Spectral Comparison of Large Urban Graphs

Ref 039

Sean Hanna
University College London, Bartlett School of Graduate Studies, London, United Kingdom
s.hanna@ucl.ac.uk

Keywords
graphs; graph spectra; comparative analysis; cities; memes

Abstract
The spectrum of an axial graph is proposed as a means for comparison between spaces, particularly for measuring between very large and complex graphs. A number of methods have been used in recent years for comparative analysis within large sets of urban areas, both to investigate properties of specific known types of street network or to propose a taxonomy of urban morphology based on an analytical technique. In many cases, a single or small range of pre-defined, scalar measures such as metric distance, integration, control or clustering coefficient have been used to compare the graphs. While these measures are well understood theoretically, their low dimensionality determines the range of observations that can ultimately be drawn from the data. Spectral analysis consists of a high dimensional vector representing each space, between which metric distance may be measured to indicate the overall difference between two spaces, or subspaces may be extracted to correspond to certain features. It is used for comparison of entire urban graphs, to determine similarities (and differences) in their overall structure.

Results are shown of a comparison of 152 cities distributed around the world. The clustering of cities of similar properties in a high dimensional space is discussed. Principal and nonlinear components of the data set indicate significant correlations in the graph similarities between cities and their proximity to one another, suggesting that cultural features based on location are evident in the city form and that these can be quantified by the proposed method. Results of classification tests show that a cityâ€™s location can be estimated based purely on its form.

The high dimensionality of the spectra is beneficial for its utility in data-mining applications that can draw correlations with other data sets such as land use information. It is shown how further processing by supervised learning allows the extraction of relevant features. A methodological comparison is also drawn with statistical studies that use a strong correlation between human genetic markers and geographical location of populations to derive detailed reconstructions of prehistoric migration. Thus, it is suggested that the method may be utilised for mapping the transfer of cultural memes by measuring comparison between cities.

1. Introduction

Quantifiable links have been suggested between comparative spatial differences as expressed by graph representations and the cultural or geographic differences in which the space is situated, particularly for smaller buildings such as houses (Hillier et al. 1987; Conroy-Dalton and Kirsan 2008). A number of methods have also been used in recent years for comparative analysis within much larger data sets of urban areas, both to investigate properties of specific known types of street network (Peponis et al. 2007) or to propose a taxonomy of urban morphology based on an analytical technique (Figueiredo and Amorim 2007). In many cases, a single or small range of pre-defined, scalar measures such as metric distance, integration, control (Hillier and Hanson, 1984) or clustering coefficients (Watts and Strogatz 1998) have been used to compare the graphs of these spaces, but while these measures are well understood theoretically, their low dimensionality

determines the range of observations that can ultimately be drawn from the data. Comparison of spaces is limited to the scales determined by the chosen measures, and if an indication of cultural or other specific traits is sought these scales may not be the most relevant - the traits in question may be more complex.

Analysis of cities is a specific instance of the wider premise that cultural traits in general are expressed in the artefacts produced and that these traits have a local influence that can be seen to vary over distance. In the fields of genetic and linguistic prehistory, for example, statistical studies by Cavalli-Sforza, Menozzi and Piazza (1994) use a strong correlation between human genetic markers and geographical location of populations to derive detailed reconstructions of prehistoric migration. The measurement of similar comparisons between cities invites comparison with these linguistic and genetic measures, and thus offers a similar mapping for the transfer of cultural memes. Genetic measurements, however, are highly multivariate and are also accessible directly by DNA sampling, whereas evaluation of memes is always based on an interpretation of their effects.

Graph spectra have been used to effectively index, classify and retrieve complex, high dimensional data in pattern recognition and image classification applications (Luo et al. 2003; Robles-Kelly and Hancock 2003), and are suggested here as a means to analyse spaces due to their ability to represent graphs in many dimensions simultaneously, and their derivation directly from the graph. They have been shown as applicable to the representation of axial and similar graphs of fewer than 100 nodes in (Hanna 2007), but are suited to comparative measurement between much larger graphs. To demonstrate and test the method, a very large data set of 152 entire urban graphs will be used, each of a mean size of approximately 6,500 nodes. Similarities (and differences) between the overall structures of the cities they represent will be determined from the graph spectra alone.

