinants of tourists length of stay: A hurdle count data approach. *Journal of Travel Research, 58*(6), 977–994.

Brida, J. G., Cortes-Jimenez, I., & Pulina, M. (2016). Has the tourism-led growth hypothesis been validated? A literature review. *Current Issues in Tourism, 19*(5), 394–430.

Brida, J. G., & Scuderi, R. (2013). Determinants of tourist expenditure: A review of microeconometric models. *Tourism Management Perspectives, 6*, 28–40.

Brida, J. G., Scuderi, R., & Seijas, M. N. (2014). Segmenting cruise passengers visiting Uruguay: A factor–cluster analysis. *International Journal of Tourism Research, 16*(3), 209–222.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society, 46*(2), 255–308.

Carmignani, F., & Moyle, C.-l. (2019). Tourism and the output gap. *Journal of Travel Research, 58*(4), 608–621.

Carstens, C. J., & Horadam, K. J. (2013). Persistent homology of collaboration networks. *Mathematical Problems in Engineering, 2013*.

Chen, C.-M., & Chang, K.-L. (2012). The influence of travel agents on travel expenditures. *Annals of Tourism Research, 39*(2), 1258–1263.

Chirieleison, C., & Scrucca, L. (2017). Event sustainability and transportation policy: A model-based cluster analysis for a cross-comparison of hallmark events. *Tourism Management Perspectives, 24*, 72–85.

Chung, M. K., Hanson, J. L., Ye, J., Davidson, R. J., & Pollak, S. D. (2015). Persistent homology in sparse regression and its application to brain morphometry. *IEEE Transactions on Medical Imaging, 34*(9), 1928–1939.

De Silva, V., & Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology, 7*(1), 339–358.

Dimitrovski, D., & Todorović, A. (2015). Clustering wellness tourists in spa environment. *Tourism Management Perspectives, 16*, 259–265.

Dogru, T., Sirakaya-Turk, E., & Crouch, G. I. (2017). Remodeling international tourism demand: Old theory and new evidence. *Tourism Management, 60*, 47–55.

Dolnicar, S. (2003). Using cluster analysis for market segmentation-typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research, 11*(2), 5–12.

Dolnicar, S., & Grün, B. (2008). Challenging factor–cluster segmentation. *Journal of Travel Research, 47*(1), 63–71.

Drucker, P. (1984). *Marketing management*. Englewood Cliffs, NJ: Prentice Hall.

Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2000). Topological persistence and simplification. In *Foundations of computer science, 2000. Proceedings. 41st annual symposium on* (pp. 454–463). IEEE.

Edelsbrunner, H., & Morozov, D. (2013). *Persistent homology: Theory and practice*.

Ernst, D., & Dolnicar, S. (2018). How to avoid random market segmentation solutions. *Journal of Travel Research, 57*(1), 69–82.

Falk, M., & Lin, X. (2018). Sensitivity of winter tourism to temperature increases over the last decades. *Economic Modelling, 71*, 174–183.

Fasy, B. T., Kim, J., Lecci, F., & Maria, C. (2018). *Tda: Statistical tools for topological data analysis. R package version 1.6.2*. included GUDHI is authored by Clement Maria, V. R. T., by Dmitriy Morozov, D., by Ulrich Bauer, P., Kerber, M., and Reininghaus, J.

Ferrari, G., Jimenez, J. M., & Secondi, L. (2018). Tourists expenditure in tuscany and its impact on the regional economic system. *Journal of Cleaner Production, 171*, 1437–1446.

Ferrer-Rosell, B., Coenders, G., & Martínez-Garcia, E. (2016). Segmentation by tourist expenditure composition: An approach with compositional data analysis and latent classes. *Tourism Analysis, 21*(6), 589–602.

Fortin, N., Lemieux, T., & Firpo, S. (2009). Unconditional quantile regression. *Econometrica, 77*(3), 953–973.

Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing, 20*(1), 270–281.

Frechtling, D. C., & Horvath, E. (1999). Estimating the multiplier effects of tourism expenditures on a local economy through a regional input-output model. *Journal of Travel Research, 37*(4), 324–332.

Fritz, H., Garcia-Escudero, L. A., & Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software, 47*(12), 1–26.

Glover, P., & Prideaux, B. (2009). Implications of population ageing for the development of tourism products and destinations. *Journal of Vacation Marketing, 15*(1), 25–37.

Goh, C. (2012). Exploring impact of climate on tourism demand. *Annals of Tourism Research, 39*(4), 1859–1883.

