doi: 10.1111/j.1365-246X.2008.03997.x

Delineation of shallow seismic source zones using *K*-means cluster analysis, with application to the Aegean region

Graeme Weatherill and Paul W. Burton

Seismic Risk Group, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, *UK. E-mail: G.Weatherill@uea.ac.uk*

Accepted 2008 October 13. Received 2008 September 29; in original form 2008 February 14

GJI Seismology

SUMMARY

The selection of specific uniform seismic source zones for use in probabilistic seismic hazard analysis is often controversial. Recognizing that a consistent approach to source model development is not always possible, as the information available relating to geology and seismotectonics can vary from region to region, the *K*-means algorithm for hierarchical cluster analysis can be used to partition regions based on observed seismicity. The Aegean [incorporating Greece, Albania, Former Yugoslav Republic of Macedonia (F.Y.R.O.M.), southern Bulgaria and western Turkey], with its varied seismotectonics and generally high seismicity, is used as an important area of seismicity in which to develop and demonstrate the application of *K*-means. Two types of algorithm are considered. The first is a point-source *K*-means that can be used to partition a catalogue of earthquake hypocentres. The second is a novel line-source development of the algorithm, appropriate in seismology as these are analogues for the traces of active faults, which is then applied to a catalogue of known fault ruptures in the Aegean. The common problems of the *K*-means methodology are also addressed. Ensemble analyses are used to identify better choices of initial estimates for the cluster centres. A cluster quality index is used to identify the optimum number of clusters, and its robustness assessed when considering different subsets of the observed earthquake catalogue. An alternative approach is also implemented: Monte Carlo seismic hazard analysis is used to compare models with different numbers of clusters with the observed seismicity of the 20th century. Considerable variation is found in the optimum number of clusters identified either by the quality index or by stochastic seismic hazard analysis. Ultimately the *K*-means partitions of seismicity are developed into source models and their representation of Aegean seismotectonics assessed. The result is that models containing between 20 and 30 clusters emerge as the most appropriate in capturing the spatial variation in hypocentral distribution and fault type in the Aegean.

Key words: Persistence, memory, correlations, clustering; Spatial analysis; Seismicity and tectonics; Europe.

1 INTRODUCTION

The issue of seismic source delineation is often a controversial one in the practise of seismic hazard analysis, both deterministic and probabilistic. There can be substantial disparity in the way in which seismic sources are characterized. This depends, not immaterially, on the varying levels of knowledge of the seismotectonics of different regions across the globe. It has been common practice since the development of probabilistic seismic hazard analysis (PSHA) by Cornell (1968) and McGuire (1976), to utilize areal source zones of seismic homogeneity (Kramer 1996; Abrahamson 2006). This is often the case when the objective of the analysis is to produce a seismic hazard map. The shortcomings of areal zones are well recognized (Bender & Perkins 1987). It has often become an objective of PSHA, in many countries, to develop a sufficient knowledge of the fault systems within a region so as to move away from uniform areal zones. Despite the rapid expansion of knowledge of earthquake sources across the globe, however, there are surprisingly few places where this approach is applied. Although the arguments against areal source zoning are abundant, in many locations there are few other options.

Of the many criticisms against the use of uniform seismic source zones, one of the most outstanding is the absence of any recognized formal and consistent procedure for developing and evaluating zone models. It is common, even on extensive seismic hazard assessment projects, to make a choice of source models based on expert opinion. Whilst this has the clear benefit of allowing information from different sources (geology, seismotectonics, geodesy, etc.) to be incorporated into the model, the output can be subjective and lack transparency to the user. Where uncertainty exists in the applicability of a particular source model, different models may be incorporated into a logic tree analysis (Grünthal & Wahlström

2001; Barani *et al.* 2007). Consistency between different source zone models for a particular region can be hard to achieve in the absence of well-constrained seismotectonic information (fault location, focal mechanism, slip rate, etc.). In some locations absence of such information may be due to inadequate investigation of the sources. Elsewhere, and particularly in intraplate and low seismicity regions, geophysical investigation of the seismic source may simply not reveal enough information about the seismic source. Where physical parameters of seismic sources are not characterized, a substantial seismic hazard may still exist. In such circumstances it is the distribution of observed and historical seismicity that reveals the most about the nature of the seismic source.

If development of a uniform seismic source model incorporating physical faulting characteristics is often controversial, then source zonation using hypocentral distribution exclusively can be even more so. Decisions regarding the boundaries between groups of earthquakes can be highly subjective and may often appear arbitrary, even if great consideration has been given to them. As noted by Beauval *et al.* (2006), for regions where faulting is not well-characterized, 'different experts often provide very different maps that characterize somewhat different zonation schemes, based on their differing interpretation of the meagre data that exist.' Where some consistency between models is found, there may be a common factor that influences the zonation. This may arise not from the true variation of seismicity in a region, but from variability in seismic network coverage and completeness (Papazachos 1990). Such partitions may be useful for analysing variation in parameters of seismic behaviour (e.g. *b*-value, $M_{\max}$). Delineation of zones of seismic homogeneity on such a basis, however, is an approach that should be treated with caution.

Given the shortcomings, including an often perceived subjectivity of the commonly used approach to seismic source zonation, especially in areas where hypocentral distribution is the only indication of earthquake behaviour, then it is not unreasonable to search for an alternative objective, replicable and transparent approach. What is meant by this is an approach that would produce an ostensibly similar partition (not necessarily identical, though this is ultimately desirable as data sets become more informative) of a set of earthquake hypocentres. By expressing this as a partitioning problem, it becomes possible to address the issue of source zonation with the use of hierarchical cluster analysis techniques, in particular the *K*-means algorithm (Hartigan 1975).

To develop the techniques of cluster analysis in ways applicable to seismicity and seismic zoning, the Aegean region shall be used both as a case study and significant target area in itself. Aegea contains the highest seismicity in Europe, and is of particular interest as it encompasses areas of seismic activity where active faults are well-characterized, as well as those where the seismic sources are poorly defined. The study area extends from 18° to 32°E and 33° to 43°N, which includes all of Greece and Albania, Western Turkey and parts of the Former Yugoslav Republic of Macedonia (F.Y.R.O.M.) and Bulgaria. Of particular note in this region is the disparity in fault models across national borders, and even within countries themselves. This is despite being an area of intensely studied high seismicity, with many carefully geologically mapped visible fault ruptures and extensive earthquake catalogues, both old and new. We use the Aegean region, largely because it encompasses areas where active faults are well-defined (e.g. the Gulf of Corinth or the Sea of Marmara), as well as areas where seismicity is lower but with sporadic large events (e.g. northern Greece, southern Balkans). This allows for comparison of the partitions with the distribution of recognized active sources, as well as the application to areas that may not be adequately captured by existing source models.

The seismotectonic situation of Greece and its surrounding area is highly varied. With the exception of the trace of the North Anatolian fault east of the Sea of Marmara, many of the most active seismogenic faults are located offshore (Danciu & Tselentis 2007). However, damaging earthquakes have been known to occur away from the main plate boundaries in areas that have been classed as low seismicity. A good example of this is the 1995 Kozani-Grevena earthquake (6.5 $M_{\rm w}$), which occurred in such an area (Stiros 1998).

Previous seismic source models for the Aegean have been developed by Hatzidimitriou *et al.* (1985) [HZ1985], Papazachos (1990) [PP1990] and Papaioannou & Papazachos (2000) [PP2000]. Furthermore, seismic zones covering Greece and the southern Balkans can be found in the Turkish seismic source model of Erdik *et al.* (1999) [ED1999]. An additional set of areal sources can be seen in Jimenez *et al.* (2001) [JI2001], as part of a unified source model for the Mediterranean region. This model was constructed by geometrical homogenization of the ED1999 and the PP2000 models. Refer to Jimenez *et al.* (2001) for more details regarding this procedure. The evolution of the Aegean source models form HZ1985 to PP2000 illustrates the adaptation of the source models as new information becomes available.

2 THE *K*-MEANS METHODOLOGY

K-means cluster analysis (Hartigan 1975; Hartigan & Wong 1979; Jain *et al*. 1999) is an example of a hard partitioning algorithm. A set of *N* data ($x_1, x_{2,} \ldots, x_N$) in *d* dimensions is partitioned into *K* clusters, where each element in the data set is allocated entirely to a particular cluster. It is an iterative process whereby the data are initially partitioned, the mean position of each group calculated, and then the data partitioned again by allocating each datum to its nearest mean cluster position. The procedure terminates when no datum changes cluster or when the number of iterations reaches a pre-defined maximum (usually 100 iterations, as is the case here). The algorithm is described by the flowchart in Fig. 1. In this research the *K*-means algorithm used is a modified version of the *K*-means code provided by Nabney (2002), implemented in Matlab 7.1.

In most applications of *K*-means clustering the preferred distance metric is the Euclidean square distance (Hartigan & Wong 1979). Other types of distance metric may be used instead, but for many practical applications of *K*-means cluster analysis the Euclidean-square metric is still the most common (Jain *et al.* 1999). Theoretically there is no upper limit to the number of dimensions of data that can be clustered using *K*-means, though we shall only be considering 2- and 3-D spatially distributed data here. Since the objective of this application is to partition a set of data distributed across a fixed crustal volume, more complicated distance metrics may not be appropriate. They also offer little compensation for the additional computation required. When using Euclidean-square distance as a metric, however, there is a risk that the largest scaled features will dominate the clustering process. This is usually overcome by normalization of the data or by implementing weighting schemes to compensate for dominance of one particular dimension. In the study of seismicity, an acceptable partition or set of clusters should provide viable seismic zones for seismic hazard quantification and analysis. This is a further long term aim.

Assessment of the 'quality' of a partition is an important consideration in cluster analysis. The most common measure of cluster quality (for known *K*) is the total within-cluster sum of squares

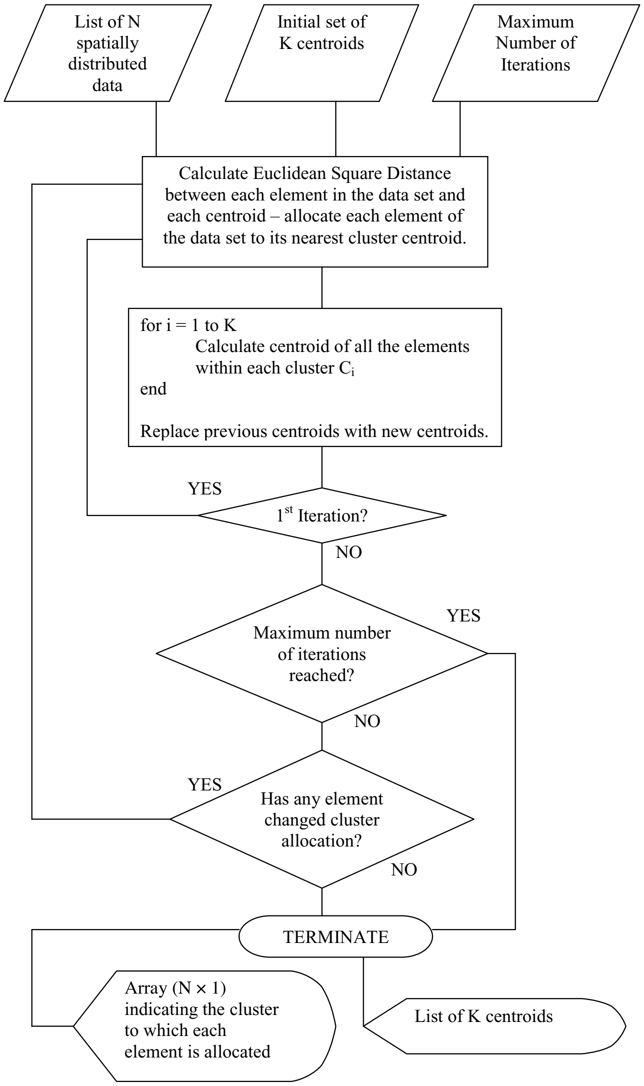
List of N spatially distributed data
Initial set of K centroids
Maximum Number of Iterations
Calculate Euclidean Square Distance between each element in the data set and each centroid – allocate each element of the data set to its nearest cluster centroid.
for i = 1 to K
Calculate centroid of all the elements within each cluster C$_i$
end
Replace previous centroids with new centroids.
YES
1st Iteration?
NO
YES
Maximum number of iterations reached?
NO
YES
Has any element changed cluster allocation?
NO
TERMINATE
Array (N × 1) indicating the cluster to which each element is allocated
List of K centroids

Figure 1. Flowchart describing the *K*-means algorithm.

(TWCSS), also referred to as squared error or clustering error (Likas *et al.* 2003). The TWCSS is defined as:

$$TWCSS = \sum_{i=1}^{N}\sum_{k=1}^{K} I\left(x_i \in C_k\right)\left\|x_i - m_k\right\|^2 , \tag{1}$$

where m_k is the mean of cluster C_k and I(X) is 1 if statement X is true, 0 otherwise. An alternative to this particular measure is the pooled within-cluster sum of square distances (WK). This metric is defined by Tibshirani *et al.* (2001) as:

$$D_k = \sum_{i,i' \in C_k} d_{ii'} \tag{2a}$$

$$WK = \sum_{k=1}^{K} \frac{1}{2n_k} D_k , \tag{2b}$$

where n_k is the number of elements within cluster C_k and d_{ii}' is the Euclidean-square distance between point x_i and x_i'. For a cluster of n_k elements, TWCSS and WK can be visualized in the manner of Fig. 2.

