A Guest Mix Approach to Analysing City Tourism Competition

8

Christian Buchta and Josef A. Mazanec

8.1 Introduction

Chapter 8 builds on previous results published in the first edition of this reader. Section 4.1, Part IV, of the first edition presented ‘A guest mix approach’ to assessing and visualising the competitive relationships among 16 European tourist cities as reflected in the guests’ distribution by nationalities. If two cities A and B exhibit very similar proportions of their guest nationalities it is likely that the CTO managers of A pay the same attention to each of these guest nations as the managers of B. In other words, the analysis rests on the assumption that the CTOs base their marketing effort on a geographical segmentation approach. This does not seem to be a severe restriction as it corresponds to customary strategy guidelines followed by many tourist organisations.

In the following the authors demonstrate how bednight statistics may be exploited to classify and position city destinations in a competitive space defined by tourism generating countries. Particular emphasis is on tackling the missing data problem. International organisations such as the UNWTO or the EU, the European Travel Commission or European Cities Tourism have made great efforts to harmonise and complement the international statistics on arrivals and bednights. Despite all these initiatives international tourism statistics are still plagued by inconsistency and/or lack of data. Tourism researchers as well as managers are constantly challenged by how to overcome these insufficiencies. In this case example they will find advice on data pre-processing steps that may turn out to be instrumental when facing incomplete data. If a limited amount of missing data is tolerated or replaced the analyst is particularly responsible for drawing cautious conclusions. When interpreting results we will remind ourselves of this principle and sort out spurious effects likely to be attributable to missing values.

8.2 The city database

Let us initially point out that we intend to compare our findings with previous results. Therefore, we use the same destinations and markets as analysed in Mazanec (1997). This may come at the price that markets and destinations that were not important back then could be important now but are not included in the study and vice versa. However, as will be shown below, the present data are incomplete. As the previ-

Table 1 Number of observations by destinations and type of accommodation for 17 markets of origin in 1995–2007

	Type of accommodation					
City	NA	NAS	NG	NGS	NZS	Total
AMS (Amsterdam)	0	0	198	0	0	198
BER (Berlin)	221	0	34	0	0	255
BRU (Brussels)	219	0	0	0	0	219
BUD (Budapest)	217	0	0	0	0	217
HEL (Helsinki)	221	0	0	0	0	221
LIS (Lisbon)	0	0	207	0	0	207
LON (London)	0	0	0	0	187	187
MAD (Madrid)	0	168	0	45	0	213
OSL (Oslo)	202	0	0	0	0	202
PAR (Paris)	0	0	179	0	0	179
PRG (Prague	179	0	0	0	0	179
ROM (Rome)	156	0	149	0	0	305
STO (Stockholm)	219	212	0	0	0	431
VIE (Vienna)	153	221	34	0	0	408
ZAG (Zagreb)	199	0	0	0	0	199
ZUR (Zurich)	181	0	0	0	0	181
Total	2,167	601	801	45	187	3,801

Table 2 Summary of NAS/NA ratios for Vienna in 1995–2007

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.007	1.029	1.039	1.056	1.071	1.261

ous study was kept free of missing values we also regard it as a reference for validating current results. Changing the definitions of markets or the set of destinations would not alter the methodology and is therefore left to future work and the reader's own exercise.

The city bednight data were obtained from the TourMIS database available online at www.tourmis.info in 8/2008 (and cross-checked for missing values again in 3/2009). As the previous study ended in 1994 we retrieved data from 1995 to 2007. Data based on different definitions of the city area as well as the type of accommodation are available. See Table 1 for a summary for the period 1995–2007. For example, 11 of the 16 cities report figures for 'Type' equals 'Bednights in all paid forms of accommodation establishments in city area' (abbreviated 'NA' in TourMis) amounting to a total number of 2,167 observations across 17 markets of origin for 1995–2007.

For cities where we had a choice we decided to use the category most complete in terms of observations, as in the case of Stockholm (NA), or Vienna ('Bednights in all paid forms of accommodation establishments in greater city area' = NAS). Note that although the latter definition is widening, i. e. encompasses the former, Vienna's market position will not be overstated as evidenced by Table 2.

For Rome it turned out that the data for NA and 'Bednights in hotels and similar establishments in city area only' (= NG) are identical up to 2002. As there were fewer missing observa-

Table 3 Number of missing observations by cities and years for 17 countries of origin

	Year													
City	95	96	97	98	99	00	01	02	03	04	05	06	07	Total
AMS	2	2	2	3	3	3	2	2	2	1	1	0	0	23
BER	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRU	0	0	0	0	0	0	0	0	0	1	1	0	0	2
BUD	0	0	0	0	0	1	1	1	1	1	0	0	0	5
HEL	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LIS	3	3	5	3	1	1	0	0	0	0	0	0	0	16
LON	0	0	0	16	0	0	0	0	0	0	0	1	17	34
MAD	1	1	0	0	0	0	0	0	0	10	17	17	17	63
OSL	0	0	0	0	0	0	0	0	0	9	0	0	10	19
PAR	7	6	5	5	5	1	1	1	1	1	1	1	7	42
PRG	0	0	0	0	0	17	0	15	0	0	0	0	10	42
ROM	0	0	0	0	0	0	0	0	0	0	0	11	10	21
STO	1	1	0	0	0	0	0	0	0	1	0	0	0	3
VIE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ZAG	2	2	1	17	0	0	0	0	0	0	0	0	0	22
ZUR	8	8	8	8	8	0	0	0	0	0	0	0	0	40
Total	24	23	21	52	17	23	4	19	4	24	20	30	71	332

tions for NG after 2002 we settled for NG and filled in the missing values from 1995 to 1997 with the figures from NA.

For Madrid we decided not to use the 'Bednights in hotels and similar establishments in greater city area' (= NGS) figures to complement the NAS category as the latter are on average 9% higher and therefore its use might introduce spurious effects over time.

For the cities Amsterdam, Lisbon, and Paris only the hotel categories are reported, so we have to assume that they represent the dominant part of a destination's market. Finally, London reports on 'Bednights in all accommodation establishments including visiting friends and relatives in greater city area' (= NZS), which means that unpaid forms of accommodation are included. Therefore, its definition is wider than NAS.

Despite the fact that the database we compiled for further analysis is based on partly different definitions the important point to note is that for each destination the data are consistent over time. Thus, a city's market position may be biased overall but changes in relative positions over time, if any, can be expected to be consistent indicators of true shifts in market structure.

As a result based on the selections discussed 3,801 available observations were reduced to 3,204 observations. However, as we need a total of 16 × 17 × 13 = 3,536 observations there are 332 observations missing (or 9.4% of the data). Table 3 shows the distribution of missing observations across cities and years. The maximum total number per year occurs in 2007. This suggests dropping the year 2007 from further analysis. For Madrid, London,

Table 4 Number of 1, 2,... consecutive periods with missing values in a market series by cities for 1995–2006

	Runs												
City	1	2	3	4	5	6	7	8	9	10	11	12	Total
AMS	0	0	0	0	0	2	0	0	0	0	1	0	23
BER	0	0	0	0	0	0	0	0	0	0	0	0	0
BRU	0	1	0	0	0	0	0	0	0	0	0	0	2
BUD	0	0	0	0	1	0	0	0	0	0	0	0	5
HEL	0	0	0	0	0	0	0	0	0	0	0	0	0
LIS	2	0	0	2	0	1	0	0	0	0	0	0	16
LON	17	0	0	0	0	0	0	0	0	0	0	0	17
MAD	0	8	10	0	0	0	0	0	0	0	0	0	46
OSL	9	0	0	0	0	0	0	0	0	0	0	0	9
PAR	1	1	0	0	4	0	0	0	0	0	0	1	35
PRG	32	0	0	0	0	0	0	0	0	0	0	0	32
ROM	11	0	0	0	0	0	0	0	0	0	0	0	11
STO	1	1	0	0	0	0	0	0	0	0	0	0	3
VIE	0	0	0	0	0	0	0	0	0	0	0	0	0
ZAG	16	1	0	1	0	0	0	0	0	0	0	0	22
ZUR	0	0	0	0	8	0	0	0	0	0	0	0	40
Total	89	24	30	12	65	18	0	0	0	0	11	12	261

Prague, and Zagreb we observe years where no data at all are available. In total, Madrid shows the highest number of missing values, followed by Paris and Prague. Whereas for Prague the missing values accumulate in the years 2000 and 2002, for Paris they are scattered over all periods. Similarly, Amsterdam shows only a single period without a missing value. Zurich comes in third place missing 40 observations up to the year 2000.