Gretzel, U., Fesenmaier, D. R., Formica, S., & OLeary, J. T. (2006). Searching for the future: Challenges faced by destination marketing organizations. *Journal of Travel Research, 45*(2), 116–126.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100–108.

Huang, W., & Ribeiro, A. (2018). Network comparison: Embeddings and interiors. *IEEE Transactions on Signal Processing, 66*(2), 412–427.

Jang, S. S., & Ham, S. (2009). A double-hurdle analysis of travel expenditure: Baby boomer seniors versus older seniors. *Tourism Management, 30*(3), 372–380.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.

Khalilzadeh, J., & Wang, Y. (2018). The economics of attitudes: A different approach to utility functions of players in tourism marketing coalitional networks. *Tourism Management, 65*, 14–28.

Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, 33–50.

Liang, H., & Wang, H. (2017). Structure-function network mapping and its assessment via persistent homology. *PLoS Computational Biology, 13*(1), e1005325.

Li, S., Chen, T., Wang, L., & Ming, C. (2018c). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management, 68*, 116–126.

Ligges, U., & Mächler, M. (2003). Scatterplot3d - an R Package for visualizing multivariate data. *Journal of Statistical Software, 8*(11), 1–20.

Li, K. X., Jin, M., & Shi, W. (2018b). Tourism as an important impetus to promoting economic growth: A critical review. *Tourism Management Perspectives, 26*, 135–142.

Line, N. D., & Wang, Y. (2017). Market-oriented destination marketing: An operationalization. *Journal of Travel Research, 56*(1), 122–135.