Though the *K*-means algorithm is popular in several branches of science (e.g. image processing, pattern recognition and genetics), it suffers from two major problems that, despite the development of research into the procedure, remain unresolved. The first problem is that of determining the optimal number of clusters in a set of data. This is a classic issue in cluster analysis and many different approaches have been suggested in attempts to solve it (Tibshirani *et al.* 2001; Feng & Hamerly 2006, and references within). These approaches can be subdivided into three different classes: validity indices that do not correlate with *K* (e.g. Krzanowski and Lai 1988; Kaufmann & Roeusseuw 1990; Tibshirani *et al.* 2001); cluster splitting depending on a failure criterion (Hamerly & Elkan 2003; Feng & Hamerly 2006; Welling & Kurihara 2006); genetic programming methods with fixed (Krishna & Murty 1999) or variable *K* (Sheng & Liu 2006). The cluster splitting methods tend to work well if the clusters are well-separated. They also tend to assume that clusters are ostensibly Gaussian in nature, which is not always the case. Where clusters may appear to overlap, the optimum *K* estimated is strongly dependent on the statistical test used and the probability

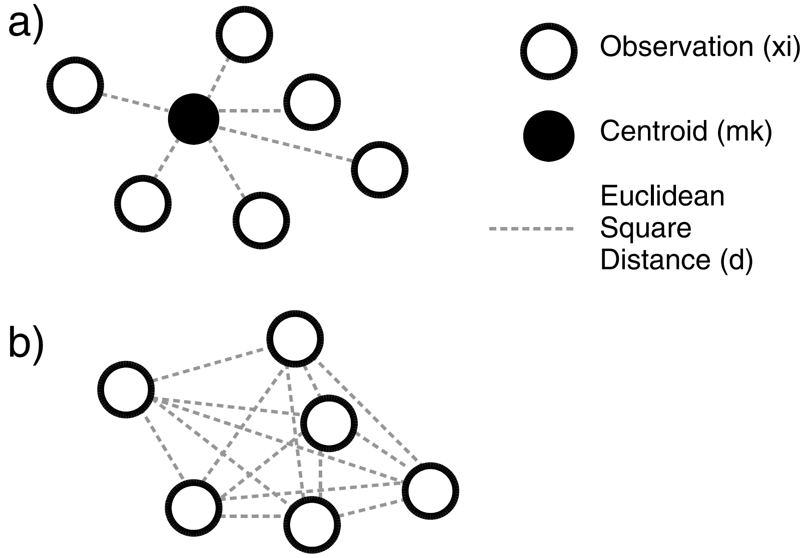

Figure 2. Visualization of within-cluster quantity indices: (A) total within cluster sum of squares, (B) pooled within-cluster sum of square distances.

level at which the test is passed (Feng & Hamerly 2006). Similarly, validity indices may vary in the appropriateness for the data set, with some performing better than others depending on whether the clusters are well-separated. Ultimately, the choice of K may still be an expert decision based on the output from different indices and methods.

The second problem inherent in the K-means algorithm is the sensitivity of the partition to the choice of initial cluster centres. K-means is an example of a hill-climbing algorithm that may stabilize at a local optimum rather than the global optimum partition. Many of the approaches to overcome this employ stochastic techniques to identify the optimum set of initial cluster centres. Alternatives to stochastic techniques tend to require either exhaustive searches for the optimum initial seeds (Likas *et al*. 2003), which can be computationally impractical even on high performance computers, or a change in distance metric or definition of the cluster centre (Zhang *et al.* 1999).

The simplest approach to identify the optimum initial clusters is an ensemble analysis (Peña *et al*. 1999; Kuncheva & Vetrov 2006). The algorithm is repeated a large number of times, with different initial seeds (randomly selected) on each trial. The partition that produces the smallest TWCSS is the optimum. As is common in stochastic search procedures, this optimum may still only be a local rather than global optimum. However, if the size of the ensemble is large enough, it may be sufficiently close to the global optimum for practical purposes. Krishna & Murty (1999) and Lu *et al.* (2004) extended the stochastic approach to a Genetic K-means algorithm, which converges toward the global optimum partition using the evolutionary computing approach. This tests a large number of random solutions and selects the better fitting ones as a basis for a new population of solutions.

3 EARTHQUAKE CATALOGUES

The earthquake catalogue used in this analysis is that of Burton *et al.* (2004). This catalogue consists of 5198 earthquakes with $M_w \geq 4.0$ for the period 1900–1999 and region 18°–31°E and 33°–43° N. Only shallow earthquakes, with focal depths less than 60 km, are considered here. Where no focal depth has been determined 15 km is assumed, approximately equal to the mean depth of shallow earthquakes in the Aegean region. The completeness magnitude M_c for the entire period covered by this catalogue is 5.2 M_w, which is consistent for the period 1900–1963. From 1964 onward M_c drops to 4.8 M_w.

To supplement this catalogue two more catalogues are used. The first is the catalogue of Papazachos & Papazachou (1997), which contains 427 earthquakes for the period 550 B.C.–1899 (the combined catalogues can be seen in Fig. 3). This catalogue is constructed

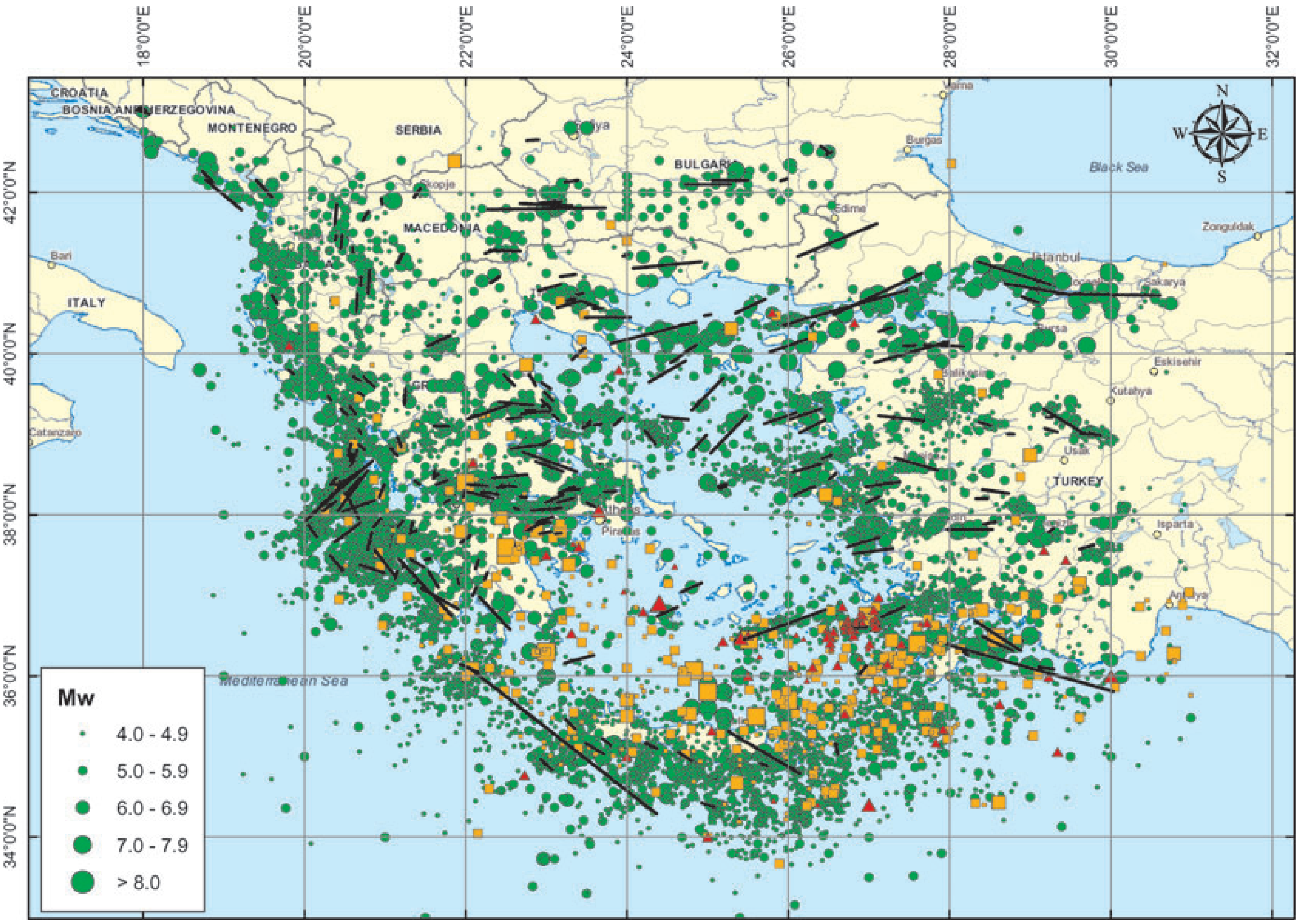

Figure 3. Earthquakes in the Aegean region [Burton *et al.* (2004) and Papazachos and Papazachou (1997) catalogues combined]. Shallow earthquakes (Depth ≤ 60 km) indicated by green circles, intermediate depth earthquakes (60 < Depth (km) ≤ 120 km) by orange squares and deep earthquakes (Depth > 120 km) by red triangles. Black lines indicate the location and strike of the ruptures in the Papazachos *et al.* (1999) catalogue, with lengths calculated using eq. (3) (Wells & Coppersmith 1994).

with M_w as the preferred magnitude unit. The second additional catalogue is taken from the National Observatory of Athens and spans the time period 2000 A.D.– 2005 A.D. This adds a further 3698 events with $M_w \geq 4.0$. Criteria for homogenizing the National Observatory of Athens data with the remainder of the catalogue are found in Burton *et al.* (2004), and were followed for this subsequent period.

The presence of aftershocks may be very influential in cluster analysis; hence comparison needs to be made between the original catalogues and those that have been purged of aftershocks using a catalogue declustering algorithm. The preferred algorithm for removal of aftershocks is that of Reasenberg (1985).

Also considered in this analysis is a catalogue of known ruptures for historical earthquakes in the Aegean (Papazachos *et al.* 1999). The ruptures in this catalogue are expressed in terms of their source parameters, including strike, dip and rake. These earthquakes are all assumed to be shallow events whose rupture planes have been distinguished using macroseismic intensities, surface observations or aftershock distribution. The catalogue found in Papazachos *et al.* (1999) contains 150 events up to 1995.This catalogue is cut-off at 6.0 M_w. Using the atlas of isoseismals for Greece (Papazachos *et al.* 1997), a further 64 fault ruptures are identified for earthquakes with magnitudes in the range $5.5 \leq M_w \leq 6.0$. For all but a few of the events in this catalogue the exact shapes of these faults remain largely unknown. An approximation to a line source is therefore made. The orientation of the lines is indicated by strike and the lengths calculated using the empirical fault scaling relation of Wells & Coppersmith (1994). The equation used is the 50th percentile of the regression of the common logarithm of subsurface rupture length against M_w for earthquakes of all fault types:

$$\log(RLD) = (0.59M_w - 2.44) \pm 0.16. \tag{3}$$

This catalogue is supplemented further with an additional eight ruptures from significant earthquakes in the period 1996–2006: Dodecanese Islands, 1996 July 20 (6.2 M_w); Southern Greece, 1997 October 13 (6.4 M_w); Southern Ionian Islands, 1997 November 18 (6.6 M_w); Izmit, 1999 August 17 (7.6 M_w); Athens, 1999 September 7 (5.9 M_w); Skyros, 2001 July 26 (6.4 M_w); Lefkada, 2003 August 14 (6.2 M_w) and Kythera, 2006 January 8 (6.7 M_w). The fault dimensions (i.e. strike, dip, rake and subsurface rupture length) have been determined for four of these events: Izmit (Barka 1999; Reilinger *et al.* 2000; Wright *et al.* 2001; Barka *et al.* 2002), Athens (Papadopoulos *et al.* 2000; Pavlidis *et al.* 2002), Skyros (Papadopoulos *et al.* 2002) and Lefkada (Papadopoulos *et al.* 2003). Faulting parameters for the Dodecanese Islands, Southern Greece and Southern Ionian Islands events were taken from the Global Centroid Moment Tensor Database (www.globalcmt.org/CMTsearch.html), with subsurface rupture lengths estimated using eq. (3). Both the catalogue of ruptures and of hypocentres can be seen in Fig. 3.

The Kythira event is the only intermediate depth event included in the rupture catalogue. The initial focal depth was estimated to be approximately 60 km. Subsequent moment tensor inversions suggest a range of between 55 and 65 km, with aftershocks possibly as shallow as 45 km (Konstantinou *et al.* 2006). The depth cut-off used for the Aegean earthquake catalogue is 60 km, clearly making the Kythira event a borderline case for inclusion. It has been included here for two reasons. First, given the magnitude of this event and the depth range of the aftershocks it is clear that the rupture associated with this event penetrates well into the seismogenic crustal depth assumed previously (60 km). Secondly, although an intermediate depth event, damage from this event was significant and widespread, with an epicentral intensity of VIII on the Modified Mercalli Scale (Konstantinou *et al.* 2006). Clearly, this event is significant for hazard analysis in the Aegean region; the context in which the *K*-means algorithm is applied.