An underlying working hypothesis will be required to test the spectral representation. Typically an alternative and established means of measurement is used as a datum in applications of graph comparison; previous experiments on graph spectral distance use either an existing database of samples such as images that can be evaluated by simple visual similarity (Robles-Kelly and Hancock 2003), or generate new samples with an expected variation in distance by parametric or random means (Zhu and Wilson 2005). Conroy-Dalton and Kirsan (2008) use cultural difference to label samples from two local but separate communities in testing the implicit assumption that this correlates to graph distance. A similar method will be used here, in which cultural difference is equated with geographical distance, but with two caveats. First, it is acknowledged that this is a simplification in that a number of important cultural variables are ignored - not least of which is the age of the city, or even distinct neighbourhoods within a city. Unfortunately, data was unavailable for these variables, but this means that results will be necessarily limited and would only be improved by the availability of better data. Second, due to natural geographic avenues of communication, colonisation, political division and other factors it is almost certain that some cultural similarities exist across great distances while some differences exist more locally. In fact it is this sort of cultural pattern, if viewed as a departure from a uniform rate of geographical change, that this method may help illuminate. The experiments here therefore test the hypothesis that there is a correlation between a city's form - as represented by its spectrum - and its geographical location.

2. Spectra of axial graphs

The spectrum of a graph is the ordered set of eigenvalues of its connectivity matrix, and can be used to representing the graph as a single feature vector. This can be recommended for several reasons. First, it remains invariant under all permutations of the original matrix, and is therefore identical for all isomorphic graphs. Additionally, it has been suggested that almost all graphs of a substantial size may be uniquely determined by their spectrum (Van Dam and Haemers 2002). Zhu and Wilson (2005) have exhaustively tested all graphs up to a node count of 20 and determined that the number of cospectral graphs increases up to a node count of between 10 and 15, then declines. This indicates for the node counts of even small axial line graphs, and especially for the

city graphs examined here (up to 79740 nodes), spectra may be considered nearly unique to their graphs. Most importantly, the Euclidian distance between graphs has been shown to correlate highly with differences between graphs generated by random edge changes (Zhu and Wilson 2005) and parametric changes to plans generating the graphs (Hanna 2007).

2.1 Defining the Spectrum for Plan Representation

An unlabeled graph with a set of nodes V and a set of edges E can be represented in matrix form by an adjacency matrix *A*, a $|V| \times |V|$ matrix defined by:

$$A(i,j) = \begin{cases} 1 \text{ if } (i,j) \in E \\ \text{or} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

The spectrum of the graph is found by taking the eigendecomposition of this matrix - eigenvalues λ and eigenvectors φ for *A* are given by solving for

$$A = \Phi \Lambda \Phi^{T} \quad (2)$$

$$\Phi = (\varphi^{1} \mid \varphi^{2} \mid \ldots \mid \varphi^{|V|})$$

$$\Lambda = \text{diag}(\lambda^{1}, \lambda^{2}, \ldots, \lambda^{|V|})$$

$$\{ \lambda^{1}, \lambda^{2}, \ldots, \lambda^{|V|} \}. \quad (3)$$

$$|\lambda^{1}| > |\lambda^{2}| > \ldots > |\lambda^{|V|}|$$

2.2 Assembling the Feature Vector

Graphs to be compared may vary considerably in size (46 to 79740 nodes in the set examined here) but their spectra must be of identical dimensionality to permit measurement of Euclidian distance. In most applications (e.g. Luo et al. 2003; Robles-Kelly and Hancock 2003), values are sorted by absolute magnitude such that $|\lambda^{1}| > |\lambda^{2}| > \ldots > |\lambda^{|V|}|$, and the vector is composed of the first n values:

$$S = (\lambda^{1}, \lambda^{2}, \ldots, \lambda^{n})^{T}. \quad (4)$$

Graphs produced by plan adjacencies (axial or segment graphs of street networks or minimal axial maps of any plan space) are naturally sparse, as connections between nodes can only occur where lines intersect locally. This is beneficial, as Arnoldi iteration may be used to estimate only the largest n eigenvalues required, rather than the whole set for large graphs. In this paper, ARPACK, accessed by Matlab function eigs(), was used to perform the estimate of the largest 100 eigenvalues.

It is essential that the spectral eigenvalues be ordered consistently for all graphs. Sorting by magnitude as mentioned above can be problematic when S contains several values that are of the same magnitude, either positive or negative, and the resulting sort yields a different order for identical graphs. Sorting by actual value, including the sign such that $\lambda^{1} > \lambda^{2} > \ldots > \lambda^{|V|}$, avoids this problem and is the method used here.

3. Comparison of cities by their spectra

Axial graphs were obtained for all 152 cities in the data set. The spectra were produced from these based on adjacency matrices in which each axial line is represented by a graph node, and vertices indicate intersections with other axial lines. To enable measurement in a uniform space, all feature vectors used a dimension of $n=100$ (as in equation 4.)