Lin, V. S., Mao, R., & Song, H. (2015). Tourism expenditure patterns in China. *Annals of Tourism Research, 54*, 100–117.
Li, X., Pan, B., Law, R., & Huang, X. (2017b). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66.
Li, H., Song, H., & Li, L. (2017a). A dynamic panel data analysis of climate and tourism demand: Additional evidence. *Journal of Travel Research, 56*(2), 158–171.
Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018a). Big data in tourism research: A literature review. *Tourism Management, 68*, 301–323.
Luo, Q., & Zhong, D. (2015). Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites. *Tourism Management, 46*, 274–282.
Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: A systematic literature review. *International Journal of Contemporary Hospitality Management, 30*(12), 3514–3554.
Marrocu, E., Paci, R., & Zara, A. (2015). Micro-economic determinants of tourist expenditure: A quantile regression approach. *Tourism Management, 50*, 13–30.
Mathies, C., Gudergan, S. P., & Wang, P. Z. (2013). The effects of customer-centric marketing and revenue management on travelers choices. *Journal of Travel Research, 52*(4), 479–493.
MATLAB Optimization Toolbox. (2017). *Matlab optimization toolbox*.
McCamley, C., & Gilmore, A. (2018). Strategic marketing planning for heritage tourism: A conceptual model and empirical findings from two emerging heritage regions. *Journal of Strategic Marketing, 26*(2), 156–173.
Molera, L., & Albaladejo, I. P. (2007). Profiling segments of tourists in rural areas of South-Eastern Spain. *Tourism Management, 28*(3), 757–767.
Olya, H. G., & Mehran, J. (2017). Modelling tourism expenditure using complexity theory. *Journal of Business Research, 75*, 147–158.
ONS. (2017). *International passenger survey [data collection] SN: 8016*.
Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science, 6*(1), 17.
Pereira, C. M., & de Mello, R. F. (2015). Persistent homology for time series and spatial data clustering. *Expert Systems with Applications, 42*(15–16), 6026–6038.
Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P. J., et al. (2014). Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface, 11*(101), 20140873.
Pike, S., & Page, S. J. (2014). Destination marketing organizations and destination marketing: A narrative analysis of the literature. *Tourism Management, 41*, 202–227.
R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Ramires, A., Brandão, F., & Sousa, A. C. (2018). Motivation-based cluster analysis of international tourists visiting a World Heritage City: The case of Porto, Portugal. *Journal of Destination Marketing & Management, 8*, 49–60.
Rashidi, T. H., & Koo, T. T. (2016). An analysis on travel party composition and expenditure: A discrete-continuous model. *Annals of Tourism Research, 56*, 48–64.
Rigall-I-Torrent, R., & Fluvià, M. (2011). Managing tourism products and destinations embedding public good components: A hedonic approach. *Tourism Management, 32*(2), 244–255.
Robinson, R. N., Getz, D., & Dolnicar, S. (2018). Food tourism subsegments: A data-driven analysis. *International Journal of Tourism Research, 20*(3), 367–377.
Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., et al. (2019). Clustering algorithms: A comparative approach. *PloS One, 14*(1), e0210236.
van Rossum, G. (1995). *Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica*. Amsterdam: CWI.
Rudkin, S., Sharma, A., et al. (2017). Enhancing understanding of tourist spending using unconditional quantile regression. *Annals of Tourism Research, 66*, 188–191.
Shahzad, S. J. H., Shahbaz, M., Ferrer, R., & Kumar, R. R. (2017). Tourism-led growth hypothesis in the top ten tourist destinations: New evidence using the quantile-on-quantile approach. *Tourism Management, 60*, 223–232.
Siliverstovs, B., & Wochner, D. S. (2018). Google Trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization, 145*, 1–23.
Sousa, M. J., & Rocha, Á. (2019). Skills for disruptive digital business. *Journal of Business Research, 94*, 257–263.
Tausz, A., Vejdemo-Johansson, M., & Adams, H. (2014). JavaPlex: A research software package for persistent (co)homology. In H. Hong, & C. Yap (Eds.), *Proceedings of ICMS 2014, lecture notes in computer science 8592* (pp. 129–136).
Thrane, C. (2014). Modelling micro-level tourism expenditure: Recommendations on the choice of independent variables, functional form and estimation technique. *Tourism Economics, 20*(1), 51–60.
Veisten, K., Haukeland, J. V., Baardsen, S., Degnes-Ødemark, H., & Grue, B. (2015). Tourist segments for new facilities in national park areas: Profiling tourists in Norway based on psychographics and demographics. *Journal of Hospitality Marketing & Management, 24*(5), 486–510.
Vogt, C. A. (2011). Customer relationship management in tourism: Management needs and research applications. *Journal of Travel Research, 50*(4), 356–364.
Wang, L., Fang, B., & Law, R. (2018a). Effect of air quality in the place of origin on outbound tourism demand: Disposable income as a moderator. *Tourism Management, 68*, 152–161.
Wang, L., Fong, D. K. C., Law, R., & Fang, B. (2018b). Length of stay: Its determinants and outcomes. *Journal of Travel Research, 57*(4), 472–482.
Wang, Y., Li, X., & Lai, K. (2018c). A meeting of the minds: Exploring the core–periphery structure and retrieval paths of destination image using social network analysis. *Journal of Travel Research, 57*(5), 612–626.
Weinberger, S. (2011). What is… persistent homology? *Notices of the AMS, 58*(1), 36–39.
Wilkes, R. E. (1995). Household life-cycle stages, transitions, and product expenditures. *Journal of Consumer Research, 22*(1), 27–42.
Wong, I. A., Fong, L. H. N., & Law, R. (2016). A longitudinal multilevel model of tourist outbound travel behavior and the dual-cycle model. *Journal of Travel Research, 55*(7), 957–970.
Wu, L., Zhang, J., & Fujiwara, A. (2013). Tourism participation and expenditure behaviour: Analysis using a scobit based discrete–continuous choice model. *Annals of Tourism Research, 40*, 1–17.
Xia, K., & Wei, G.-W. (2014). Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering, 30*(8), 814–844.
Xia, K., Zhao, Z., & Wei, G.-W. (2015). Multiresolution persistent homology for excessively large biomolecular datasets. *The Journal of Chemical Physics, 143*(13), 10B603_1.
Yang, Y., Fik, T. J., & Altschuler, B. (2018). Explaining regional economic multipliers of tourism: Does cross-regional heterogeneity exist? *Asia Pacific Journal of Tourism Research, 23*(1), 15–23.
Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry, 33*(2), 249–274.
Zomorodian, A., & Carlsson, G. (2008). Localized homology. *Computational Geometry, 41*(3), 126–148.

Dr. Tristan W. Chong is an Associate Professor and Area Head of Marketing at the SP Jain School of Global Management. His research interests are in the areas of B2B marketing, marketing analytics and topological data analysis and applications. His research has been continuously funded by both internal and external funding bodies including the Natural Science Foundation of China. He is also a Visiting Scholar at the National University of Singapore; Deakin University, Australia; Heriot Watt University, UK; and South University of Science and Technology of China.

Dr Simon Rudkin is a Senior Lecturer in Economics at the Swansea University School of Management. Motivated by the ability of Economics and Statistics to bring new insight to market understanding, his research interests include consumer sentiment, traveller behaviour and the power of big data within tourism.