4 APPLICATION TO AEGEAN SEISMICITY

4.1 Application to hypocentres

K-means cluster analysis can be applied to the hypocentral distribution of observed earthquakes in any region. There are some important considerations when using earthquakes, which may not necessarily apply in other applications of *K*-means analysis. The most obvious concern is that of the characterization of an earthquake source. Cluster analysis techniques are typically applied to data characterized as point source. It shall be shown how this can be overcome as *K*-means is extended to apply to line data. This is appropriate because the assumption of earthquakes being a point source is not valid over the entire magnitude range considered in most earthquake catalogues. The point-source approximation may be valid and necessary, though undesirable, for earthquakes with $M_w < 6.0$. For large earthquakes whose rupture lengths are of the orders of tens to hundreds of kilometres, a point-source approximation is unrealistic.

To extend the *K*-means algorithm beyond the original assumption of identical point sources, it is necessary to introduce modifications that would incorporate information regarding the physical parameters of the earthquakes being partitioned. The most obvious feature to consider, and one that is of particular significance in seismic source modelling, is earthquake size. Therefore, a modification to the standard *K*-means algorithm is to use the weighted centroid of the earthquakes within a cluster in place of the mean. The modification is simple. When clustering in Euclidean space, and using Euclidean square distance as the metric, the mean of the N data points (x_i) in the cluster is calculated by:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N}. \tag{4}$$

However, if each point is associated with a weighting (W_i), the centroid becomes:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{N} \mathbf{w}_i \mathbf{x}_i}{\sum_{i=1}^{N} \mathbf{w}_i}. \tag{5}$$

This modification will pull the cluster centroid towards the location of the largest events in the catalogue. Where many powerful earthquakes have occurred within a small area, the zone may become smaller in size but would contain many events, which is a typical characteristic of seismic source models. The enhanced influence of strong earthquakes in the cluster analysis should improve the stability of the algorithm.

If implementing a weighting scheme, the question arises as to what facet of the earthquake should form the basis for weighting. Earthquake size can be measured in terms of a magnitude unit (here M_w), seismic moment or by dimension of the fault (length, area, slip, etc). The choice of weighting metric should not be arbitrary. Those suggested can span a range of scales, which may have an appreciable influence on the behaviour of the clustering algorithm.

Here, the preferred weighting metric is fault length. This is chosen as it is consistent with the Euclidean space in which clustering is taking place, and it also scales over a sufficiently great range (1–300 km) as to elucidate stronger events from smaller ones. It is not so great a range, however, as to place nearly all the weight in a small number of large events, as would occur with moment weighting.

4.2 Application to ruptures

The weighted centroid in cluster analysis has the advantage of being compatible with the existing schemes for estimating the optimum partition and optimum K. It remains unphysical in that it is still treating earthquakes as point sources, albeit of varying masses. What is needed is a method that takes into account the size and spatial orientation of the seismic source. For this the modification to K-means is less straight forward. To achieve this, a new K-means algorithm is proposed for clustering a set of straight lines as spatial analogues for faults: Line K-means (Fig. 4).

The application of K-means to finite lines rather than points is a novel one, but is computationally expensive. Furthermore, there are limits to which existing clustering indices, developed for conventional K-means, can be applied to the partition of line segments. Those indices that are functions of either TWCSS or WK are still viable for use. To calculate the sum of pairwise Euclidean square distances (WK) for a cluster of line segments the algorithm of Allen *et al.* (1993) is used. Here, the distance between two line segments is defined as the shortest Euclidean square distance at any point between the two segments. If line segments intersect, this distance is equal to zero.

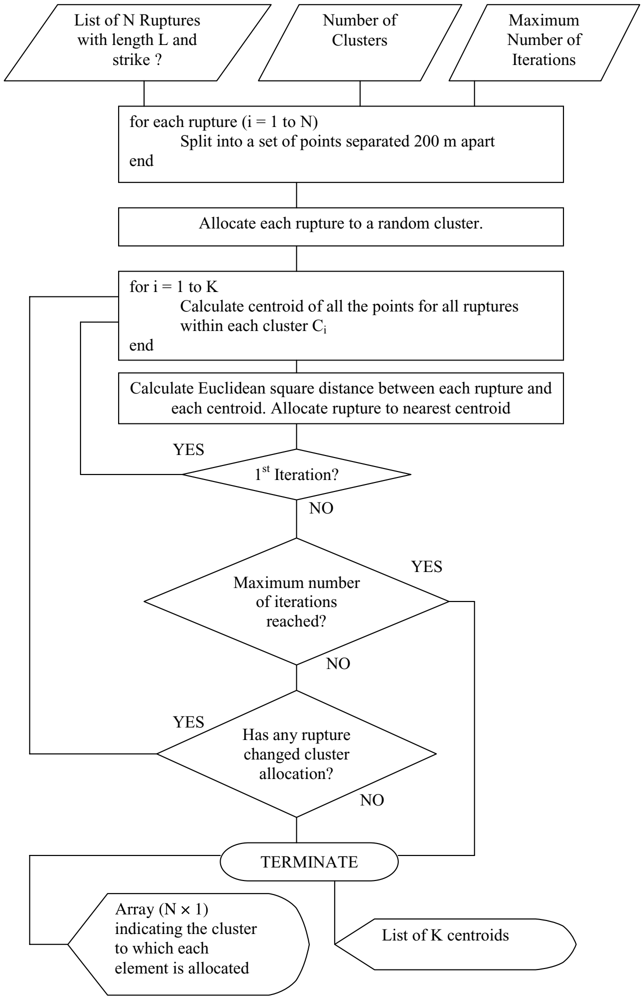
List of N Ruptures with length L and strike ?
Number of Clusters
Maximum Number of Iterations
for each rupture (i = 1 to N)
Split into a set of points separated 200 m apart
end
Allocate each rupture to a random cluster.
for i = 1 to K
Calculate centroid of all the points for all ruptures within each cluster C$_i$
end
Calculate Euclidean square distance between each rupture and each centroid. Allocate rupture to nearest centroid
YES
1st Iteration?
NO
YES
Maximum number of iterations reached?
NO
YES
Has any rupture changed cluster allocation?
NO
TERMINATE
Array (N × 1) indicating the cluster to which each element is allocated
List of K centroids

Figure 4. Flowchart to indicate operation of Line K-means algorithm.

4.3 Illegal partitions

An illegal partition is defined as a partition for a specified K that will, after iteration, produce an empty cluster. This is a common hazard of cluster analysis and the likelihood of illegal partitions increases with higher K. In the earliest K-means algorithms (Hartigan & Wong 1979), an illegal partition would produce an error, which in ensemble analysis would be flagged and excluded from further consideration. The frequency of illegal partitions for high values of K is problematic, as it reduces the number of valid partitions from which the global optimum is determined. Although one could simply proceed with iteration, albeit with $K - 1$ clusters; when identifying the optimum K this is undesirable. Instead a 'singleton' procedure is invoked. This procedure requires that, on production of an empty cluster, the data point furthest from its allocated cluster centre is then selected as a new cluster, upon which the iteration proceeds. This has the impact of slowing down convergence of the K-means algorithm, but this is a satisfactory cost if it allows for more stable partitions at higher K.

4.4 The optimum set of initial clusters

The K-means algorithm is designed to settle at a local optimum partition depending on the choice of initial clusters. In this application a stochastic ensemble of K-means algorithms is conducted. The number of trials in each ensemble is set to 100, and the iteration producing the lowest TWCSS is taken as being the optimum. This is not a global optimum. Given the size of the data set being considered, however, it is assumed that this optimum is sufficiently close to the global optimum for this application.

There are several ways in which the set of initial cluster centres can be constructed in an ensemble analysis. Here we shall be utilizing random partition initialization (Peña *et al*. 1999). This will randomly allocate each point X_i to a cluster C_k and define the initial set of K seeds as the centroids of the randomly allocated clusters. To ensure the stability of the RP initialization N/K points are allocated to each cluster. Where K is not a factor of N, the remaining points are allocated at random in the manner of Krishna & Murty (1999). K-means partitions initialized using RP tend to be less strongly influenced by outliers in the data set (Bradley & Fayyed 1998).

4.5 Identifying the 'optimum' number of clusters in an earthquake catalogue

To assist in the identification of a partition or partitions that are most appropriate for the spatial distribution of seismicity in the Aegean region, several indices were compared. These were the index of Xie & Beni (1991), the silhouette index (Kaufmann & Roeusseuw 1990), the index of Calinski & Harabasz (1974), that of Krzanowski & Lai (1988) and the 'gap statistic' (Tibshirani *et al.* 2001). These indices were tested on a set of artificial data in three dimensions with a known number of clusters in addition to the earthquake catalogue itself. The silhouette index and Calinski & Harabasz index tended to correlate with increasing K, whilst the Xie & Beni index and the 'gap statistic', which performed well when the synthetic clusters were well separated, performed poorly when the clusters were not well separated. Only the Krzanowski & Lai index showed no obvious correlation with K and frequently identified the correct number of clusters within a synthetic data set. The optimum K is that which maximizes the Krzanowski & Lai index (KL):

$$KL(K) = \left| \frac{DIFF(K)}{DIFF(K+1)} \right|, \tag{6a}$$

where

$$DIFF(K) = (K-1)^{2/d}\, WK_{K-1} - K^{2/d} WK_K \tag{6b}$$

d refers to the number of dimensions of the data; WK_{K-1} and WK_K the pooled within-cluster sum of squares (WK) of the $K-1$ and K partitions, respectively.

The performances of all the indices indicated here depend on the data set that is used. All are influenced by the separability of the clusters, with performance improving with increasing cluster separability. Where these particular indices are deficient is in their failure to test the null set of data. This means that they cannot calculate the index for the $K = 1$ case. The 'gap statistic' (Tibshirani *et al.* 2001) does test this case, but performs poorly when identifying the correct number of clusters in tests on synthetic earthquake data sets weighted by rupture length. It is implicitly assumed by the application of the indices that the optimum number of clusters in a data set is greater than 1.

There are many more indices of partition quality used in many applications of clustering algorithms. While the Krzanowski and Lai index may identify a particular value or values of K, it is still sensitive to small changes in the data partition itself. Unfortunately the failure of the ensemble analysis to always identify the global optimum partition for a specific value of K can have an impact. There are several ways in which this can be taken into consideration. The first is to simply increase the number of trials in the ensemble analysis. This assumes that for a finite set of data and K number of clusters, an increasing number n of trials will produce the following result: $\min(\mathrm{TWCSS_n}) \rightarrow \min(\mathrm{TWCSS_{global}})$. The second option is to test not only the value of K that maximizes the Krzanowski and Lai index, but also other values of K that may also stand out as producing good partitions. Krzanowski and Lai recommend testing several of the values of K that produce better partitions for comparison.

When applying data partitioning methods such as K-means to a set of data, there are two considerations that need to be addressed. The first is the problem of identifying the most appropriate number of clusters. The second problem is that of identifying the optimum partition of the data set, for a specified number of clusters. Often, developments in the field of cluster analysis treat these problems independently. The former assumes that for each value of K, the optimum partition for each K is known *a priori*; the latter assumes that K is known *a priori*, but that the optimum partition for the specified K remains to be found. Here, as in many practical applications of cluster analysis, these two problems are not independent. It is not assumed that the optimum number of clusters for either the earthquake catalogue or rupture catalogue is known. Implementation of the cluster analysis therefore, is a two step procedure. Initially for a defined value of K, the optimum partition is found using an ensemble K-means analysis. The Krzanowski and Lai index for the specified K is calculated from the partition that produces the lowest TWCSS from the ensemble. This procedure is then repeated for values of K in the range $2 \leq K \leq 50$.

5 ASSESSING THE SOURCE MODELS USING STOCHASTIC SEISMIC HAZARD ANALYSIS

To identify, both quantitatively and objectively, the seismicity dependent zone models that are more suitable for the region under consideration, it is appropriate to assess how they perform in a seismic hazard analysis. The Monte Carlo method of PSHA (Shapira 1983; Musson 1999; Giardini *et al*. 2004) is a useful method for testing the source dependence in seismic hazard analysis. A basic procedure is suggested in Musson (2004) whereby synthetic earthquake catalogues are produced using the zone model and forecasts are compared to those from the observed earthquake catalogue by χ^2 analysis. A similar approach is adopted here. It is extended further, however, in that the χ^2 analysis is derived from the expected ground motion over a set of data points.