To learn about the temporal structure of missing observations we computed so-called runs statistics. These are counts of the number of consecutive periods a missing value occurs in a time series for a city's country of origin. Table 4 shows the figures for the reduced dataset 1995–2006. For example, among the Madrid series there are ten runs with three consecutive periods of missing values. Going back to Table 3 we see that no data were reported in 2005–2007. Note that a series may be counted more than once, e. g., for Prague there are 32 single-period runs, 17 in 2000 and 15 in 2002. The row totals show again the total number of missing values for a city and the column totals the total number corresponding to a run, i. e. the column sum times the length of a run.

For Paris there is not one observation for the Australian market and for further five markets (Finland, Greece, Norway, and Sweden) observations are missing for 1995–1999. Zurich did not report data on eight markets (Australia, Belgium, Canada, Finland, Greece, Netherlands, Norway, and Spain) up to the year 2000 (see Table 4).

In sum, 66% of the missing values pertain to consecutive instances of missingness and, therefore, most likely have to be attributed

Table 5 Sum of median bednights, total market share and percentage of missing values (PM) by cities for 1995–2006

	Bednights	Share	PM
AMS	6,400,350	3.65	10
BER	10,098,254	5.76	0
BRU	3,503,169	2.00	1
BUD	2,869,911	1.64	2
HEL	1,939,898	1.11	0
LIS	2,744,018	1.56	7
LON	79,377,500	45.25	8
MAD	9,460,248	5.39	21
OSL	2,286,094	1.30	4
PAR	26,780,746	15.26	16
PRG	4,919,282	2.80	14
ROM	12,056,062	6.87	5
STO	3,596,378	2.05	1
VIE	6,942,307	3.96	0
ZAG	258,566	0.15	10
ZUR	2,206,420	1.26	18

Table 6 Robust trend estimates and medians for Madrid and Zurich 1995–2006

	Trend		Median	
	MAD	ZUR	MAD	ZUR
AT	815	1,324	34,756	51,450
AU	3,578	3,092	45,340	33,846
BE	4,290	–496	81,766	21,957
CA	5,613	ns	47,994	40,148
CH	2,398	32,366	52,113	709,008
DE	12,240	17,292	318,136	390,238
ES	301,227	4,323	6,170,039	62,574
FI	2,402	ns	14,211	11,186
FR	22,225	2,423	370,290	74,156
GR	–1,061	ns	46,887	16,594
IT	14,751	1,342	466,342	75,120
JP	–13,421	–3,771	326,042	107,908
NL	5,206	ns	102,287	51,303
NO	2,420	ns	22,150	12,258
SE	1,628	ns	46,266	28,076
UK	32,440	6,757	468,195	187,304
US	80,372	ns	847,434	333,293

to structural problems in the data acquisition process. Note that with the elimination of 2007 the total percentage of missing observations drops to 8%.

How do we deal with such structural deficiencies in a dataset? Possible routes to take will be discussed step by step in Section 3. Let us first introduce a practical approach to computing the overall market shares of the cities. The median is a robust statistical measure of the central value of a collection of data points. With missing values we have varying numbers of observations available, making the estimates more or less reliable. Further, the statistics could be biased. For example, if values are missing at the ends of a series with a trend the estimates would be biased up- or downwards. E. g., assume a positive trend. Then the values at the beginning are lower and, if missing, the estimate will be biased upwards and vice versa. A robust measure can only ease the problem but not eliminate it.

Table 5 shows the sum of the median number of bednights across all markets, the total market share, and the percentage of missing values (PM). We see that London and Paris dominate the market. Disregarding the systematic concerns raised above the London figures are not less reliable than those of the lower end destinations Lisbon or Zagreb. Although Paris has double the percentage of missing values its share is more than double the share of Rome. The numbers for Madrid and Zurich could be biased down- and upwards as the missing values are substantial. Table 6 shows robust (linear regression) trend estimates and medians for both cities, where 'ns' indicates that an estimate (model) was insignificant at the 5%

Table 7 Market shares of 17 countries of origin and diversification index (entropy) by cities 1995–2006

City	Market																	
	AT	AU	BE	CA	CH	DE	ES	FI	FR	GR	IT	JP	NL	NO	SE	UK	US	DI
AMS	1	2	2	2	2	8	6	1	6	1	6	3	14	1	1	27	17	2.28
BER	1	0	1	0	2	78	1	0	1	0	2	1	2	0	1	3	4	1.08
BRU	1	0	12	1	2	11	6	1	14	2	6	4	9	1	2	19	9	2.44
BUD	5	1	2	1	3	23	8	3	6	2	12	5	4	2	4	9	13	2.48
HEL	1	1	1	1	2	6	2	51	2	1	3	4	2	3	7	8	6	1.89
LIS	2	1	2	2	3	13	23	1	10	1	12	3	4	2	2	10	10	2.37
LON	1	5	1	3	2	7	4	1	7	1	5	2	2	1	2	38	17	2.14
MAD	0	0	1	1	1	3	65	0	4	0	5	3	1	0	0	5	9	1.42
OSL	0	0	0	0	1	4	2	1	1	0	2	2	1	68	6	5	5	1.37
PAR	0		2	1	2	5	5	0	39	1	7	7	2	0	1	11	16	1.99
PRG	3	1	2	1	2	27	8	2	6	1	13	3	4	2	3	11	9	2.40
ROM	1	2	1	1	1	7	4	0	3	1	36	11	1	1	2	6	21	2.03
STO	0	0	1	1	1	4	2	3	2	0	2	2	1	3	66	6	6	1.45
VIE	22	1	1	1	4	27	4	1	4	1	10	5	2	1	1	5	9	2.24
ZAG	9	2	3	4	2	21	3	1	8	1	12	3	4	1	3	11	13	2.47
ZUR	2	2	1	2	32	18	3	1	3	1	3	5	2	1	1	8	15	2.15

level. Overall, the trends are positive except for Japan (Belgium). For Zurich six of the series with missing values (Canada, Finland, Greece, Netherlands, Norway, and Sweden) have insignificant estimates. However, the share of the sum of the median values of these markets in the overall market share value is only 7%.

Let us finally take a look at the median market shares of the countries of origin by destinations. The market shares in Table 7 denote the generating countries' relative contributions to a city's inbound tourism. In nine of the destinations the home market is dominant: Berlin (78%), Helsinki (51%), London (38%), Madrid

(65%), Oslo (68%), Paris (39%), Rome (36%), Stockholm (66%), and Zurich (32%). Keep in mind that the home markets of Budapest, Lisbon, Prague, and Zagreb are not represented in the database. Germany (DE) is particularly dominant in the destinations Budapest (23%), Prague (27%), and Zagreb (21%). Spain (ES) is dominant in Lisbon (23%). Amsterdam draws visitors mainly from the United Kingdom (27%); Vienna does equally from Germany (27%) but has a strong home market (22%) too. According to the diversification index DI (measuring entropy across the cities' market shares) Budapest has the most balanced guest mix, followed by Zagreb and Brussels. Note that the dominant home market cities have the lowest index values.

Hint

The computational steps used to prepare the database can be found in the scripts *prepare.R* and *trend.R* available from http://www.wu.ac.at/itf/downloads/software/guestmix.