3.1 Comparison of subgroups

The measurement of distance between graphs facilitates the analysis of assumed cultural subgroups within the data set, to determine both their internal homogeneity and their relative distinction from other groups. This approach is similar to that adopted by Conroy-Dalton and Kirsan (2008) on small graphs and by Cavalli-Sforza, Menozzi and Piazza (1994) on genetic and linguistic data. The basic underlying assumption is that if local cultural traits do manifest themselves in the artefacts studied, samples within a given cultural group will be more homogeneous than those across groups.

A number of criteria may be used to subdivide the set-linguistic, national, topological, etc. As data originally supplied was already labelled by regions roughly corresponding to topological and linguistic divisions, these were used to provide five classes: NOR (English speaking North America: USA only in the data), LAT (Latin America, including Mexico, central and South America), EUR (Europe), ARA (presumably labelled to indicate Arabic speaking countries: actually extending from Arabic speaking Africa through Western Asia including Iran), and ASP (Asia-Pacific, also including New Zealand). The homogeneity or variance of a single group was calculated as the mean difference between all possible pairs of spectra within that group, and its difference from other groups as the mean difference between all spectra in one group and all spectra of the other. In general:

$$(5) \qquad d_{A,\,B} = \sum_{i=1:n} \sum_{j=1:m} (d_{Ai,\,Bj} / (n \times m))$$

where d is measured distance. Because dimensions in the spectra are considered independent of one another, L1 (Manhattan) distance was used instead of L2 (Euclidian).

Table 1 shows this difference between all subgroups in the set. Columns are normalised so that values are relative to the internal variance within each group-i.e. the first column (NOR) indicates that the mean difference between a city in LAT and a city in NOR is 1.41 times that of two cities in NOR; the mean difference between a city in EUR and a city in NOR is 1.45 times that of two cities in NOR; etc. The fact that for the table as a whole, most values off the diagonal are greater than one indicates that the groups are typically more homogeneous internally than their resemblance to other groups. In the most extreme case, a city in ARA is, on average, 4.65 times more similar to another city in ARA than it is to a city in NOR. With the groups listed by rough geographical location from west to east, it is also noticeable that the distinction between groups tends in general to increase with geographic distance. Cities in ARA, for example, are most similar to EUR then ASP, with which they share a land mass, more distinct from LAT and finally the most distant NOR.

	NOR	LAT	EUR	ARA	ASP
NOR	1	1.47	2.94	4.65	2.13
LAT	1.41	1	1.58	2.39	1.18
EUR	1.45	0.81	1	1.53	0.87
ARA	1.21	0.65	0.81	1	0.70
ASP	1.66	0.96	1.37	2.07	1
External mean:	1.50	1.07	0.97	1.39	1.05

Table 1

Relative spectral distinction d between groups

	NOR	LAT	EUR	ARA	ASP
Spectral variance ($d_{A,\,A}$)	1136	1085	559	296	882
Location variance (km)	1931	1742	1358	1202	2931

Table 2

Spectral and location variance within groups

The relative distinction between groups is not entirely consistent however, as several of the values for d fall below 1.0. It is only somewhat surprising that Latin American cities appear more distinct from North American cities than from those in Europe or Asia, but it seems rather more questionable that they are more distinct from one another than they are from foreign cities. The external means in Table 1 represent the relative distinction of each group with respect to all other cities in the set, and while North American cities are on average 1.5 times more distinct from any others, cities in Europe are slightly more similar to cities around the world. Considering the initially somewhat arbitrary group divisions, this is explained by the internal homogeneity or heterogeneity of the groups themselves. Table 2 lists the actual spectral variance dA, A within each group; these are the actual values by which the columns in Table 1 are normalised. It can be seen that the group that shows the greatest values of relative distinction from other groups - ARA - is also the most internally homogeneous, with a variance of only 296. Similarly LAT, with apparently more distinction internally than with respect to several other groups, can be seen to have this because its own internal variance of 1085 is so high. These variances are themselves what would be expected if the geographical spread of each sample is taken into account. The variance of geographical location, expressed as mean distance in km between cities within the group, is given in the lower row of the table and corresponds very highly with spectral variance. The most homogeneous group of cities - ARA - is also the most highly clustered geographically, while as heterogeneity increases, so does the geographic spread of the group. Excluding ASP, which is the only group to span large (but easily navigable) expanses of ocean, the spectral and location variances correlate almost exactly, with a correlation coefficient of 0.98.

As might be expected, the comparison of groups overall shows the closest resemblances between Eurasian groups, particularly between ARA and EUR. LAT shows almost as high a degree of similarity with ARA and EUR, while NOR is both the most heterogeneous and distinct from other groups. A good degree of correlation between city spectra and geographical location, both within and among groups, can thus be seen at this level of detail.