For a set of partitioned earthquakes, synthetic earthquake catalogues can be created in different ways. On the assumption that the earthquakes used in the K-means analysis are Poissonian, the parameters of each synthetic earthquake can be selected in the following ways. Hypocentres can be randomly sampled (with replacement) from the observed data set (Ebel & Kafka 1999), or from a uniform source zone (Musson 1999; Giardini *et al*. 2004). Magnitudes can be sampled from the observed data set or from the cumulative distribution function (CDF) of the Gutenberg & Richter (1944) relation bounded at upper ($M_{\max}$) and lower magnitudes ($m_{\min}$) (Kramer 1996):

$$F(M_w) = \frac{1-\exp\left[-\beta\left(M_W - m_{\min}\right)\right]}{1-\exp\left[-\beta\left(M_{\max} - m_{\min}\right)\right]}, \quad \text{where } \beta = b\ln(10). \tag{7}$$

When synthetic catalogues are created by random resampling of the observed catalogue (Ebel & Kafka 1999) then a partitioned data set is useful. For example, Ebel & Kafka (1999) created synthetic catalogues for the Northeastern United States by randomly resampling from the entire catalogue covering the whole study region. This assumed that seismicity is effectively homogeneous, at least in terms of fault mechanism and strain rate. When considering a region as tectonically diverse as the Aegean, this assumption is less reasonable. The K-means partition allows for the regional catalogue to be broken down into smaller subsets. Creating synthetic catalogues by resampling earthquakes within each subset ensures that the synthetic catalogue maintains the same regional differences in earthquakes as the observed data set.

A set of points partitioned as one cluster is not akin to an areal or spatial zone; thus using the K-means partition to develop a model of seismic source zones is a more complex matter. In this analysis, and for the purposes of automation, zones are created by partitioning (without weighting) the entire source region around the centroids of each cluster, in a manner similar to centroidal Voronoi tessellation within a finite region (Du *et al.* 1999). This is achieved by creating a grid of discrete points (spaced at 0.02° × 0.02°) across the Aegean region. Each gridpoint is then allocated to its nearest cluster centre, using Euclidean square distance as the metric. The zone is then represented as the collection of gridpoints allocated to the respective centres. In the Monte Carlo analysis, synthetic epicentres within a zone are generated by random sampling, with replacement, the gridpoints within each zone. To create a zone from the gridpoints each sampled epicentre is then adjusted by adding random scatter around the sampled gridpoint.

The tessellation method of zone delineation, as implemented, here is arguably the most simplistic approach to the problem of translating partitions into source zones. It has the advantage of transparency and computational efficiency. The former is particularly important if comparing several models in an analysis of epistemic uncertainty. The gridpoints are allocated to each centre with a uniform weighting. This does not necessarily have to be the case as the centroids could be weighted by number of events or within-cluster average moment or magnitude. Tessellation is, however, a simple approach, which will delineate zones in a mosaic pattern within the region being considered. This may appear counter intuitive as it may often be that case that source zones do not appear elongated along the largest ruptures in the zones. Instead, the creation of geometric zones of uniform seismicity around these clusters may be done by inspection. This allows the analyst to create zones that may reflect more accurately the seismotectonic variability of the region, whilst preserving the partitions identified by the K-means cluster analysis. Some consideration of alternative automated approaches to delineating source zones is given in the discussion (Section 8).

Regions where seismicity is trivial (i.e. no earthquakes greater then M_c are recorded in the catalogue) are excluded from the source model to avoid extending zones well offshore into regions of very low seismicity (e.g. the Black Sea, western Libyan Sea). This automated zonation process is used for comparison of source models with the number of zones in the range $2 \leq K \leq 50$.

To quantitatively assess how well the source model represents the observed seismicity of the Aegean, the region is overlaid by a new set of NG grid of points. Using the earthquake catalogue of Burton *et al.* (2004) (cut-off below the magnitude of completeness) the maximum ground motion observed in the period 1900–1999 is calculated for each gridpoint from an appropriate attenuation relation. The source model is then used to simulate 100 yr of seismicity and the maximum ground motion at each gridpoint determined for the 100 yr of synthetic data. This simulation and calculation is repeated a large number ($N_{\mathrm{syn}} = 100$) of times and the spread of maximum observed ground motions for the different 100 yr simulations is developed. Using the geometric mean ($\bar{X}$) and standard deviation (σ) of the N_{syn} maximum simulated ground motions in 100 yr at each gridpoint, it is possible to the calculate a χ^2 for the model:

$$\chi^2 = \sum_{i=1}^{NG} \frac{\{\max[\mathrm{Obs}(\log_{10}[SGM_i])] - \bar{X}_i\}^2}{\sigma_i^2}. \tag{8}$$

To allow for spatial variation in the performance of model, the normalized difference between the observed and expected ground motion is used. Models producing a lower χ^2 value indicate a better fit to the observed data.

Traditional methods of seismic hazard analysis use peak ground acceleration (PGA), peak ground velocity (PGV) and peak ground displacement (PGD), and their spectral ordinates, as the preferred strong motion parameters. For the χ^2 calculation here, ground motion quantified by a single parameter (e.g. PGA) is needed. Whilst PGA may seem the more likely candidate, for the purposes of assessing maximum ground motion for a 100 yr period, it has some shortcomings. In stochastic seismic hazard analysis, the greatest PGA can come from small earthquakes in the near-field, even if such earthquakes may not be especially damaging. Instead, a duration-dependent parameter of strong ground motion is used: Arias intensity (I_{a}) (Arias 1970) defined as:

$$I_a = I_{xx} + I_{yy} = \frac{\pi}{2g}\int_0^{t_0} [a_x(t)]^2\,\mathrm{d}t + \frac{\pi}{2g}\int_0^{t_0} \left[a_y(t)\right]^2\mathrm{d}t, \tag{9}$$

where a_x and a_y are the horizontal accelerations in the x and y directions, respectively, t_0 is the duration of strong shaking and g is the acceleration due to gravity (≈981 cm s^{-2}).

The attenuation relation used is that of Danciu & Tselentis (2007), which is derived from 335 strong motion records from 151 Greek earthquakes. The strong motion records span an epicentral distance of 1–150 km, and a moment-magnitude range of $4.5 \leq M_{\mathrm{w}} \leq 7.0$.

$$\begin{aligned} Log_{10}(I_a) = &-2.663 + 1.125M_{\mathrm{w}} - 2.332\log_{10}\sqrt{R^2 + 13.092^2} \\ &+ 0.028S_0 + 0.200F_0 + 0.524\sigma, \end{aligned} \tag{10}$$

where, R is epicentral distance, S_0 indicates engineering soil class (0 for rock, 1 for stiff soil, 2 for soft soil), F_0 indicates fault mechanism (0 for normal faulting, 1 for strike-slip/thrust faulting) and σ the total standard error including inter and intraevent variability. This attenuation relation is preferred above the global Arias intensity attenuation relation of Travasarou *et al.* (2003) because of its emphasis on Greek earthquakes.

6 LINE K-MEANS ALGORITHM: APPLICATION TO FAULT RUPTURES

6.1 Identification of the optimum number of clusters

The Line K-means algorithm is applied to the modified catalogue of fault ruptures. An ensemble of 100 trials is conducted for each value of K in the range $2 \leq K \leq 35$. The optimum partition from the ensemble for each value of K is obtained and the Krzanowski and Lai index determined (Fig. 5). A reasonable estimate for $K_{\max}$ in many applications of cluster analysis is $K_{\max} \leq \sqrt{N}$ (Sheng & Liu 2006). Given the small number of fault ruptures being considered, a $K_{\max}$ greater than 35 might not seem appropriate and would most likely produce singleton clusters. In this particular application, however, this is tolerable as several earthquakes may be attributable to a single rupture, from which seismic hazard parameters can still be defined. The uniform zones used in the stochastic seismic hazard analysis are created by partitioning a defined grid of points (spaced at 0.02° × 0.02°) around the centroids output from the Line K-means analysis. The non-zoned approach will partition the shallow 20th century catalogue ($M_{\mathrm{w}} \geq 5.2$) around the centroids, then simulating hypocentres for each cluster by random resampling, with replacement, within the cluster. The variation in $\boldsymbol{\chi}^2$ with K is shown in Fig. 6. Seismic hazard parameters for each zone are calculated using the shallow Aegean 20th Century catalogue ($M_{\mathrm{w}} \geq 5.2$), with the full (Figs 6A and B) and Poisson declustered (Figs 6C and D) catalogues compared.

The Krzanowski and Lai index shows a clear peak at $K = 15$ clusters, with other 'good' partitions at $K = 10$, 13 and 22. As expected, there is a considerable amount of variation in the $\boldsymbol{\chi}^2$ analysis. It is not clear to what extent this is due to the quality of fit of the cluster model or to the inherent variability of the Monte Carlo method of seismic hazard analysis. Yet since the variability of the Monte Carlo simulation is accounted for by the σ term in the $\boldsymbol{\chi}^2$ equation, the impact of this variability should not be that great.

The $\boldsymbol{\chi}^2$ analyses that implement uniform source zones, created using the automated zonation technique, clearly show a large degree of variability. There are, however, some values of K that consistently produce low $\boldsymbol{\chi}^2$ values regardless of whether the earthquake catalogue is declustered and whether uniform zones are used. These values are $K = 27$ and 29. When uniform zones are used several values of $\boldsymbol{\chi}^2$ in the range $5 \leq K \leq 10$ stand out, but these are inconsistent when comparing the full and declustered catalogues. When the non-zoned Monte Carlo method is used, $\boldsymbol{\chi}^2$ generally appears to decrease with K below $K = 30$, before then appearing to increase above $K = 30$.

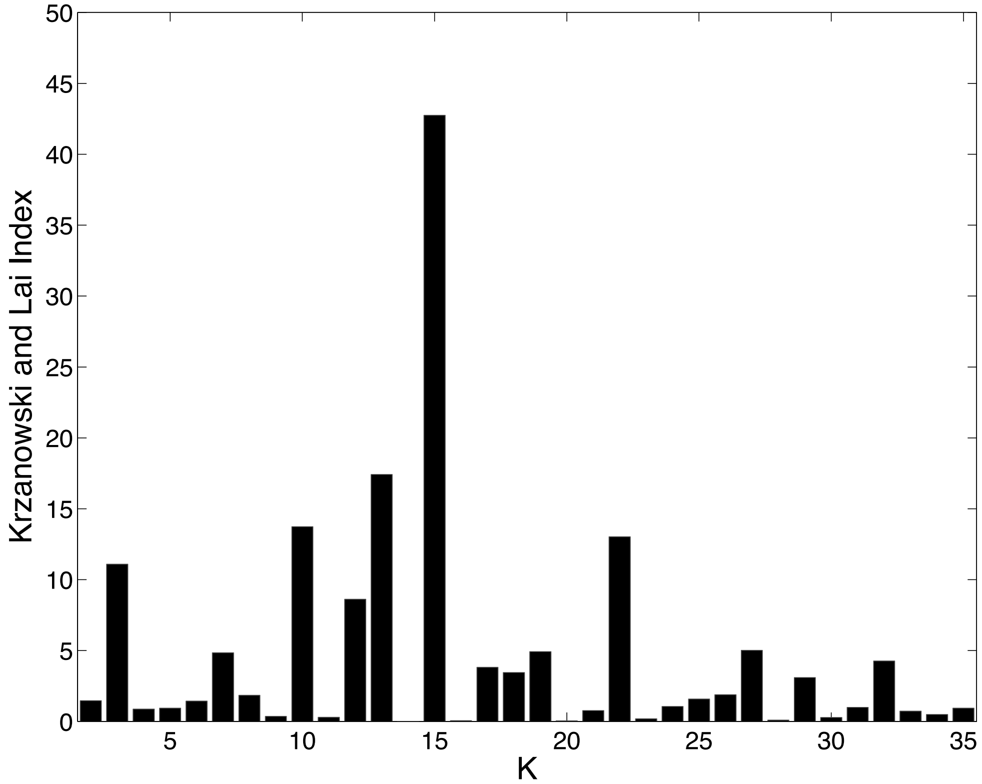

Figure 5. Variation in clustering index with *K* for the Line *K*-means algorithm and modified rupture catalogue of Papazachos *et al.* (1999).

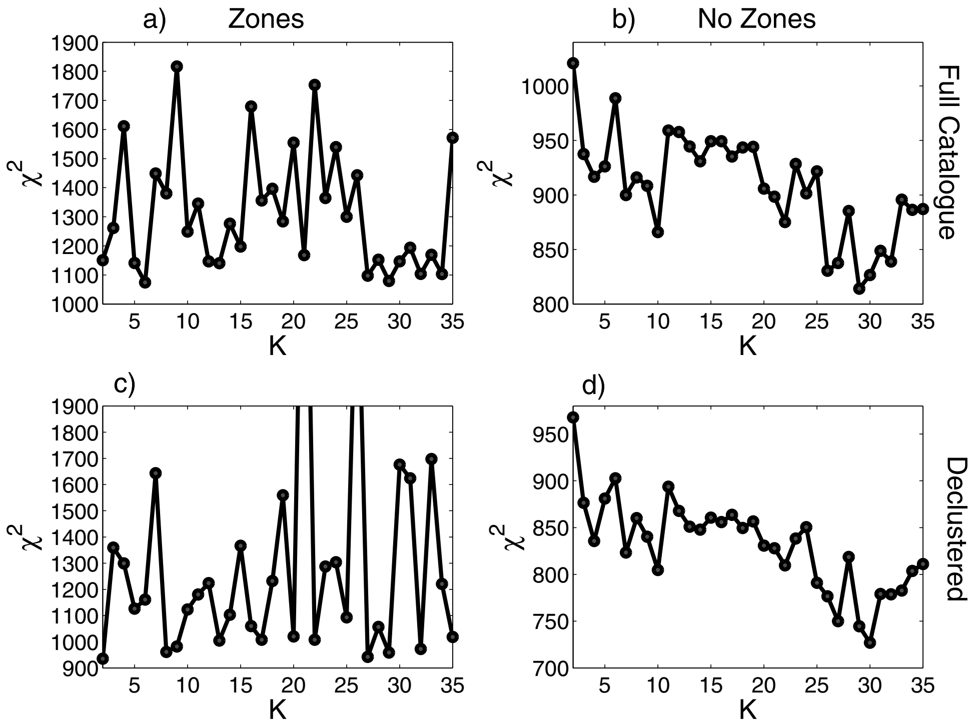

Figure 6. Variation in χ^2 values with increasing *K* for the Line *K*-means algorithm. The rupture catalogue is the same and the reference earthquake catalogue is the Aegean 20th century catalogue of shallow earthquakes ($M_w \geq 5.2$), (A) Full catalogue with zones, (B) Full catalogue without zones, (C) Declustered catalogues with zones and (D) Declustered catalogue without zones.