8.3 The missing value problem

How to deal with incomplete data? In the best case we might use some statistical model of missing observations. However, the analysis of the structural defects in our data led to the conclusion that this is not recommendable. What distributional assumptions should we make? Do we at least have reliable information on the minimum, maximum, or expected values? Such information does not come from incomplete data. Nevertheless, as predicting from existing data is a common practice we take a closer look at some of these methods. The interested reader will appreciate the overviews given in Lemieux and McAlister (2005), Allison (2001), or Schafer (1997).

There are two broad categories of data-driven models we could use, univariate or multidimensional time series models and multidimensional cross-sectional models. In the first category, once data are missing over an extended period of time, predictions from a time series model without exogenous variables become too unreliable or even infeasible. On the other hand, we could exploit correlations across or within destinations. This is more involved but again has its limits if either the target or explanatory series contain missing values over an extended period of time. At the other extreme, modelling each missing value separately, possibly using expert knowledge on the destination or market, is not an option either as we have a total of 261 missing values (see Table 3; recall that 2007 was excluded).

In the second category, we must assume that the destination profiles hold information about each other. First we have to determine which profiles to use for prediction. In other words, we have to identify the neighbours of a multidimensional data point. Note that we therefore must decide on some concept of proximity, too. Then we can compute the averages across the neighbours for prediction. Depending on how we define a neighbourhood this imputation method provides both local and global predictions. A local model uses only close neighbours assuming that similar profiles provide more reliable information on missing values than less similar ones. However, the less information a profile provides due to missing values, the less reliably it can be assigned to a neighbourhood. Information here means the relevant information as quantified by market share. Thus, information on important markets should not be missing. But there is a fundamental catch: with increasing dimensionality 'close' neighbours becomes a scarce commodity. This is one aspect of what is known in the literature as 'the curse of dimensionality' (Hastie, Tibshirani and Friedman, 2001, p. 22).

Recall Table 3 and verify that only eight city profiles are missing more than half of their market values. Therefore, a cross-sectional approach would be viable. However, as mentioned above, important differences between a profile and its neighbours might exist. As the analysis of the competition among destinations depends on this information we might not gain more insight. On the contrary, due to the averaging, we might blur the picture. Therefore, if we fill in missing values we have to determine the possible influence the size of the neighbourhood has on the analysis. So why fill in missing values at all? Unfortunately, some class of models, for example linear regression, do not work with missing values. Cases with missing values must be excluded entirely even if a single component is missing. Clearly, for such models there is a trade-off to accept between either dropping a case or possibly biasing the available information with inaccurate predictions of missing values. Unfortunately, the multidimensional-scaling models we are going to use do not allow for excluding missing cases (although they are linear models) implying that we must predict missing values.

The concept of locality provides the basic ideas for pursuing a simple approach. Local estimators are typically obtained with *k* nearest-neighbour methods, where the optimal *k* can be determined by minimizing the prediction error. However, we cannot determine *k* as we cannot compute the prediction error of missing values, and further, a uniform *k* may not be appropriate. Therefore, we suggest using cluster analysis: it provides neighbourhoods (i. e. the clusters) but it does not rely on prediction errors. Also, the optimal *k*, the proper number of clusters, can be based on various measures of information of a solution. The information is contained in the representatives of the clusters which are vector-based measures known collectively as centroids, for example, the vector of means or medians. They are local estimators and using all the available information to compute them is the best we can do. In short, what we suggest is to modify a clustering method's proximity and summation functions to omit missing values instead of omitting data points. As the former measures use pairs of values both observations must be available to be taken into account. City profiles that do not provide sufficient information should be collected in a separate cluster.

The 1997 study used market share profiles to represent a destination's guest mix. As the database of the present study is incomplete and we prefer a methodology that does not fill in missing values we cannot use market share profiles. If a single market value is missing the total cannot be computed. Therefore, we need a proximity measure that is appropriate for market share data but does not depend on missing values. The cosine similarity is such a measure and it is well known in the classification and neurocomputing literature (Caudill, 1993, p. 18 f.). It maps equally directed vectors to the same value but does not take the lengths of the vectors into account. Thus, if two market share profiles are the same the cosine similarity is maximal, i. e. attains the value 1, irrespective of whether the data are in absolute or relative terms. In other words, differences in destination market size are ignored. Missing values may be handled by mapping vectors into a common subspace. Consider an example with one missing value. The vectors (2, NA, 2) and (1, 1, 1) map to the common subspace with dimensions 1 and 3 (NA indicates 'not available'). In this subspace both have the same direction. Note that the first vector has the same direction as the vector (1, 2, 1) but for any real value of NA both cannot be true. This approach is equivalent to removing cases from the input data as indicated above.

Another approach is to make most use of the available data and compute the numerator and denominator in the cosine similarity formula separately (see Meyer and Buchta, 2008). This is similar to setting NA = 0 except for special cases. The intention of this approach is to factor in the unreliability due to missing values:

the similarities in the above example decrease to .82 and .58 respectively as the lengths of the vectors increase. We expect that this will assist a clustering algorithm in assigning profiles with missing values to lower-dimensional clusters.

In whatever way we tackle the missing value problem we should keep in mind that there are only likely neighbourhoods and, therefore, we must be careful when interpreting results.

Question

For which value of NA is the vector (2, NA, 2) equally similar to both of the complete vectors (1, 1, 1) and (1, 2, 1)?

8.4 Clustering guest mix profiles

The data matrix consists of the guest mix profiles for 16 European cities. As the profiles range over a time period of 12 years (1995–2006) each city contributes 12 repeated measurements yielding a total of 16 × 12 = 192 data records. The working step outlined in this section seeks to identify typical guest mix patterns in this database. Having generated a number of such guest mix prototypes we may ask which city is represented by which pattern in which calendar year.

Cluster analysis is a very mature but complex field (Aldenderfer and Blashfield, 1984; Tan et al. 2006). Among the well-known approaches are hierarchical and partitioning methods. The former have the advantage of producing deterministic results while the latter usually do not. For example, the *k*-centroids algorithm improves a solution in two steps: first it computes the centroids of the current partition, e. g., the mean vectors of the clusters. Then it computes the new partition by assigning each data point to its closest centroid. The algorithm stops if there is no further change in partition. *K*-centroids is simple and fast and it can deal with large data sets, but there is no guarantee a solution is a global optimum. Therefore, it is recommended to try different initial solutions and retain the best. If the initial solutions are chosen randomly this amounts to performing random search and therefore the final solution may be random too. That is, if we repeat the procedure we may obtain considerably different results. This could complicate interpretation.

On the other hand, the hierarchical methods are limited by the necessity to compute the proximities for all possible pairs of data points and the number of pairs is a square function of the number of data points. Here, we use the two approaches in combination: we compute an initial solution for *k*-centroids by a hierarchical clustering method. This turned out to provide more stable results than an extensive random search.

R function for computing *k*-means with missing value data[1]

For readers interested in computational details the listing at the bottom of this box shows a code snippet for computing *k*-means with missing value data with the statistical software *R*. The first input argument, *x*, is a data matrix with market profiles in the rows. The second, *k*, is a vector of cluster labels corresponding with the rows of *x*. Now, we apply the function *tapply* to each column (by setting MARGIN = 2) of *x*; *tapply* partitions a column (vector) according to *k* and applies the function on lines 3 to 6 to each subset. On line 4 we compute the mean,

[1] *R* is an open source system available via http://cran.r-project.org/.

mandating that missing values be removed from the computation (by setting na.rm = TRUE). On line 5 we test if the result is not a number (NaN) which will be obtained if the removal of missing values results in a vector of zero-length, which in turn results in a division by zero. In that case we return *R*'s missing value code NA, as we don't know what the result is and otherwise the mean. *R* pieces the results from the (nested) function calls together and returns a vector or a matrix with the centroids in the columns and the markets in the rows.[2] For example, try *mean.k(as.matrix(c(1,NA,1,NA)), c(1,1,2,3))* and verify the result.

```
1  mean.k ← function(x, k)
2        apply(x, MARGIN = 2, tapply, k,
3              function(x) {
4                    x ← mean(x, na.rm = TRUE)
5                    if (is.nan(x)) NA else x
6              }
7        )
```

[2] Note that this adaptation is essential, as setting NA = 0 would underestimate the mean. In other words, we cannot just set NA = 0 to work around an implementation of a clustering algorithm that cannot handle missing values.