3.2 Individual cities and their geographic location

The relationship between geographic location and spectral similarity can be examined further by examining the data not as subgroups but as individual cities. Although each city is initially represented by a vector in a 100 dimensional space, a number of methods exist to reduce this either to a lower-dimensional space for visualisation or a single scale for analysis. Principal components analysis (PCA) is a linear transformation and is the most commonly recognised, but a number of linear and nonlinear methods were also tested, including factor analysis, kernel PCA, isomap and landmark isomap. All were found to perform similarly, with PCA representing close to the mean, and it is therefore the method described here.

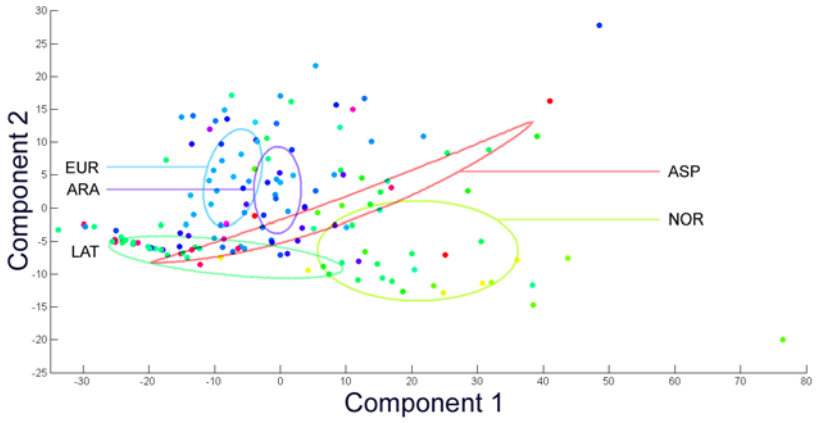

Figure 1

First two principal components of city spectra. Colour indicates longitude.

All cities in the data set are plotted by their first and second principal components in Figure 1; point colour represents the longitude, from yellow (west) through blue to red (east). Most of the identities of the geographical subgroups in the previous section can be made out by their colour: NOR are the yellow and light green points mainly toward the right of the diagram; LAT are the green points to the left; EUR are cyan and blue; ARA are indigo; and ASP are magenta and red. Although considerable overlap is evident, all groups except ASP can be identified with an approximate region in the projected two-dimensional space.

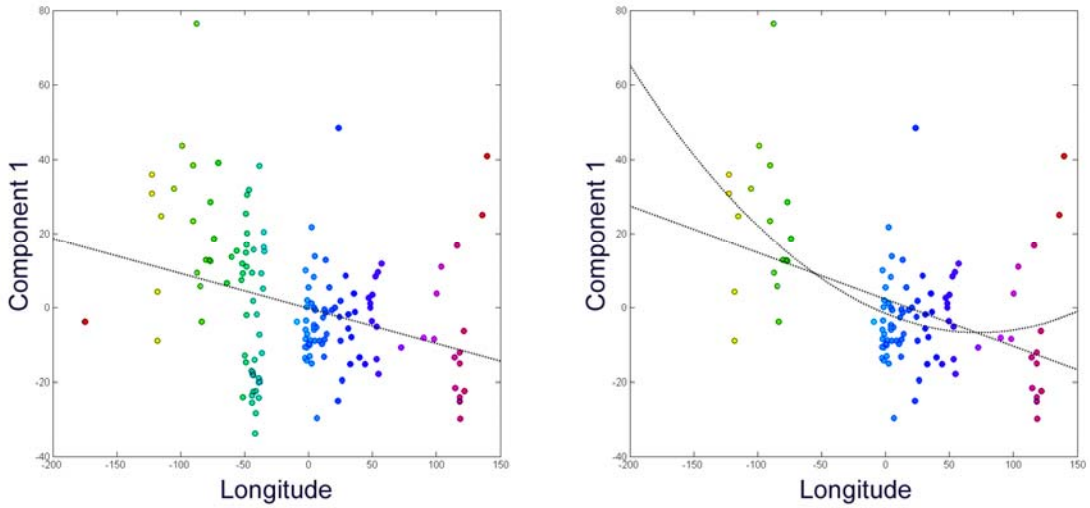

Figure 2
First principal component of city spectra (vertical) against longitude (horizontal). Correlation is low for the entire data set (left), moderate for the northern hemisphere only (right).