It is necessary to note that the χ^2 values produced when using the non-zoned approach are considerably lower than those produced using the zoned approach. As described previously, in the non-zoned approach the locations of the synthetic epicentres are determined by randomly resampling, with replacement, from within each cluster, although Gaussian scatter (with $\sigma = 10$ km) is used to allow for greater variability in the epicentral location. When implementing a zoned approach, the epicentres are sampled from a uniform distribution across the entire zone. Consequently, the spatial distribution of synthetic epicentres will display a much greater similarity to the observed distribution when using the non-zoned approach, than for the zoned approach. Given the non-linear decay of strong ground motion with distance, and the definition of χ^2 given in 8, this will result in higher χ^2 values, and a greater degree of variability, when the zoned approach is implemented.

It is not unexpected that the Krzanowski and Lai index and the χ^2 method of identifying optimum K should produce radically different results. The former would appear to suggest that the catalogue of ruptures is best represented by partitions with perhaps 10–15 clusters. The latter, however, suggests that, when compared to the observed seismicity in the region, the rupture catalogue is best partitioned using 27–30 clusters. Except for those partitions where χ^2 is anomalously high, given the number of parameters used in defining the seismic hazard at a location, it is not possible to conclusively identify a partition that is significantly better than the others in a strictly statistical sense. It can be argued that those partitions consistently producing lower χ^2 may be more representative of the seismotectonics of the Aegean.

6.2 Fault rupture partitions and observed seismotectonics of the Aegean

The previous section has shown how the catalogue of fault ruptures in the Aegean may be represented using partitions with several numbers of clusters. The variability in both the Krzanowski and Lai index and χ^2 mean that it is not possible to identify, categorically, a single optimum value of K using just the algorithmic implementation of K-means. Instead, it is more appropriate to review several of the better partitions, to compare these with existing knowledge of the seismotectonics of the Aegean. The following partitions will be scrutinized in further detail: $K = 10, 13, 15, 22, 27, 29$ and 30 (Figs 7A–G, respectively).

In the $K = 10$ partition, the Line K-means algorithm has produced a clear distinction between some of the major tectonic provinces of the Aegean. Of particular note is the plate margin that extends from the Adriatic coast, through the Ionian Islands and along the Hellenic arc. This is described by the five clusters, which broadly correspond to the NW–SE striking compressional zone in Northern Albania and Montenegro, the SW–NE orientated transform fault zone in the Ionian Islands, two zones of NW–SE striking thrust faults along the western Hellenic arc and one more in the eastern Hellenic arc. East to west striking normal faulting on mainland Greece is mostly separated into two clusters: the first spanning most of central Greece and the highly active Gulf of Corinth region, the second encompassing much of the faulting associated with the Mygdonia basin (Tranos *et al*. 2003; Vamvakaris *et al.* 2006).

Where the 10-cluster partition is largely incompatible with the seismotectonics of the Aegean region is in western Turkey and the North Aegean Sea. Seismicity in this region is driven predominantly by two mechanisms: dextral strike-slip movement along the western extension of the North Anatolian fault (NAF), which gradually gives way to back-arc extension in the North Aegean Sea and southwestern Turkey. Several large fault ruptures in the Sea of Marmara and North Aegean Sea define the trace of the NAF, yet the 10-cluster partition separates this region into three large clusters, often separating neighbouring events.

The 13- and 15-cluster models resolve some of the incompatibilities between the partitions of ruptures in western Turkey and the observed seismotectonics. Here we begin to see the trace of the NAF appear, via a series of clusters running from the eastern coast of Evia Island, across the North Aegean Sea and Sea of Marmara, and terminating around the Izmit region. The 13-cluster model is especially pertinent as it partitions the northern branch of the NAF into three clusters (North Aegean Sea, Western Marmara and Eastern Marmara) and adds an additional cluster to account for faults attributed to the southern branches of the NAF in western Turkey. Elsewhere, the differences in fault type remain consistent with the 10-cluster model, albeit with some regions divided into smaller clusters. Of note is the separation of the NW–SE striking thrust faults along the Montenegro coast, from the N–S striking thrust faults and increasing N–S extension seen in eastern Albania and western F.Y.R.O.M.

The 22-cluster model begins to partition, more closely, the extensional faulting observed in central and northern Greece and the North Aegean Sea. Clear distinction is made between the western Gulf of Corinth, the Eastern Gulf of Corinth and the Gulf of Volos. This is in keeping with the differences in geodetic movement between the regions (Clarke *et al.* 1997; Reilenger *et al.* 2006). Where there is some disparity is in the propensity of this model to partition the southern branch of the NAF into may small clusters running east to west, whilst combining transform and oblique faults on the northern branch with the less active region in Northern Greece and Southern Bulgaria. It is remarkable, however, how robust the partitions along the Africa–Eurasia plate margin have been when the number of zones is increased from 10 to 22.

The remaining models ($K = 27, 29$ and 30) begin to resolve many of the concerns of previous models, which arose due to conflicting fault types being partitioned into the same cluster. In all three models the broad trace of the North Anatolian Fault is emerging as a series of smaller clusters that run east to west across the Sea of Marmara, and then turning southwest into the North Aegean Sea. In doing this the K-means method has segmented the NAF along its length making it harder to detect the traces of the northern and southern branches. This is representative of a dilemma that exists in zoning this part of the Aegean. Some source models (PP1990, PP2000) have chosen to segment the western end of the NAF, whilst others (ED1999) prefer to keep the NAF as a single coherent structure. That the K-means method segments the fault is simply a manifestation of the discrete distribution of fault segments from the Papazachos *et al.* (1999) catalogue. From a seismotectonic perspective, the partitioning of the NAF into northern and southern branches, and the increasing extensional component of slip as it enters the North Aegean basin, would strongly suggest that modelling the entire fault as a single entity is not necessarily the most appropriate approach.

The 29- and 30-cluster models generally agree along the length of the plate margin, from the Adriatic Sea to the Dodecanese Islands. The principal difference being that the 30-cluster model splits the NW–SE striking thrust faulting along the Montenegro coastline into two clusters (one of which is a singleton), instead of one. For the purposes of considering seismic hazard in the Aegean region, this distinction may not be that relevant. Perhaps the most interesting distinction between these two models arises in the eastern Corinth region. The Thessaly region is consistently divided into two clusters,

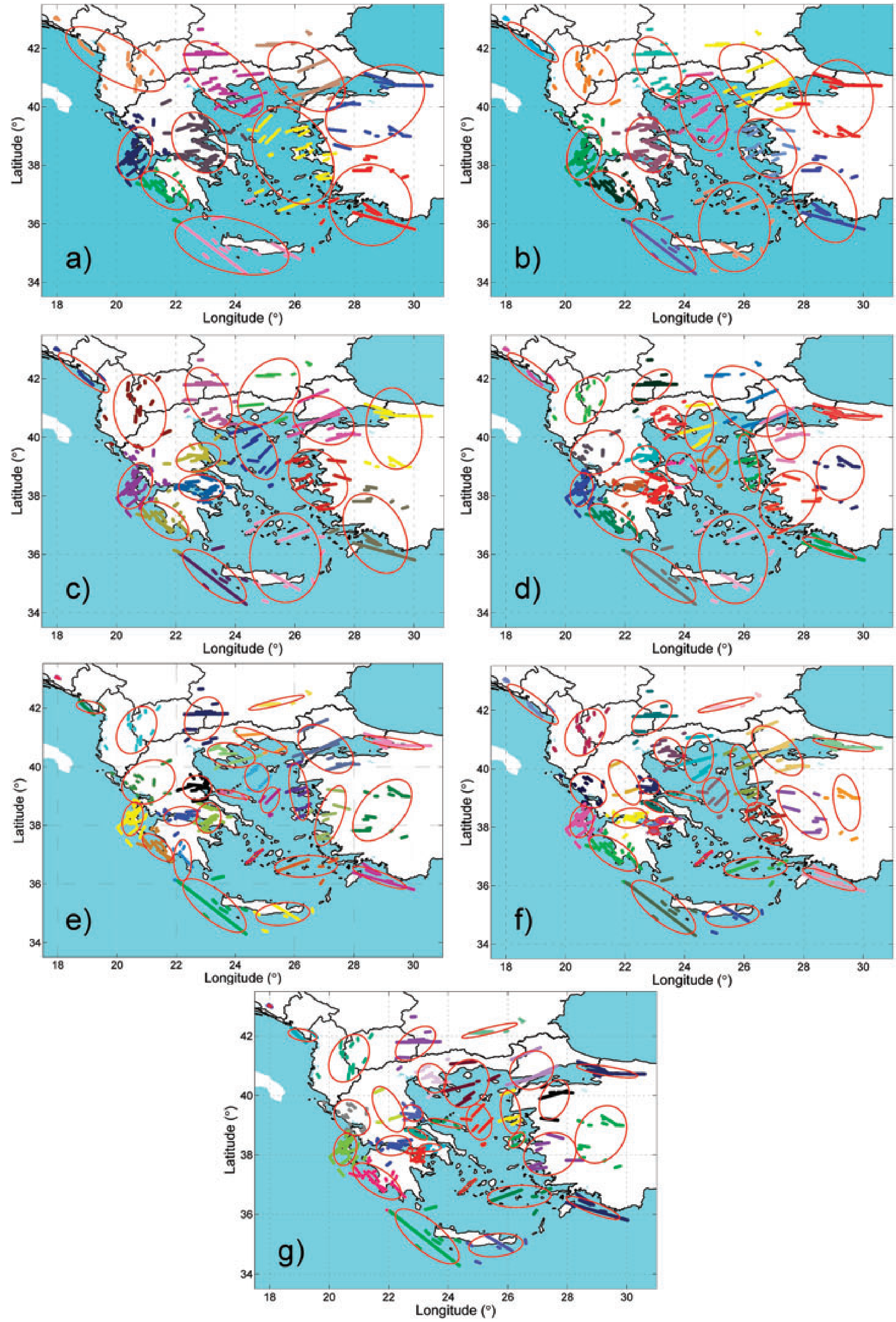

Figure 7. Partitions of the modified catalogue of known ruptures. (A) $K = 10$, (B) $K = 13$, (C) $K = 15$, (D) $K = 22$, (E) $K = 27$, (F) $K = 29$ and (G) $K = 30$. Ellipses are used as markers and are not indicative of a source zone.

one containing the high seismicity around the Pagasitikos Gulf, the other the lower seismicity extending from the Kardista basin northwards to include the 1995 Kozani-Grevena earthquake rupture. Furthermore, both models clearly identify the Atalandi fault zone as one cluster, which is dominated by the ruptures associated with the two large earthquakes of 1894 (6.7 M_w and 7.2 M_w). This disparity arises in the region around the city of Corinth, where surprisingly the 29-cluster models splits the ruptures into two clusters whilst the 30-cluster model keeps it as one. The same distinction between the western Gulf of Corinth and the Attica region remains.

An important question to consider is whether a database of 223 ruptures is sufficiently representative of the tectonics of the Aegean region as to form a reliable basis for zonation using cluster analysis. There are some regions of significant seismic activity that are poorly represented in the catalogue of known fault ruptures. Of these, perhaps the greatest mismatch between observed seismicity and known fault ruptures is found along the western coast of Northern Greece and Albania. Similar mismatches can be found around the Trikala province of central Greece and the areas of offshore seismic activity around the island of Karpathos and immediately to the north of Crete. Conversely, the rupture associated with the A.D. 365 Gortyna earthquake ($\approx 8.3M_w$) dominates the Hellenic arc to the west of Crete; an area that more recently has experienced lower seismicity than along the rest of the arc. Given the considerable uncertainty associated with this earthquake's magnitude and location, and consequently with the rupture length, this may have an unduly large influence over the K-means partition in the Hellenic arc. It is certainly recommended that the K-means algorithm is repeated as more ruptures are discovered; be it in the course of an earthquake or by re-evaluation of historical events.

For interest, a comparison of the 15-, 22-, 29- and 30-cluster partitions is made with the 67-zone PP2000 shallow seismic source model and can be seen in Fig. 8. The shapes of the automated uniform zones can be seen in Fig. 9. Clearly the difference in clusters is such that exact comparison is not possible. Similarities are apparent, however. Especially of interest is the 22-cluster model and the manner in which it partitions the segments of the NAF. The groups of ruptures appear to be separated along the length of the fault using breaks that are close to the zone boundaries of the PP2000 model in that region. Similar occurrences can be found in the Ionian Islands, central Greece and the eastern Dodecanese islands.