Question

What needs to be changed in the code for computing k-medians?

We determine the number of clusters by a combination of visual inspection of the cluster dendrogram and goodness of fit measures of partitions. A dendrogram is a tree. At the bottom each data point is in a separate cluster and at the top all the data points are in the same cluster. The interior nodes represent the merging steps of the clustering algorithm. The height of a node quantifies the homogeneity of the corresponding cluster. We use the average within-cluster similarity for the height. The loss in detail that comes with merging can be seen from the changes in height, which therefore provide the basis for the decision where to cut the tree. Figure 1 shows the dendrogram of the guest mix data. Note that the heights represent average cosine distances instead of similarities. Cutting the tree at a height of less than .1 results in a partition with 17 or more clusters. The rectangles at the bottom represent a partition with 17 clusters. From a practical point of view we must consider the possibility that the destinations could be quite distinct and therefore using less than 16 clusters does not seem appropriate. An additional cluster may be needed to collect profiles with missing values.

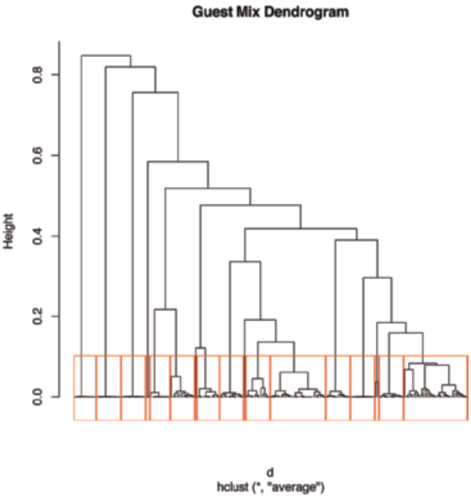

Fig. 1 Hierarchical clustering results for the 192 guest mix profiles

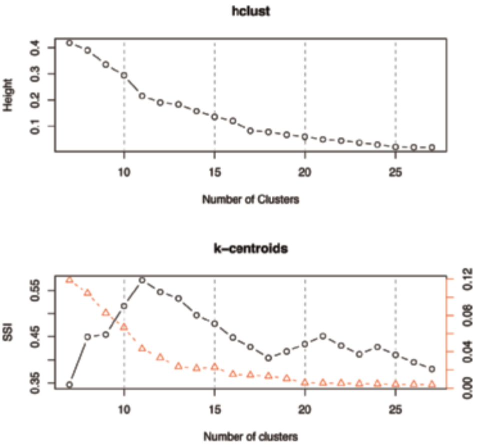

Fig. 2 Goodness-of-fit measures for solutions with 7 to 27 clusters by clustering methods, hierarchical (top) and centroid-based (bottom)

However, from a quantitative point of view, fewer clusters may be recommended.

Figure 2 (top) shows the height for solutions with between 7 and 27 clusters. Overall, the height decreases with increasing numbers of clusters. As indicated above, it is recommended to look at the differences, where adding a further cluster (adding more detail to the picture) does not lead to a significant reduction in average within-cluster distance. Thus, solutions with 11 and 17 clusters are possible candidates.

The average within-cluster distance is also the optimisation criterion of the *k*-centroids method. However, with cosine distances the procedure described earlier is not guaranteed to maximise this measure. Without elaborating the details (see Leisch, 2006), we assume that it produces good results. After all the whole approach we suggest is heuristic in nature. We used the simple structure index (Mazanec, 2001) as a further measure to assess the proper level of detail to use in our market analysis. That is, the average dimension range across the cluster profile adjusted for the number of profiles that contain an extreme value on at least one dimension (guest nationality). For example, for the hypothetical profiles (1, 0) and (0, 1) the index is 1 (maximal) but with (1, 1), instead of the first or second, it drops to 1/2 as this set of profiles does not provide distinctive information on the first dimension. Further, if we added (1/2, 1/2) to a solution with (1, 0) and (0, 1) the index would drop to 2/3 as the additional profile does not provide any distinctive information.

Figure 2 (bottom) shows the figures for the average within-cluster distance (triangles) and the simple structure index (circles). According to the former, the 13, 15, and 20 cluster solutions are possible candidates for further analysis. The simple structure index, on the other hand, attains a global maximum at the 11 cluster solution, and has two further local maxima at 21 and 24 clusters. Note that the index values are never close to 1, which indicates that the cluster profiles contain redundant information.

Now, if we are interested in less detail we could use 11 clusters and if we are interested in more fine-grained results we could use 17, 21, or 24. However, fewer clusters also means bigger clusters (larger neighbourhoods), and vice versa, and therefore more reliable predictions.

Closer inspection of the city market maps of the candidate solutions reveals that the 21 and 24 cluster solutions are too detailed and that the 11 cluster solution is too coarse. Therefore, we present the solution with 17 clusters as it nicely illustrates the trade-off to be made between quantitative criteria and the necessities of missing value analysis. Table 8 shows the distribution of guest mix profiles across clusters. For example, Amsterdam and London are represented with 12 and 11 profiles in cluster 1. The remaining London profile is separated out into cluster 15. 10 of the profiles of Prague are in cluster 4, together with 12 of the profiles of Budapest. One of its profiles is singled out into the uninformative cluster (last column) and another can be found in cluster 2, together with the 12 profiles of Berlin. Besides for London, profiles are singled out for Oslo (1), Rom (1), Zagreb (2 and 1), and Madrid (1). Zagreb

Table 8 Distribution of cities across clusters

	Cluster																	
City	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**NA**
AMS	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BER	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRU	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BUD	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HEL	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
LIS	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0
LON	11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
MAD	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	2
OSL	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	1	0	0
PAR	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0
PRG	0	1	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ROM	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	1	0
STO	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0
VIE	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0
ZAG	0	0	1	8	0	0	0	0	0	0	0	0	0	2	0	0	0	1
ZUR	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0
Total	23	13	13	30	12	12	10	11	12	11	12	12	12	2	1	1	1	4
Cities	2	2	2	3	1	1	1	1	1	1	1	1	1	1	1	1	1	3

also occupies cluster 4 with 8 of its profiles. Remember that each profile has a year tag and that the profiles of a cluster are more similar to their own centroid than to any other. Thus, cluster 4 indicates that Budapest, Prague, and Zagreb provide a similar guest mix across almost all periods of the study, and cluster 1 indicates this for Amsterdam and London. The remaining destinations seem to provide a unique guest mix.