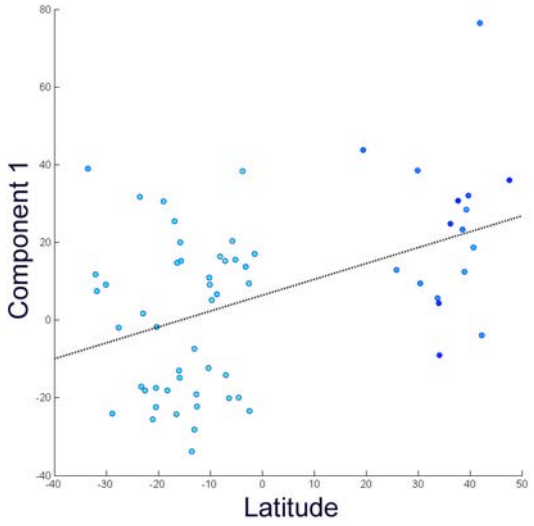

Figure 3
First principal component of city spectra (vertical) against latitude (horizontal). There is a moderate degree of correlation (coefficient 0.44).

Although some information is lost in the remaining 99 dimensions, the first principal component appears to describe a good deal of the variation in longitude and can be used to estimate a lower bound on the degree to which individual city spectra are determined by geographic location. This is measured in Figure 2, which plots the first principal component of all city spectra (vertical)

against their longitude (horizontal). The data set as a whole exhibits a low but not insignificant degree of correlation (0.32) between longitude and spectral component. The low value might be expected for two reasons: as can be seen in Figure 1, the LAT cluster is distinct from EUR and ARA only in the second component; and while longitude provides a reasonable indication of geographic location for most of the set within the northern hemisphere it is useless at indicating the north-south variation of the Americas. Figure 2 (right) plots the same points for the northern hemisphere only. More significant, moderate levels of correlation are evident, with a coefficient of 0.46 to a linear regression, increasing to 0.55 if a quadratic function is used. Japan appears to be a notable outlier in this plot (the upper two red points are Tokyo and Kyoto) - if it too is removed from the set the linear correlation jumps to 0.57.

Latitude is more significant than longitude in the location of cities in North and South America. Figure 3 plots the first component for LAT and NOR against their latitudinal location, also showing a moderate correlation of 0.44.

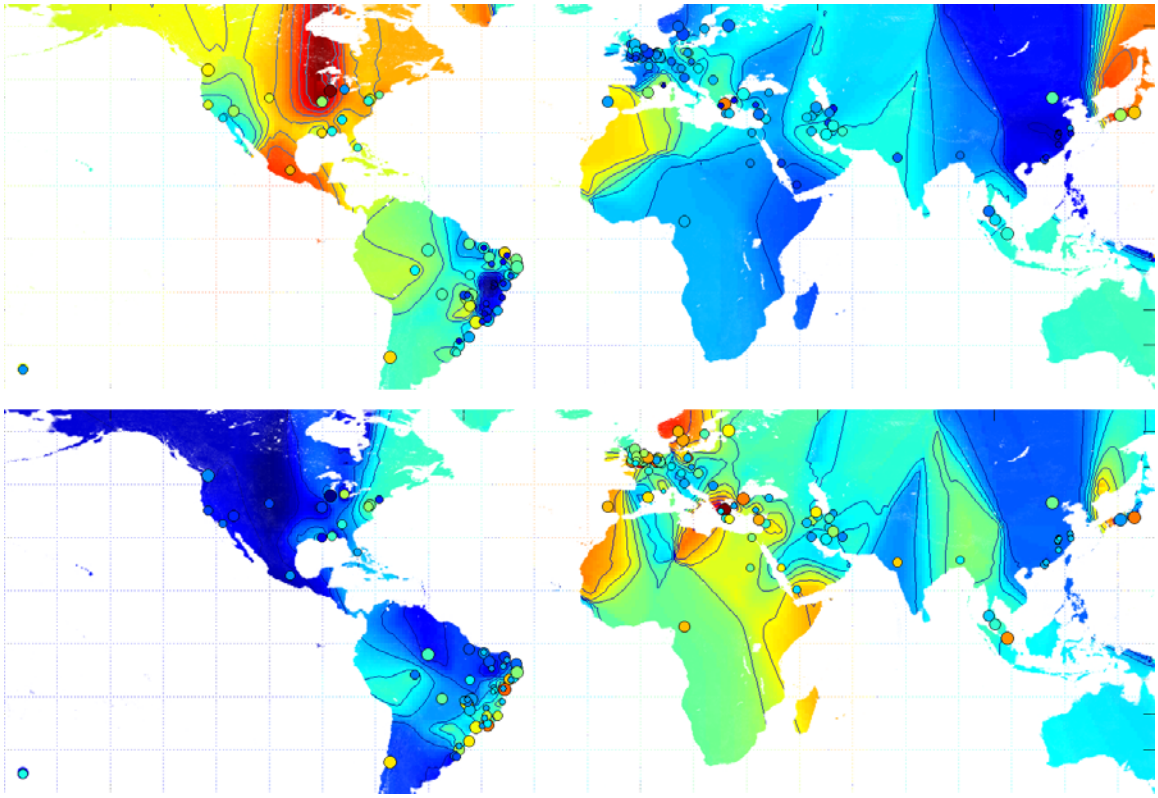

Figure 4
First (top) and second (bottom) principal components of city spectra indicated on the world map. Radius of circles is proportional to the log of the graph size.