7 K-MEANS ALGORITHM: APPLICATION TO AEGEAN EARTHQUAKE CATALOGUES

7.1 Identifying the optimum number of clusters

The K-means algorithm is applied to six subsets of earthquakes from the combined Aegean catalogues of Burton *et al.* (2004) (20th Century), including the supplementary period 2000–2005 (homogenized from the NOA catalogue) and Papazachos & Papazachou (1997) (Pre-20th century).

Subset (1) 20th Century Events: Shallow (Depth $\leq$ 60 km): $M_w \geq 5.2$.
Subset (2) 20th Century Events: Shallow: $M_w \geq 5.2$: Declustered (Reasenberg 1985).
Subset (3) Up to 2000 A.D.: Shallow: $M_w \geq 5.2$.
Subset (4) Up to 2000 A.D.: Shallow: $M_w \geq 5.2$: Declustered (Reasenberg 1985).
Subset (5) 1900 A.D.–2005 A.D.: Shallow: $M_w \geq 5.2$.
Subset (6) 1900 A.D.–2005 A.D.: Shallow: $M_w \geq 5.2$: Declustered (Reasenberg 1985).

The variation in Krzanowski and Lai index with K for each of these catalogues of earthquakes can be seen in Fig. 10. It is certainly clear that the optimum K values determined by this index are not robust when different subsets of the catalogue are considered. While it is reasonable to expect the optimum number of clusters to change depending on whether the catalogue has had aftershocks removed, the change in the optimum number of clusters is so large as to attract concern.

Situations where partition quality indices such as the Kraznowski and Lai index fail to clearly and robustly identify a single optimum K are common in practical application of cluster analysis. Within any clustering criterion implicit assumptions are made as to the size, shape and separability of the clusters (Krzanowski & Lai 1988). Furthermore, an assumption is made that the partitions for a given set of data are global optimum partitions. This may not be the case if (1) the K-means algorithm has reached its maximum number of iterations before the partition stabilizes or (2) the input centroids produce a stable partition, but not a global optimum. The implementation of the K-means algorithm (Nabney 2002) will produce a warning message in the former case. No warnings were observed in the implementation of the algorithm, confirming that all partitions had stabilized before reaching the maximum number of iterations. It is possible, however, that the global optimum partition was not output from the ensemble analysis. Repetitions of the analysis with a greater number of ensembles produced minor variations in KL index for only a few values of K and maintained the same optima shown here. This suggests that the partitions produced for each K are stable, and that the absence of an optimum K that is identified regardless of the subset used arises from the application of the index to poorly separated clusters. All of which adds further impetus to ensure that partitions are scrutinized carefully in the context of Aegean seismotectonics.

The only value of K that appears as a 'good' partition in more than one index is $K = 36$, which is identified as the optimum for Subset 4, and a 'good' partition for Subset 6. Also, $K = 4$ stands out as the optimum partition for both Subset 1 and Subset 3 although we believe a partition with so few clusters to be incompatible with the seismotectonics of the Aegean.

The optimum-K indices shown do not identify a consistent number of clusters. This lack of robustness in the optimum K, as defined by the Krzanowski and Lai index, may suggest that seismicity in the Aegean is not readily characterized by a clearly identifiable number of well-separated clusters. The next step is to apply χ^2 to identify the number of zones that will most closely reproduce the observed hazard. The reference catalogues used to define the observed hazard are those spanning the 20th Century: Subset 1 (aftershocks included) and Subset 2 (aftershocks removed). These catalogues are both complete above 5.2 M_w. Two types of seismic source are compared: the first produces synthetic epicentres by randomly resampling the epicentres within each cluster, the second will sample epicentres from a uniform source zone produced using the centroid method described in Section 5. Depths are sampled from a uniform distribution in the range 0 < depth (km) $\leq$ 60 (km). Magnitudes are sampled from the CDF of the truncated Gutenberg & Richter (1944) relation (Kramer 1996). Maximum magnitude (M_{max}) for each cluster is calculated using the cumulative moment method of Makropoulos & Burton (1983). This method is chosen because it is independent of b-value and cannot produce a value of M_{max} lower than the observed M_{max} in each cluster. Where a cluster contains

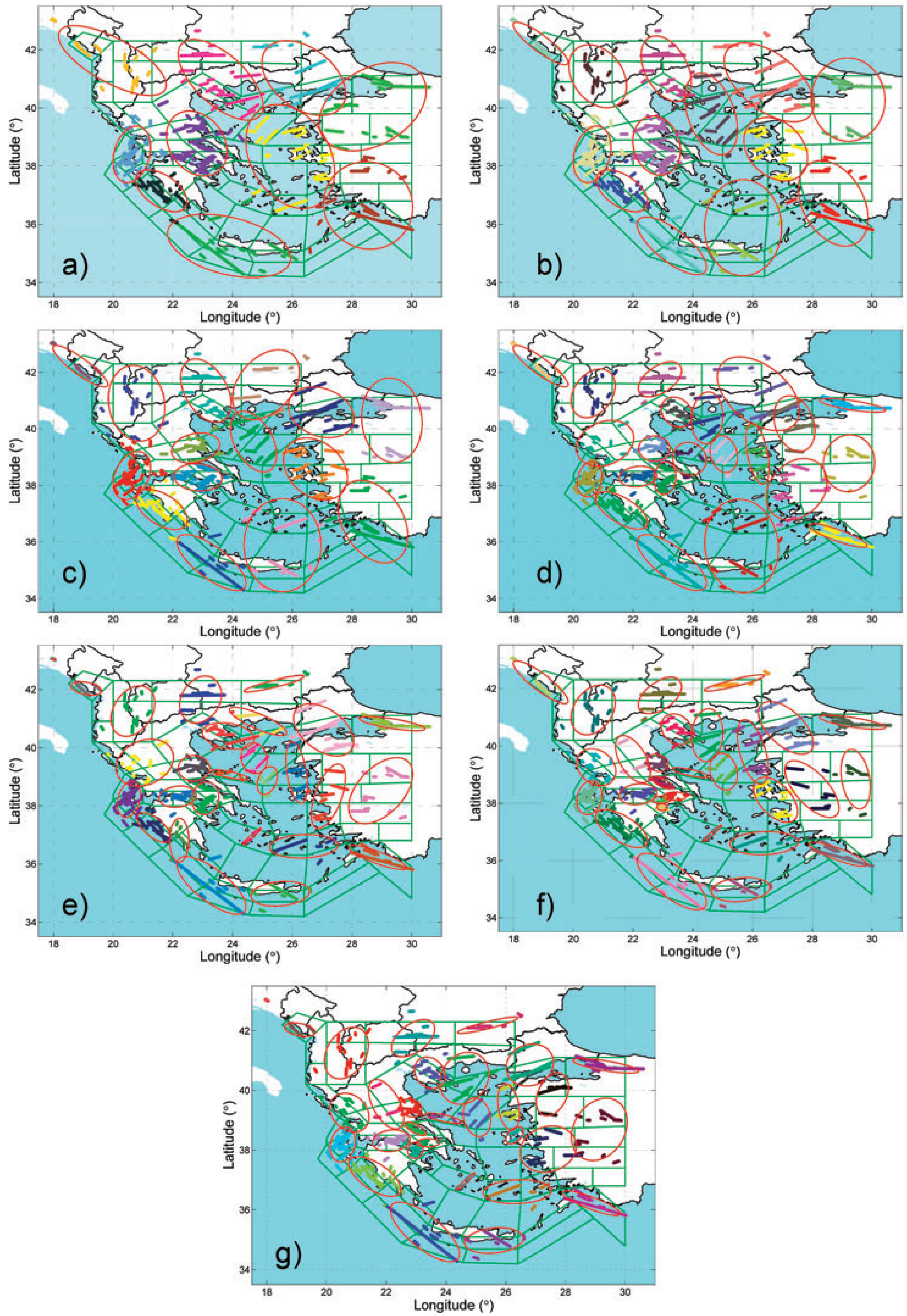

Figure 8. Comparison of the (A) 10-cluster, (B) 13-cluster, (C) 15-cluster, (D) 22-cluster partition, (E) 27-cluster, (F) 29-cluster and (G) 30 cluster with the source model of Papaioannou and Papazachos (2000) (marked in dark green). Ellipses are used as markers and are not indicative of a source zone.

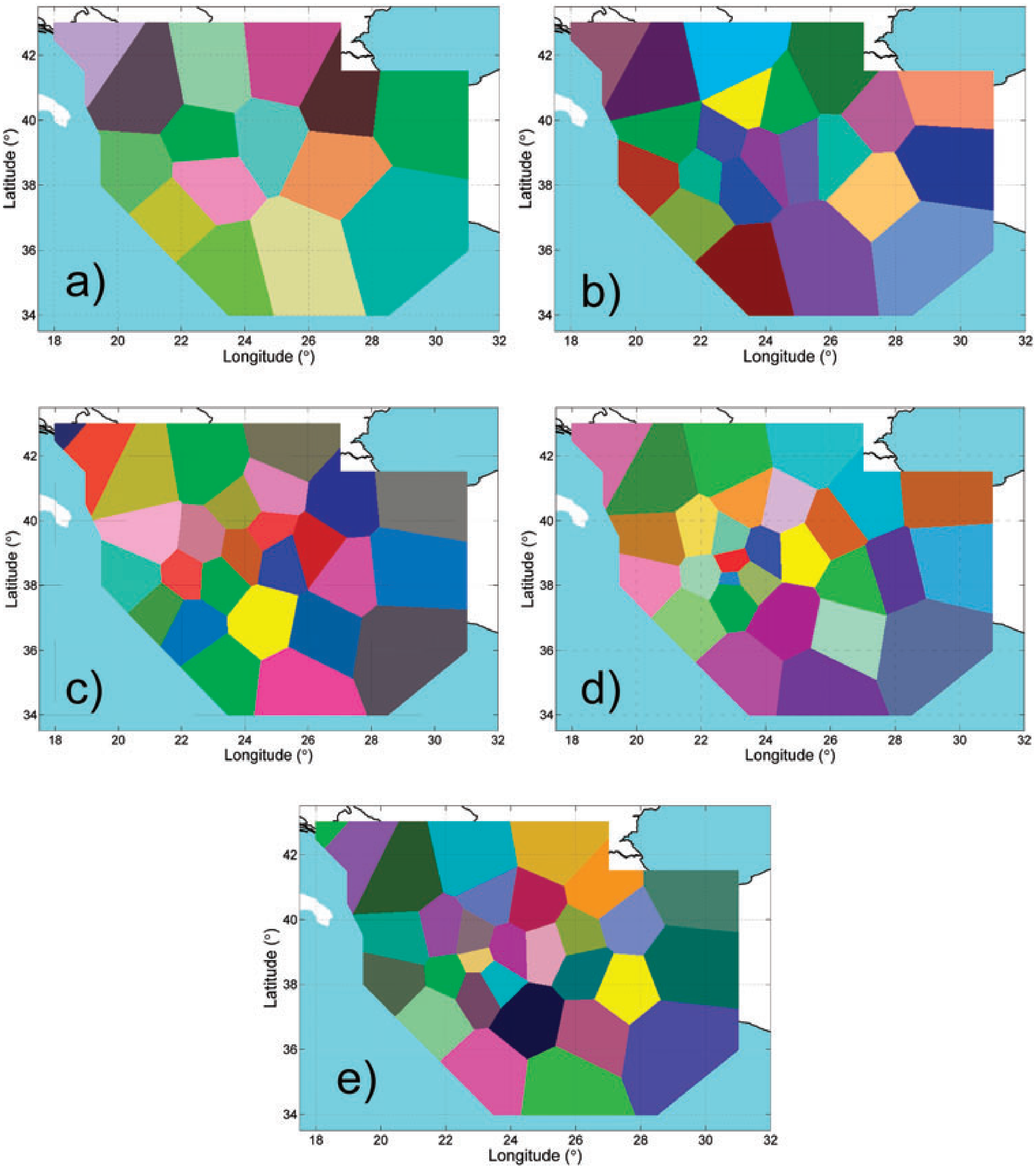

Figure 9. Zones created by partitioning around the centroids of the rupture catalogue. (A) 15-cluster, (B) 22-cluster, (C) 27-cluster, (D) 29-cluster and (E) 30-cluster.

too few earthquakes from which to determine the *a*- and *b*-value, these parameters are interpolated using Inverse Distance Weighting of the parameters of the nearest 5 clusters for which the parameters have been determined. For each *K*, hazard across a grid of points spaced at 0.5° is determined from 100 synthetic catalogues each of 100 yr in length. The maximum strong ground motion (Log I_a) in 100 yr for each gridpoint is then compared with the observed hazard from the 20th century catalogue. The results can be seen in Fig. 11.