Before we proceed to analysing the relationships among clusters we first have to secure that the profiles of a cluster are not too heterogeneous and their representatives are not too biased by missing values. Of course, the uninformative cluster and the single destination clusters, 14, 15, 16, and 17 serve a different purpose and therefore cannot be judged by these criteria. Table 9 shows the proportion of profiles with a missing value. First, observe that clusters 5, 8, 10, and 12 are unbiased, i. e. contain no missing values. As these are the exclusive positions of Helsinki, Oslo, Rome, and Vienna, their guest mixes are unbiased too. Second, only three of the single-profile clusters capture low-dimensional profiles as half or more of the values are missing. Cluster 14 representing Zagreb can be assumed to indicate a true shift in market position as neither Canada nor Greece can be expected to contribute significantly to its guest mix (see Table 7). Except for the Canadian market in cluster 9, the exclusive position of Paris, less than 50% of the values are missing at the worst. Further examination confirms that the affected markets are not the important ones. Especially, for any single-destination

Table 9 Proportion of profiles with a missing value × 100 by markets and clusters (rounded)

	Cluster																
Market	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
AT	0	8	0	0	0	33	10	0	17	0	0	0	0	0	100	100	100
AU	48	8	15	17	0	0	30	0	100	0	0	0	42	0	100	100	100
BE	26	8	0	0	0	0	10	0	0	0	8	0	42	0	100	100	100
CA	0	8	8	0	0	8	10	0	8	0	0	0	42	50	100	100	100
CH	0	8	0	0	0	0	10	0	0	0	0	0	0	0	100	100	100
DE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
ES	26	8	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
FI	0	8	0	0	0	33	10	0	42	0	0	0	42	0	100	100	100
FR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
GR	4	8	8	0	0	0	10	0	42	0	17	0	42	100	100	100	100
IT	0	8	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
JP	0	8	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
NL	0	8	0	0	0	0	10	0	0	0	0	0	42	0	100	100	100
NO	0	8	0	0	0	50	10	0	42	0	0	0	42	0	100	100	100
SE	0	8	0	0	0	0	0	0	42	0	0	0	42	0	100	0	100
UK	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
US	0	8	0	0	0	8	10	0	0	0	0	0	0	0	100	0	100
Size	12	7	7	16	6	6	5	6	6	6	6	6	6	1	1	1	1

cluster the affected profiles are still closer to this cluster than to any other thus implying that the missing information is indeed not relevant.

Table 10 shows the average and maximum within-cluster distances. Clusters 1 and 4 have the largest values and therefore are less homogeneous than the single destination clusters. Overall, the values are low, so the clusters are well represented by their mean profiles. But are the clusters well separated, too? The nearest neighbour of a data point need not be in the same cluster. The decision rule of *k*-centroids does not imply that. Therefore, we computed for each cluster the frequency distribution of the cluster indexes of the nearest neighbours of the data points. For clusters 1 to 13 the proportion of same-cluster indexes is 100 % and therefore the clusters are well separated.

Question

How can the concept of separation be further tightened?

Let the joint neighbourhood of a pair of centroids consist of the data points that are closer to both than to any other pair of centroids, i. e. closest to one and second closest to the other centroid. We use the proportion of the number of data points in a neighbourhood in the union of the pair of clusters to quantify the similarity of the clusters. To avoid spurious relationships we suggest putting a threshold on the distance to the second closest centroid. Asymmetric relationships can be quantified by the proportion of the number of data points in the intersec-

Table 10 Average and maximum within-class distance × 100.

Cluster													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
4.79	0.28	1.82	3.25	.05	0.87	0.13	.04	.34	0.60	.03	.25	0.52	1.00
10.34	3.28	7.46	9.35	.10	4.88	1.00	.09	.69	1.21	.07	.58	1.08	1.20

Table 11 Selected cluster centroids × 100 and diversification index

	Market																	
Cluster	AT	AU	BE	CA	CH	DE	ES	FI	FR	GR	IT	JP	NL	NO	SE	UK	US	DI
1	1	8	2	3	2	6	5	1	7	1	5	3	3	1	2	36	16	2.19
4	4	1	2	1	2	23	8	2	6	2	12	4	4	2	3	13	11	2.45
14	7	1	3	3	2	19	2	2	5	NA	9	1	4	2	2	8	30	2.24

tion of a neighbourhood with a cluster (in that cluster). For example, 70% of the data points of cluster 1 are in the neighbourhood with cluster 3, but for the latter 54% of its data points are closer to cluster 4, only 46% are closer to cluster 1. However, as we want to depict overall relationships in a market map we prefer the symmetric measure for drawing lines linking the cities' guest-mix positions.

Figure 3 shows a Sammon projection (Sammon, 1969) of the clusters computed from between-centroid Euclidean distances. The size of a bubble corresponds to the size of a cluster (= number of guest-mix profiles). It is coloured by the dominant market of the centroid and the label indicates the city or cities represented by a cluster. The large-font number labels indicate total market volume relative to the maximum (among all clusters). The width of a line corresponds with the symmetric link measure and the colon-separated numbers indicate the asymmetric link ratio. For example, cluster 1 corresponds to the position of Amsterdam and London, which is dominated by visitors from the UK. The total average market volume of cluster 1 is the largest among all clusters and therefore provides the basis. The neighbouring London cluster, number 15, consists of a single profile for 1998 with values missing in all markets except the UK. The high total market volume, relative to cluster 1, indicates that the centroid of cluster 1 is biased downward by the Amsterdam profiles as there the average share of the UK market is around 36% (see Table 11). The support of the neighbourhood of clusters 1 and 15 is 8/24 = 33% as indicated by the line width. The link ratio is 1:7 meaning that 7 profiles of cluster 1 are second closest to the

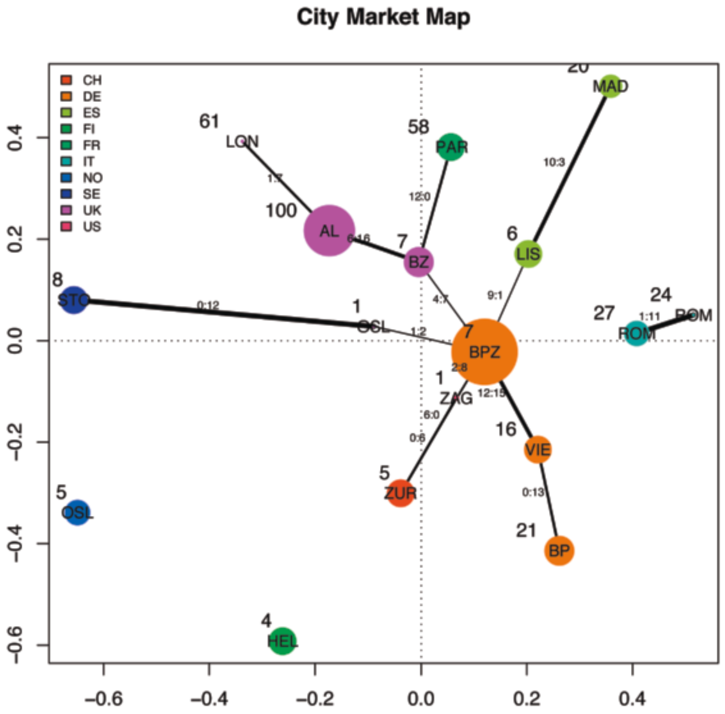

Fig. 3 Sammon projection of clusters based on between-cluster distances

centroid of cluster 15 and vice versa. However, cluster 3 with Brussels and Zagreb ('BZ') is closer to cluster 1, both in terms of map distance and support of the neighbourhood (61 %) or link ratio (6:16). Note that for Zagreb there is only a single profile for 1995. From 1996 to 1997 Zagreb occupies cluster 14 ('ZAG') with a dominant share of the US market (see Tables 8 and 11), with the value for 1996 being two- and three times the values of 1995 and 1997 respectively. Here the UK market is no longer important as in fact it decreased to half the value of 1995 by 1997.

Table 12 provides a summary of the city positions. A blank in the year column indicates the set 1995 to 2006 and a minus sign indicates exclusions from this set. Columns 'Market' and 'Share' show dominant markets (in cluster centroids) and the corresponding market shares (in average city profiles). Note that these figures may be biased upwards due to missing values. Column 'NAs' exhibits the number of missing values. Column 'Volume' shows the total market share of a city (in a cluster). Note that aggregate profiles with missing values were omitted from the computations, thus biasing the figures for the remaining cities upwards. Column 'DI' shows the diversification index of the average city profile. In case of missing values these figures are biased downwards.