Diversity of geographical relationships is clearly more complicated than any single measure of latitude or longitude can provide, and plotting the principal components of spectra against a two dimensional world map gives a better indication of the variance of cities by their location. In Figure 4 a surface is fitted to the distribution of points as an approximation of how the city spectra vary continuously across the world. The method is that used by Cavalli-Sforza, Menozzi and Piazza (1994) to plot gene frequency distributions, the principal components of which have been used to infer patterns of prehistoric settlement and migration. The value of each point on the surface is calculated as a weighted average of the data, with the weight based on relative distance to observed data points as in Shepard (1968). This results in a true interpolation rather than an approximation of points. As with genetic data, the local variance of city samples is rarely completely smooth, so the surface is fit not to actual values but to 'expected' values calculated from neighbours again by Shepard's (1968) formula (Cavalli-Sforza, Menozzi and Piazza 1994, p.

45). This gives a clearer picture of the locally dominant values in the regions with the highest density of samples.

The expected value also provides a means by which a map discrepancy can be calculated for each city individually from the difference between its expected and observed value. Discrepancies were found to approximate a normal distribution with a standard deviation of 18.5. Outliers were determined as cities with a discrepancy greater than 2.5 times the standard deviation (corresponding to the 98.75 normal percentile); seven were found and omitted from the calculation of the surface. Calculation of the gradient was on an orthogonal latitude-longitude grid, and coastlines were ignored as though this gradient were to continue between continents. The primary overall observation is that there is now a marked correlation between map expected values and non-outlier spectra. The first principal component is plotted in Figure 4, and correlates with a coefficient of 0.61 across all regions of the globe, nearly doubling from the linear longitudinal correlation of 0.32. Some of the large regional similarities noted in comparing large regions above (e.g. the similarity between Western Europe and the east coast of South America) as well as finer such details (e.g. the similarity of Mediterranean and Iranian cities) can be seen in the values of the interpolated surface. At the scale of individual cities, trends of resemblance can be seen along isogenic lines across land masses, and a city's similarity or difference from its immediate neighbours. The cities with the greatest discrepancy from their surroundings - the outliers in the set - will be discussed in the following section.

4. Classification and feature extraction

The above analyses assume no prior knowledge of city location or cultural context; they are based purely on spectral relationships alone - first pair-wise distances, then with PCA the maximum dimensions of variance of the set as a whole. However, there is no reason to believe that the dimensions measured necessarily correspond to the most meaningful features for differentiating cities from one another on a geographical or cultural basis. One of the advantages of the high dimensionality of the spectral feature vectors is that any number of arbitrary linear or non-linear subspaces may be taken through this space, separating clusters which would otherwise appear to overlap. Supervised machine learning is used in this section to derive these subspaces.

While the subgroups of the data set were seen to overlap considerably by metric distance alone in Table 1 and by principal components in Figure 2, a support vector machine (SVM, Vapnik 1995) was used to learn the mapping in which a classification between these subgroups can be made. It is thus a test of the possibility of learning the relevant dimensions in which the cultural or geographic classes can be distinguished purely by their city spectra. For each test the SVM was trained on subgroups arranged into two classes (labelled [- 1, 1]). The SVM kernel was a radial basis function with some preliminary optimisation of the parameters (Y and σ^2 a Gaussian) to the portion of the data set used, yielding the range of settings as noted in Table 3. Training was validated using leave-one-out cross validation - the SVM was trained on the set without the city in question, then that city's class determined by the trained SVM. This is the least biased method of validation (Reich and Barai 1999) as each classification by the algorithm is thus made purely on the basis of other cities that are similar in the relevant dimensions.

4.1 Classifying by subgroup divisions

The first test attempted to distinguish European cities from those in the Americas (North and South), drawing the classification on the geographical separation of the Atlantic Ocean and leaving aside for the moment the ARA and ASP portions of the data set. As noted in the discussion of mean distance measures and of principal components, the apparent similarity between the LAT and EUR groups indicate this division may not be the most relevant to the city spectra, but most cities were correctly classified nevertheless, with a validation rate of 83.0% (17% misclassified).