As is clear from Figs 10 and 11, the identification of the correct number of clusters or zones still remains a subjective process. Using the indices suggested and the χ^2 method, there is substantial discrepancy between the 'correct' number of zones identified when different elements of a catalogue are used. Though consistency in the optimum number of clusters would be desirable in the sense that one could more readily justify a particular zone model, lack of consistency is not unexpected. It is clear that each index of cluster quality will perform differently given the data set at hand, the controlling factor usually being the separability of clusters. Although spatial clustering of earthquakes is observed in the original catalogue and the declustered catalogue, there is a substantial degree of diffusivity in the distribution of hypocentres. When partitioning

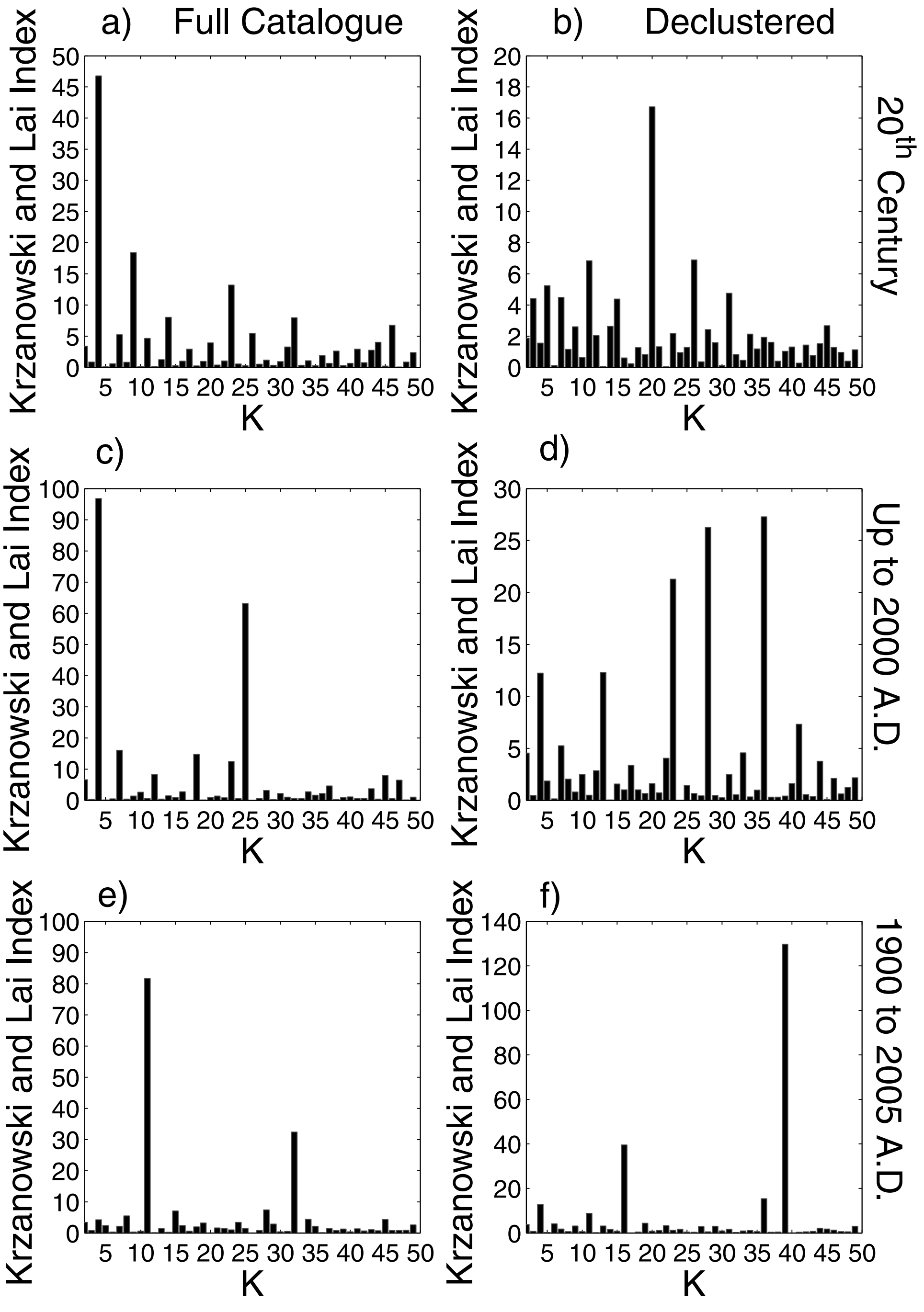

Figure 10. Variation in Krzanowski and Lai (1988) index with *K* for Aegean shallow earthquake catalogues ($M_w \geq 5.2$). (A) Subset 1, (B) Subset 2, (C) Subset 3, (D) Subset 4, (E) Subset 5 and (F) Subset 6.

with a pre-defined number of clusters, as with the *K*-means partition, the algorithm is forced to separate all the data under the same criteria. This means that compact groups of earthquakes can, and often will, be assigned to the same clusters as more widely distributed events, although this should occur less frequently in better fitting partitions.

The variation in χ^2 with increasing *K* produces interesting results. It is immediately apparent that there is little to indicate a trend of improving fit with a greater number of zones. Whilst there is no single value of *K*, or narrow range that would suggest a robust optimum number of zones, some values appear more frequently than others. In particular, when uniform source zones are used, several values of *K* in the range of 40–50 often appear to produce a good fit. Source models producing lower χ^2 values are also found with *K* in the range 30–40.

Ultimately, the decision as to how many zones to use may still lie with the user of the hazard analysis. This methodology can also be used in a logic tree or Monte Carlo analysis of uncertainty in seismic hazard. A selection of the better fitting models can be input into a layer of the logic tree, with particular models weighted in proportion to their fit to the observed catalogue. For the purpose of analysing the source models, clusters with *K* values of 5, 10,

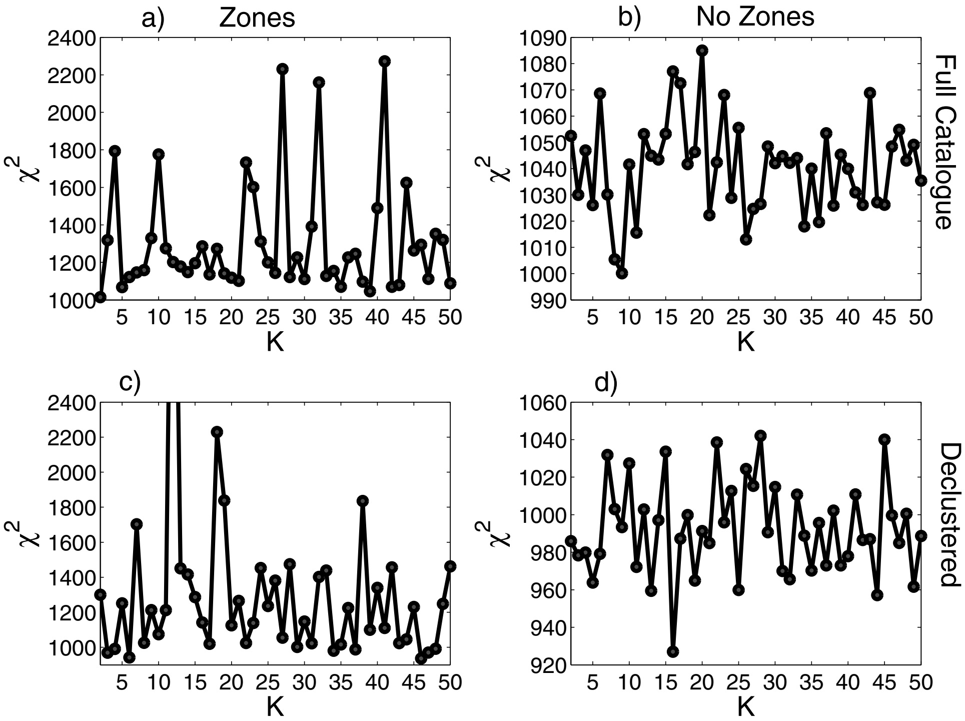

Figure 11. Variation in χ^2 value with number of clusters using random re-sampling (B and D) and zones (A and C). The catalogue used is the Aegean 20th century catalogue with only shallow (Depth $\leq$ 60 km) events, complete at M_w 5.2. (A) and (B) use the full catalogue (Subset 1), (C) and (D) the declustered catalogue (Subset 2).

20, 30, 40 and 50 will be analysed. These partitions can be seen in Fig. 12.

7.2 Comparing the partitions to the observed seismotectonics of the Aegean

Comparison of the point source partitions with knowledge of the seismotectonics of the Aegean region reveals some similarities to the partitions of the fault rupture catalogue. Where fewer clusters are used, the K-means algorithm will partition the hypocentres in a manner that loosely characterizes the differences in overall faulting pattern. The shape of the uniform source zones can be seen in Fig. 13. In the $K = 5$ example, the clusters are so large that even though different provinces of the Aegean are separated, each cluster still contains earthquakes that are a mixture of fault types. When the number of clusters is increased to 10, some of these problems are resolved. For example, the compressional faulting regime along the Adriatic coast is now separated from the E–W striking extensional faulting seen in F.Y.R.O.M. and Bulgaria. The transform and compressional faulting of the Ionian Islands is now separated from the extensional faulting found in Central Greece. Where there is still a problem in the 10-cluster model is in the characterization of the North Anatolian Fault. With so few zones, the distributions of hypocentres that broadly demarcate a linear fault structure are embedded within two clusters that cover much of western Turkey and the North Aegean Sea.

The 20- and 30-cluster partitions offer perhaps the best compromise between seismic sources that are small enough to conform to local variation in seismotectonics, whilst still having a sufficient number of earthquakes in each cluster to determine b-value and $M_{\max}$. In the 20-cluster model, there is considerable distinction between compressional, strike-slip and normal faulting regimes. Furthermore, the normal faults that typify much of mainland Greece and the southern Balkans are broken down into smaller clusters that appear well-separated by regions of low seismicity. Good examples of this are the three E–W striking clusters that encompass the Gulf of Corinth, the Gulf of Volos and the Thessaloniki-Rentina Fault Zone (Tranos *et al.* 2003), respectively. The trace of the North Anatolian fault is becoming visible by way of a series of clusters that closely follow the band of high seismic activity that extends from the eastern Sea of Marmara into the North Aegean Sea.

The 30-cluster partition makes further distinctions between groups of hypocentres, predominantly in the region of high seismicity around the Gulf of Corinth and Peloponnese. With more clusters being used, the K-means algorithm splits seismicity of the Gulf of Corinth into east and west sections, with the eastern section incorporating the normal faulting in the Parnitha region. A similar division is made in the Thessalia region of Greece, with the onshore seismicity of the Pagasitikos Gulf and surrounding region being separated from the offshore seismicity in the North Aegean Sea and northern coast of Evia island. Also, a distinction is made between seismicity in the eastern and western Sea of Marmara. Another useful result is that that the 30-cluster partition groups some of the isolated events found well-offshore in the Mediterranean into a widely-dispersed low-seismicity cluster. This particular cluster contains so few events, and is sufficiently far away from inhabited regions in the Aegean that it could reasonably be considered as having a trivial influence on engineering seismic hazard in the Aegean. At the same time, it removes some of the outliers from clusters in the Hellenic arc, making them more tightly constrained and an improved representation of seismicity along the arc. In effect, this 30-cluster model could really be considered a 29-cluster model for the purposes of seismic hazard analysis. It should be noted that

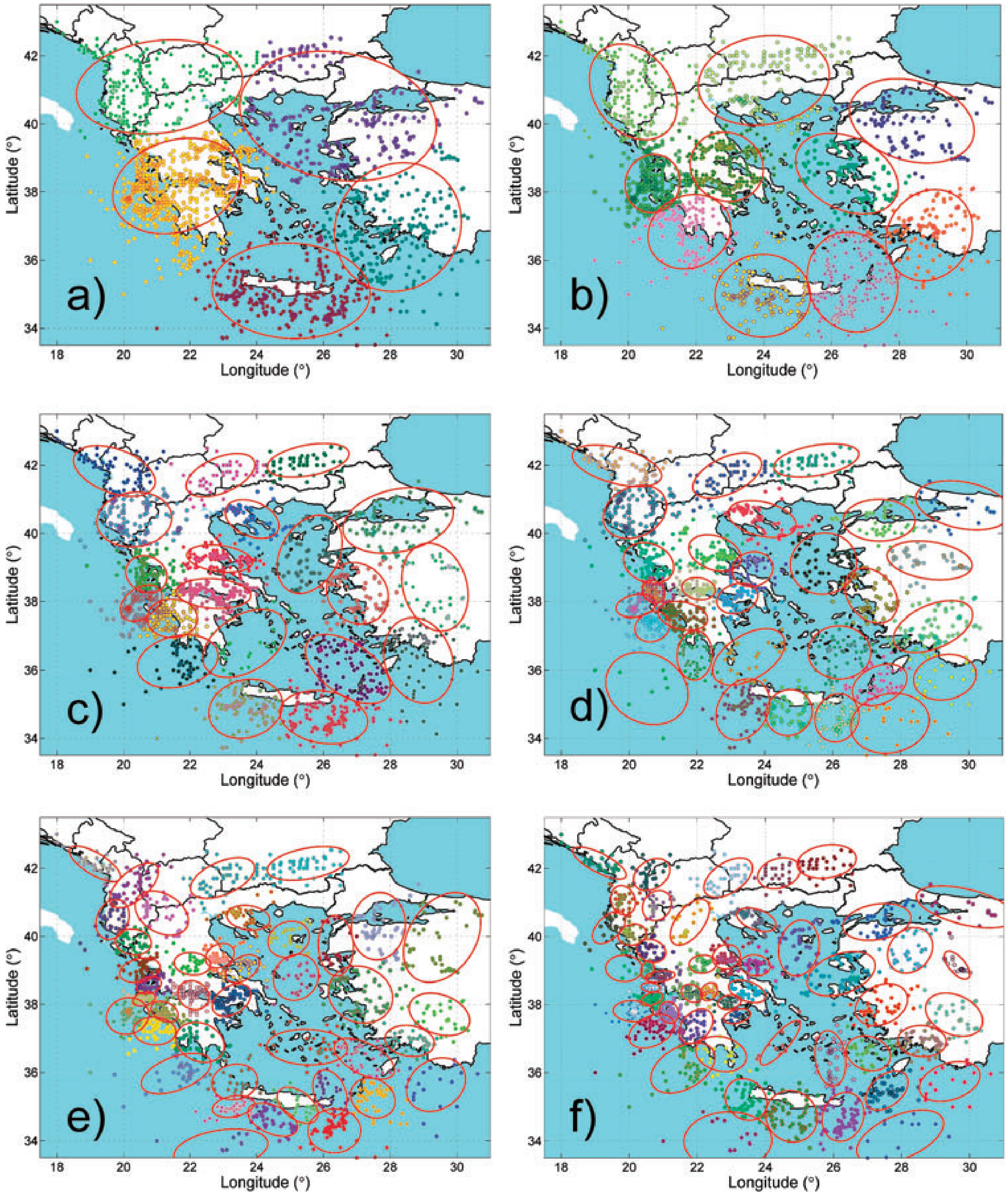

Figure 12. Partitions of the 20th Century Shallow (Depth < 60 km) Aegean Catalogue ($M_w \geq 5.2$). (A) $K = 5$, (B) $K = 10$, (C) $K = 20$, (D) $K = 30$, (E) $K = 40$ and (F) $K = 50$. Ellipses are used to distinguish between clusters and are not indicative of the size or shape of the source zone.