Continuing where we paused above we see that the 1998 profile of Zagreb is in the uninformative cluster and all profiles after 1998 are in the central cluster 4. Note that a single Prague profile with 15 missing values was assigned to cluster 2. Most likely the market share in the German (DE) market is not as high as that of Berlin. In fact, our best guess would be around 24 % as reported in cluster 4 (see also Table 7). Remember from the analysis of cluster separation that the nearest neighbour of the Prague profile must be one of the Berlin profiles. So neither the cluster (centroid) nor the nearest-neighbour (profile) can be used to

Table 12 Summary of city positions

	Cluster	City	Year	Volume	Market	Share	NAs	DI
1	1	AMS	7	UK	26.6	0	2.30	
2	1	LON	–1998	93	UK	39.0	0	2.13
3	2	BER	100	DE	76.1	0	1.16	
4	2	PRG	2002	NA	DE	78.9	15	0.51
5	3	BRU	100	UK	19.5	0	2.44	
6	3	ZAG	1995	NA	UK	16.7	2	2.43
7	4	BUD	35	DE	22.0	0	2.49	
8	4	PRG	–2000,–2002	62	DE	23.8	0	2.40
9	4	ZAG	–1995:1998	3	DE	17.3	0	2.50
10	5	HEL	100	FI	50.7	0	1.91	
11	6	LIS	100	ES	23.5	0	2.38	
12	7	MAD	–2005:2006	100	ES	65.0	0	1.42
13	8	OSL	–2004	100	NO	67.1	0	1.39
14	9	PAR	100	FR	37.8	1	2.04	
15	10	ROM	–2006	100	IT	36.9	0	2.03
16	11	STO	100	SE	66.6	0	1.45	
17	12	VIE	100	DE	27.8	0	2.25	
18	13	ZUR	100	CH	30.3	0	2.17	
19	14	ZAG	1996:1997	100	US	29.9	1	2.24
20	15	LON	1998	100	UK	100.0	16	0.00
21	16	OSL	2004	100	UK	22.5	9	1.93
22	17	ROM	2006	100	IT	52.7	11	1.44
23	NA	MAD	2005:2006	NA	<NA>	NA	17	NA
24	NA	PRG	2000	NA	<NA>	NA	17	NA
25	NA	ZAG	1998	NA	<NA>	NA	17	NA

predict the missing values. The map position of cluster 16 ('OSL') and the tie with cluster 11 ('STO') is explained by the fact that data on Oslo's home market NO is missing. Thus, the UK and the SE markets are biased upwards with the former being the dominant market. Cluster 17 ('ROM') does at least not give that kind of false impression although the indicated shift in the share of Rome's home market IT is an artefact of missing values. The remaining clusters confirm the initial picture: a clear separation between home market destinations such as Madrid, Rome, Helsinki, Oslo, and, disregarding spurious ties, Stockholm. Berlin is a home market city, too. Its asymmetric relationship with Vienna (13:0 profiles) is not stable as solutions with more clusters show. On the other hand the tie between the 'VIE' and the 'BPZ' cluster is almost symmetric (12:15 profiles). The weak link between 'BPZ' and 'LIS' is asymmetric (1:9) and most likely reflects commonalities in the IT, UK, and DE markets that are important in cluster 4 (compare Tables 11 and 7).

Note that the asymmetric link between ‘ZUR’ and ‘BPZ’ (6:0) is shadowed by the link with ‘ZAG’ (0:6). Thus, if we merged ‘ZAG’ and ‘BPZ then the link with Zurich would be symmetric. Nevertheless, Zurich is a home market destination, albeit a less prominent one than others. The ES market has the highest share in both Lisbon and Madrid, but Lisbon’s could be biased as its home market was not included in the analysis. Disregarding this, Lisbon has a diversified guest mix while Madrid has not and therefore their relationship is asymmetric (3:10). Paris is another home market destination with a one-sided relation to Brussels (12:0). Note that data on the AU market are missing in all years but, judging by the magnitude of its share in other destinations (see Table 7), it does not seem to be important. Paris is comparable to Madrid both in diversification and dominant market share. Zagreb, Budapest, and Brussels are most diversified and therefore have commonalities with more than one more-specialised destination. Thus, the least-specialised destinations (guest-mix clusters) are located in the centre of the map and the highly specialised are pulled out to the periphery.

Hint

The computational steps used in these cluster analyses are implemented in the *R* script named *cluster.R* (see http://www.wu.ac.at/itf/downloads/software/guestmix).

8.5 Longitudinal analysis of guest mix profiles

For the longitudinal analysis of guest mix profiles we use INDSCAL, a multidimensional scaling approach for repeated measurements originally developed by Carroll and Chang (1970). It maps the destinations based on their inter-profile distances into a two- or three-dimensional space (depending on our choice). The changes in distances over time are modelled as weights on the coordinate axes. Though the model has only limited capabilities of accommodating changes in city positions it seems to be a good choice given the stability of the markets we have identified earlier. However, INDSCAL can neither be adapted to handle missing values nor can we simply exclude problematic profiles. We choose the following approach:

(i) We recode problematic data profiles with too many missing values into an uninformative profile (with all values missing).

(ii) Then we compute the distances between all pairs of profiles of a year.

(iii) Finally, missing distances are replaced with the averages across years. Hence, we operate on the level of destination pairs.

Table 13 shows the number of problematic profiles for different thresholds on the number of missing values. Seven turned out to be a good choice as it does not result in erratic weight changes that we observed for higher thresholds. This means that we have to predict a total of 219 out of 120 × 12 (= 15%) missing distances.

The distribution of missing values across years is shown in Table 14. The highest numbers appear in 1998 (35%), 1995, 2004, and 2006 (24%), with Madrid and Zurich being the most affected destinations (see Table 15). Note

Table 13 Cumulative number of destination profiles with 17, 16, … or fewer missing values

17	16	15	11	10	9	8	7	6	5	3	2	1	0
4	5	6	7	8	9	14	15	16	20	26	34	59	192

Table 14 Number of missing distances per year across pairs of destinations

1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
29	15	15	42	15	15	0	15	0	29	15	29

Table 15 Number of missing distances per destination across years

AMS	BER	BRU	BUD	HEL	LIS	LON	MAD	OSL	PAR	PRG	ROM	STO	VIE	ZAG	ZUR
15	15	15	15	15	15	27	55	28	28	43	28	15	15	27	82

that a missing profile leads to a missing distance between the destination that provides no information and any other destination (whether or not its profile is missing too). Hence, in order to not distort the relations between destinations we have to fill in the missing distances with destination-specific predictions.

We determined the number of dimensions to use for the mapping space by visually examining solutions in two and three dimensions. This is feasible as the model fit provides only limited information and determining the 'correct' model (see chapter 3 in Borg and Groenen, 1997) hardly justifies the effort. A three-dimensional solution seems to be appropriate as can be seen from Figure 4. The plane spanned by the first and second dimensions running through the centre of the map is shown as a grid. The axes of this plane are coloured red. Projections of city locations (indicated by text labels) onto this plane are drawn as points which are connected through a line parallel to the third dimension. For example, we find the projections of Zagreb and Helsinki near the first axis and that of London near the second. The map corresponding to this plane looks similar to the one we found with cluster analysis (see Figure 3). Oslo, Helsinki, and Stockholm are separated from the rest; Budapest, Prague, and Zagreb are closest to the centre; Berlin is located at the periphery at some distance from Vienna, and Zurich is now closer to the centre. In the opposite direction we find Rome, Brussels, Amsterdam, London, and Paris. Lisbon is again positioned close to Zagreb, but Madrid is close to Rome instead of being at the periphery of the two-dimensional plane. However, Madrid and Lisbon extend far into the third dimension, in opposite direction to all destinations but Oslo and Stockholm. Thus, Lisbon is in fact not close to Zagreb, nor is Madrid close to Rome, and, as we suspected earlier, Madrid and Lisbon do not have that much in common. This might explain Madrid's position, but be-