The size of the data (117 cities) used for this run is reasonably comfortable but not conclusive, and access to more city spectra would be desirable. Results were seen to improve by adding the ARA subgroup to EUR such that all cities except for Pacific Asia are classified classifying for the same

(EUR) (NOR+LAT) γ=1.3 σ^2=677 Error: 17.0%	(EUR+ARA) (NOR+LAT) γ=7.8 σ^2=307 Error: 15.8%	(EUR+ARA+ASP) (NOR+LAT) γ=52.5 σ^2=459 Error: 23.7%	(EUR+ARA) (NOR+LAT+ASP) γ=812 σ^2=4718 Error: 19.7%	PCA1 Outliers
ARA				
	Gurgan Kerman	Gurgan Kerman	Gurgan Kerman	
ASP				
			Ahmedabad Auckland Dhaka Hong Kong Phuket Shanghai	
		Chengkan Hongcun Johor Bahru Kyoto Pequim Xidi Yuliang Zhanqi		
EUR				
Athens			Athens	Athens
Barcelona	Barcelona	Barcelona	Barcelona	Barcelona
Gassin	Gassin	Gassin	Gassin	
			Konya	
		Lisbon	Lisbon	
	London	London		
Manchester				
Mytilini	Mytilini	Mytilini	Mytilini	
Nafplion	Nafplion	Nafplion	Nafplion	Nafplion
			Nicosia	
Prague			Prague	
LAT				
Aracaju	Aracaju	Aracaju	Aracaju	
		Brasilia		
				Fortaleza
Maceio	Maceio	Maceio	Maceio	
Porto Alegre	Porto Alegre	Porto Alegre	Porto Alegre	
Rio de Janeiro	Rio de Janeiro	Rio de Janeiro	Rio de Janeiro	
Salvador	Salvador	Salvador	Salvador	
	Sao Luis	Sao Luis	Sao Luis	
				Sao Paulo
Florianopolis	Florianopolis	Florianopolis	Florianopolis	
Vitoria	Vitoria	Vitoria		
		Alcantara		
Cidade de Goias	Cidade de Goias	Cidade de Goias	Cidade de Goias	
		Diamantina		
		Mucuge		
Ouro Preto	Ouro Preto	Ouro Preto	Ouro Preto	
Penedo	Penedo	Penedo		
Petropolis	Petropolis	Petropolis	Petropolis	
		Pirenopolis		
NOR				
Ann Arbor	Ann Arbor	Ann Arbor	Ann Arbor	Ann Arbor
Atlanta	Atlanta	Atlanta	Atlanta	
				Chicago
		Washington	Washington	

Table 3

Errors in classification by SVM, and outliers from the first principal component map.

Atlantic division. With a combined set of 133 cities, correct validations rose slightly to 84.2%. A further classification was made of the entire data set, again using the oceans as a natural boundary and grouping together the landmasses - Americas in one class, Afro-Eurasia in the other. Although the overall number of samples is greater (by 19), in taking a geographically and culturally diverse group including EUR, ARA and ASP to be a unity it should not be expected that classification will improve. The validation error did increase, primarily in misclassifying cities in ASP, but classification is still reasonably successful with 76.3% correct. An alternative placement of ASP may be made in linking it to the Americas via the Pacific Ocean. In classifying the entire data set as two groups of (EUR + ARA) and (NOR + LAT + ASP) validation showed an improvement to 80.3% correctly classified.

4.2 Individual cities and classification errors

More detail about the relationship of individual cities to their allotted class can be gained from the list of those misclassified over the various runs of the SVM. Table 3 lists the classification errors for the four divisions described above. As mentioned, between columns one and two the classification error decreased with the addition of ARA to the EUR set. While two of the newly added cities were misclassified in the second run, the improvement is exclusively in the correct classification of two cities - Athens and Prague. As being classed with a given group suggests a similarity with that group, this indicates that these bear a meaningful resemblance to the ARA group. It is possibly significant that these are two relatively eastern cities in Europe, as they are later misclassified again (column 4) with the far eastern cities when ASP is added to the Americas class. London, by contrast, shows precisely the opposite pattern - it is correctly classified in the two cases when its class is most clearly biased toward western cities (columns 1 and 4) but grouped with the Americas when the Afro-Eurasian class is more predominantly eastern.

The entire set of errors within ASP when the Asia-Pacific cities are added to the data set are mutually exclusive - not one city is misclassified in both instances. As with the EUR city errors above, this is also likely an indication of the resemblance of these to other subgroups - the eight cities in column 3 are more like the Americas, the six in column 4 more like Eurasia.