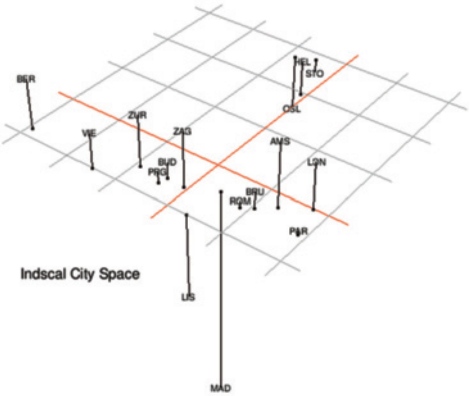

Fig. 4 Time-invariant and unweighted multidimensional scaling solution (1995 to 2006)

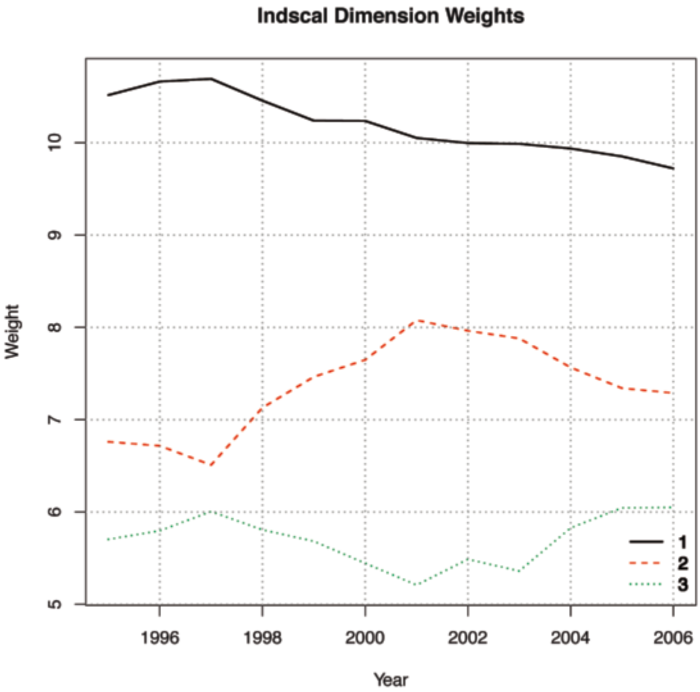

Fig. 5 Development of dimension weights (1995 to 2006)

ing second in terms of missing profiles (see Table 15) and the gap being systematic (see Table 4), biased predictions could be the cause as well. Note that the space in Figure 4 is not yet weighted; on average the weight ratios of axes 1:2:3 are 1.79:1.30:1.

Figure 5 depicts the development of the dimension weights over time. For dimension 1 there is a slight downward trend placing Oslo, Helsinki, and Stockholm closer to the centre. Note that this has no effect on the remaining destinations as they are lined up along the second dimension and, starting with 1997, the differences on this dimension become more pronounced until 2001. Over the same period the weight on the third dimension moves into the opposite direction, the net effect being that Lisbon wanders closer to Zagreb, and Madrid approaches Rome and Brussels. (The differences in the roots of the weights are relevant which are larger for the smaller weights; see Figure 6 for illustration). From 2003 to 2005 part of this development is reversed, with the weight on the third dimension returning to the level of 1997. Thus, we can qualify 1995 to 1997, 2001 to 2003, and 2005 to 2006 as the stable periods.

Table 16 demonstrates how the markets influence the coordinate values on each dimension; correlations that are insignificant at the 5% level are indicated by a left angular bracket. Note that correlations do not imply cause and effect but the mapping solution clearly reflects the data structure. Dimensions 1 and 3 are influenced by home markets such as Finland, Norway, Sweden, and Spain. The signs of the correlations are arbitrary as in total there are six equivalent solutions which can be obtained by changing each dimension’s direction independently.

As Oslo, Helsinki, Stockholm, Madrid and Lisbon are each located in the negative hemisphere of the city space the signs are correct. With increasing market share these destinations move closer to the periphery of the mapping space. The signs are also correct for the second dimension: destinations with a large

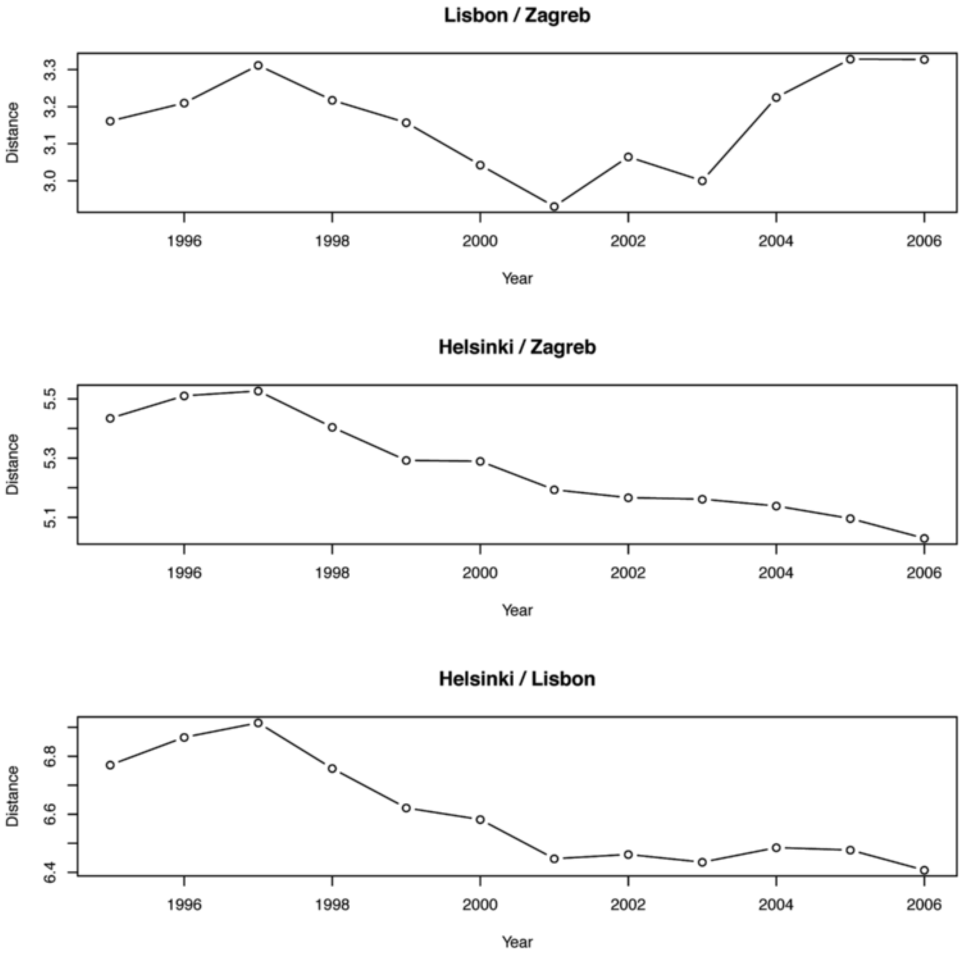

Fig. 6 Development of Euclidean distances between selected pairs of cities (1995 to 2006)

share of the German and/or Austrian market such as Berlin, Vienna, Budapest, Prague, and Zagreb are located in the positive hemisphere. On the negative side we find London, Amsterdam, and Brussels which have a large share of the UK market, or Paris with a large share in its home market France. Rome, Amsterdam, London, and Paris attain the largest shares across the US market which makes this market equally important. The Japanese market is most important for Rome and Paris. The dominance of the Italian market for Rome does not translate into a high correlation as this market is also important for Prague, Budapest, Zagreb, and Vienna (see Table 7). However, as evidenced by the figures for the small shares of the Canadian and Australian markets, correlations do not reflect the magnitudes of the market shares. Finally, on the third dimension, the high negative correlation for the Spanish market reflects the cause of the peripheral positions of Madrid and Lisbon.

For comparison with the previous study for the years 1975 to 1995 we fitted a three-dimensional INDSCAL model to this data. The results are shown in Figures 7 and 8. Note

Table 16 Correlations between market profiles and weighted map coordinates (stacked over time)

	Dimension		
Market	1	2	3
AT	0.36	0.33	0.23
AU	0.23	−0.48	0.37
BE	0.25	−0.31	0.09<
CA	0.37	−0.45	0.37
CH	0.18	0.16	0.20
DE	0.40	0.75	0.31
ES	0.24	−0.18	−0.88
FI	−0.43	0.02<	0.13<
FR	0.29	−0.57	0.01<
GR	0.51	−0.15<	0.18
IT	0.40	−0.24	0.04<
JP	0.28	−0.37	0.01<
NL	0.24	−0.33	0.32
NO	−0.58	0.17	−0.24
SE	−0.67	0.12<	−0.06<
UK	0.15	−0.62	0.34
US	0.41	−0.63	0.24

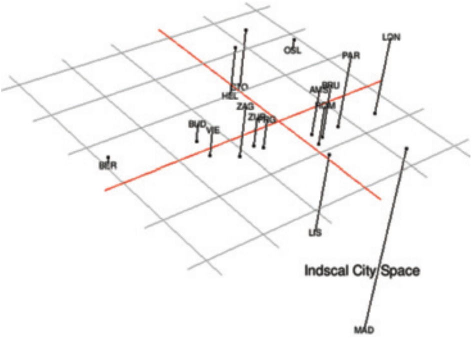

Fig. 7 Time-invariant and unweighted multidimensional scaling solution (1975 to 1995)

that this study used a different distance measure and therefore might not be fully comparable (see Mazanec, 1997). According to the positions of the cities in the map and the correlations between positions and markets (not shown) the axes correspond with the present solution. Overall, the picture is the same but with some differences in the local details for the Scandinavian, central-, and west-European groups of destinations. We do not further elab-

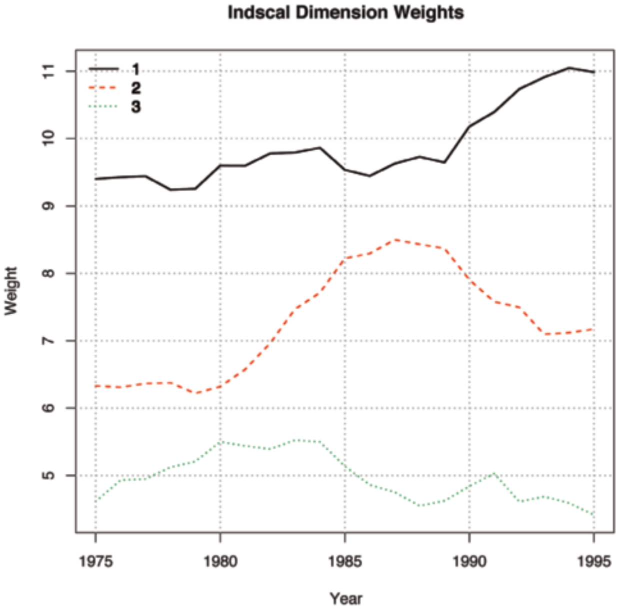

Fig. 8 Development of dimension weights (1975 to 1995)

orate these details as the periods spanned by the data are different and the data definitions do not match exactly. TourMIS provides sparse data for 1975 to 1995 and data definitions have changed. Therefore, we also refrained from fitting a model over the whole 30 year range. Nevertheless, the development of the dimension weights does not seem incompatible at the 1995 boundary and so it seems that dimensions 2 and 3 exhibit a cyclic pattern.

Hint

The computational steps of longitudinal analysis can be found in the *R* script *indscal.R*.

8.6 Conclusions

Applying two methods in parallel would not be worthwhile unless we tried to evaluate their strengths and weaknesses and spot the commonalities in the diagnostic findings. The lessons to learn refer to the substantive output and to the fine-tuning of the analytical tools.

Summarising the results of the cluster analysis we may draw the following conclusions:

(i) The city destinations exhibit a widespread dominance of the home markets.

(ii) The guest mix profiles of a city destination are remarkably stable over time. With the exception of Zagreb we did not observe changes in a destination's cluster membership that are time-related.

(iii) Too many missing values complicate the analysis as they may lead to erroneous conclusions. It is strongly recommended to go back into the raw data for examining their frequency of association with specific cities and/or time periods.

(iv) Cluster analysis manages to single out problematic profiles. In future studies we could help it along by using an information threshold beyond which we should assign a profile to the uninformative group thereby eliminating it from further analysis.

The multi-dimensional scaling analysis led to the following conclusions:

(i) The results gained from cluster analysis were confirmed.

(ii) There are temporary shifts in groups of city positions over time. This was not detected by the cluster analysis.

(iii) The correlations between map positions and markets are plausible. Especially, peripheral map positions are highly correlated with market specialisation and vice versa.

(iv) Filling in missing values on the level of distances did not distort the picture.

(v) The INDSCAL version of the multi-dimensional scaling analysis cannot handle missing values in the raw data and is sensitive to outliers but otherwise is less involved and more straightforward than cluster analysis.

Taking the gaps in the database into account and exerting all necessary caution the diagnosis of inter-city guest-mix similarity is conclusive. If a similar guest mix means anything for judging toughness of competition the cities of Budapest, Prague, and Zagreb, with Vienna and Zurich in their vicinity, must pay attention to each other's marketing strategies. A particularly large domestic market (as for Berlin) alleviates competitive stress. At least, it relaxes the need for taking care of too many generating countries all at once. If you serve a quasi-domestic region (Nordic countries) you may get company but still reap the benefit of a highly familiar marketplace (Helsinki, Oslo, Stockholm).

The guest-mix derived city positions prove to be remarkably stable. Neither changes in the business cycle nor a steady political evolution like the progress in European unification persistently disturb the guest-mix structures. There seems to be pretty strong inertia in the tourist cities' market ties as far as markets are conceived in simple terms of guest nationality.

References

Aldenderfer, M. S., Blashfield, R. K. 1984. *Cluster Analysis*, Series on Quantitative Applications in the Social Sciences, Sage University Paper.

Allison, P. D. 2001. *Missing data*, Sage, Thousand Oaks, CA.

Borg, I., Groenen, P. 1997. *Modern Multidimensional Scaling, Theory and Applications,* Springer, New York.

Carroll, J. D. and J. J. Chang 1970. Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of "Eckart-Young" Decomposition, *Psychometrika*, 35, 283–319.

Caudill, M. 1993. "A Little Knowledge is a Dangerous Thing", *AI Expert*, 8, 16–22.

Hastie, T., Tibshirani, R., Friedman, J. 2001. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer, New York.

Leisch, F. 2006. A Toolbox for K-Centroids Cluster Analysis, *Computational Statistics and Data Analysis*, 51 (2), 526–544.

Lemieux, J. and McAlister, L. 2005. *Handling Missing Values in Marketing Data: A Comparison of Techniques*, MSI Reports 05–107, Marketing Science Institute.

Mazanec, J. A. 1997. A guest mix approach, in: Mazanec, J. A. (ed), *International City Tourism, Analysis and Strategy*, Pinter, London, pp. 131–146.

Mazanec, J. A. 2001. Neural Market Structure Analysis: Novel Topology-Sensitive Methodology, *European Journal of Marketing* 35(7–8), 894–916.

Meyer, D. and Buchta, C. 2008. *R package proxy* – Distance and Similarity Measures, CRAN, http://cran.r-project.org/packages.

Sammon Jr, J.W. 1969. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers* C-18, 401–409.

Schafer, J. L. 1997. *Analysis of incomplete multivariate data*, Chapman and Hall, London.

Tan, P. N., Steinbach, M., Kumar, V. 2006. ***Intro**duction to Data Mining*, Pearson Addison Wesley.

TourMIS, 2008. TourMIS provides free access to Austrian and European tourism statistics, http://tourmis.wu.ac.at, (Site accessed 6 August 2008 and 30 March 2009).