Where the fluctuation of a classification error as classes are recomposed indicates a resemblance to the geographic subgroup of cities relabelled, a persistent misclassification suggests a city is unlike others within its own group. Other than the cities mentioned, most errors are repeated in all runs of the SVM, indicating that these cities are outliers, at least with respect to the Atlantic boundary used in these divisions. This is corroborated to some degree in column 5, which lists the seven cities previously labelled as outliers in mapping the first principal component to geography, as 4 of these correspond with cities repeatedly misclassified by SVM. The three European outliers (Athens, Barcelona and Nafplion) are confirmed as consistently atypical of their subgroup. By contrast, the fact that the two Brazilian outliers in LAT are never misclassified would suggest that these differ from their neighbours only in ways that resemble the cities of North America where much of the LAT set does not. This is in agreement with the previously noted evidence of metric distance and PCA that suggests the (NOR + LAT) grouping is not as evident as longitude and landmass would suggest.

5. Conclusions

The comparison of world cities has been used as a demonstration and test of the method of graph representation. This was based on the premise that a city's form is partially dependent on cultural factors that are transmitted locally, and therefore tested the graph spectral representation by measuring the correlation between city's geographical location and its spectrum. Although this premise is a risky simplification in that many factors other than location are certainly significant in influencing spatial morphology, the spectra were nevertheless able to demonstrate a relationship. Metric distance between groups of spectra provided a means by which internal homogeneity within - or distinction between - groups of cities were measured, and this appears to correspond with relative physical distance and geographic spread. Principal and nonlinear components of the data set indicate a correlation between graph structure as revealed by spectra and geographical location - first longitude, then the global surface - suggesting that cultural similarities due to proximity are well captured by the technique.

The high dimensionality of the resulting vector allows more detailed comparison than any single, scalar dimension. Further processing by SVM or a similar supervised learning algorithm allows the extraction of a number of arbitrary features, thus knowledge of context can be used to refine the subspace so that particular features may be sought in the data. If data is available, factors such as the age of a city would be easily used, just as geographical location has been, to refine classifications or to better understand their statistical effect on spatial form. By the same token, the high dimensionality is potentially beneficial for its utility in data-mining applications that can draw correlations with other data sets such as land use information.

The method of characterising graphs by their spectra is easily automated. Spectra are invariant to the ordering of nodes and the process from map to spectrum, as well as subsequent analyses, is mathematically well defined. As no human effort is required there is no risk of human error and the size of the data set may be very large. It is also applicable to large graphs. In this paper a maximum size of nearly 80,000 nodes was used with a computation time of 85 seconds to calculate the spectral vector, but spectra of segment maps of up to 263,215 nodes have easily been produced by the same method with modest computational expense.

References

Cavalli-Sforza LL, Menozzi P and Piazza A. 1994. *The history and geography of human genes.* Princeton: Princeton University Press.

Conroy-Dalton R and Kirsan C. 2008. Small graph matching and building genotypes. *Environment and Planning B: Planning and Design*, 35 (5): 810–830.

Figueiredo L and Amorim L. 2007. Decoding the urban grid: or why cities are neither trees nor perfect grids, *Proceedings, 6th International Space Syntax Symposium, Istanbul 2007*.

Hanna S. 2007. Representation and generation of plans using graph spectra, *Proceedings, 6th International Space Syntax Symposium, Istanbul 2007*.

Hillier B and Hanson J. 1984. *The Social Logic of Space*. Cambridge University Press.

Hillier B, Hanson J and Graham H. 1987. Ideas are in Things: an Application of the Space Syntax Method to Discovering House Genotypes, *Environment and Planning: Planning and Design*, 14: 363–385.

Luo B, Wilson RC and Hancock ER. 2003. Spectral embedding of graphs, *Pattern Recognition,* 36: 2213–2233

Peponis J, Allen D, Haynie D, Scoppa M and Zhang Z. 2007. Measuring the configuration of street networks: the Spatial profiles of 118 urban areas in the 12 most populated metropolitan regions in the US, *Proceedings, 6th International Space Syntax Symposium, Istanbul 2007*.

Reich, Y and Barai, SV. 1999. Evaluating machine learning models for engineering problems. *Artificial Intelligence in Engineering* 13 (1999): 257–272.

Robles-Kelly A and Hancock, ER. 2003. Edit Distance From Graph Spectra. Proceedings of the ninth IEEE International Conference on Computer Vision (ICCV 2003)

Shepard D. 1968. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference* (August 27 - 29, 1968). ACM, New York, NY: 517–524.

Van Dam ER and Haemers WH. 2002. Spectral Characterizations of Some Distance-Regular Graphs, *J. Algebraic Combin.* 15, pp. 189–202.

Vapnik V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Watts DJ and Strogatz SH. 1998. Collective dynamics of 'small-world' networks, *Nature*, 393: 440–442.

Zhu P and Wilson RC. 2005. A Study of Graph Spectra for Comparing Graphs. *British Machine Vision Conference 2005.*