of the six partitions shown, the 30-cluster partition produced the lowest χ^2 when using a finer resolution (0.2° × 0.2°) grid, which is marginally better than the $K = 50$ model and substantially better than all the models with $K \leq 20$.

The remaining models with higher numbers of clusters ($K = 40$ and 50) continue this trend of subdividing areas of high seismicity into smaller groups. In doing so, many sets of neighbouring clusters emerge, for which the seismotectonic properties are so similar that they may be indistinguishable when their uncertainties are taken into account. This is exacerbated by the fact that with fewer earthquakes in each cluster, the uncertainties on particular properties such as b-value, $M_{\max}$ and strike are going to be greater. By overpartitioning the hypocentres there is a greater risk that the delineation of a source zone will be made based upon a transient feature of seismicity within the observed catalogue, even when a declustered catalogue is used. Although declustering suggests a time invariant

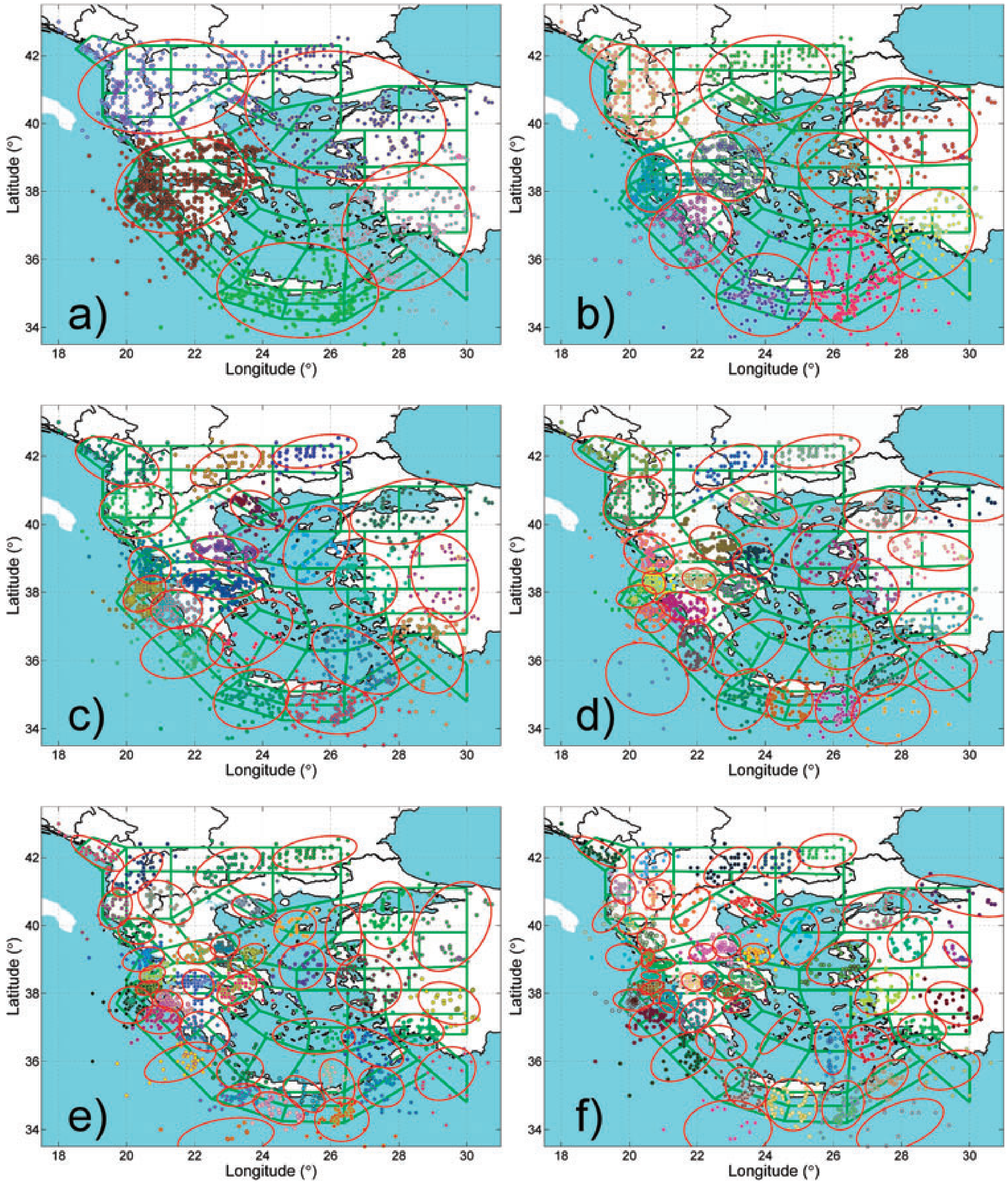

Figure 13. Comparison of the (A) 5-cluster, (B) 10-cluster, (C) 20-cluster, (D) 30-cluster, (E) 40-cluster and (F) 50-cluster partition with the source model of Papaioannou and Papazachos (2000) (marked in dark Green). Ellipses are used as markers and are not indicative of a source zone.

catalogue with properties of stationarity, it will be the case with increased subdivision that features and manifestations with intrinsic return periods beyond the observed catalogue duration will be ill-determined. As the $K = 50$ model offers no robust improvement of fit over the 30-cluster model, then the 30-cluster model should be preferred. Exception to this approach might be made where evidence for a higher cluster model comes from information pertaining to the physical properties of faulting in the region not adequately captured by the seismicity. Despite the growing knowledge of Aegean tectonics, we do not believe such an argument can be made in this example, and hence would recommend use of the 30-cluster model rather than the 40- or 50-cluster model in a seismic hazard analysis, or, if several models are used, should be weighted accordingly.

The partitions are, again, compared with the PP2000 source model (Fig. 13), and with the automated uniform zones shown in Fig. 14. Obviously, it is difficult to draw comparison with the 67-zone PP2000 model when considering partitions with very few ($K \leq 20$) clusters. Some interesting features emerge in the 30-,

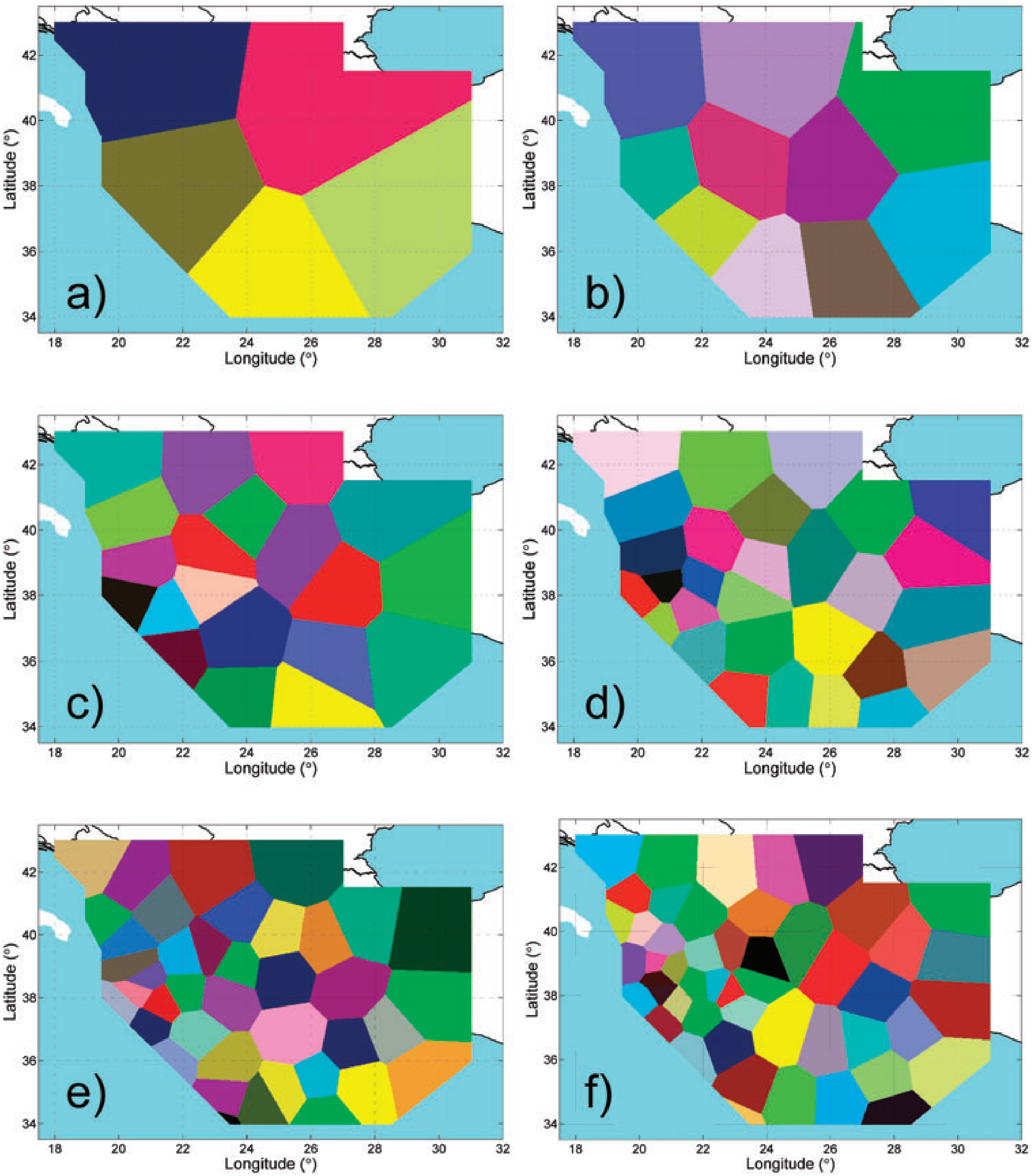

Figure 14. Zones created by partitioning around the centroids of the shallow Aegean 20th century earthquake catalogue: (A) 5-cluster, (B) 10-cluster, (C) 20-cluster, (D) 30-cluster, (E) 40-cluster and (F) 50-cluster.

40- and 50-cluster models, which may present an argument for grouping some of the PP2000 zones. In the 30-cluster partition many of the clusters demarcated in the Ionian Islands and Central Greece broadly correspond to the zone boundaries of the PP2000 model. The cluster encompassing the western Gulf of Corinth divides seismicity in the same manner as a single zone that incorporates zones 42 (Patra), 43 (Aeghio) and 39 (Agrinio) of PP2000. Similarly the cluster of hypocentres in the eastern Corinth and Parnitha region encompasses zones 41 (Thebes), 44 (Corinth), 45 (Methana) and 51 (S. Euboikos). These particular similarities persist in the 40-cluster model. There are still many discrepancies between the models, particularly in the eastern Aegean region. Furthermore, none of the cluster models are particularly adequate in capturing the narrow band of high seismicity that runs from the Cyclades islands to Western Turkey. In this region, the PP2000 model may be more suitable.

8 DISCUSSION

The source zone models presented demonstrate how the *K*-means algorithm can be used to delineate seismic sources without necessarily introducing additional information about the seismotectonics of a region. The motivation for applying this procedure is to establish a degree of consistency in the development of seismic source models. It is important to recognize that the partitions presented arise from a stochastic process, not in the *K*-means algorithm itself but in the seeding of initial centroids within the ensemble. For each model of *K* clusters, the partitions presented in Figs 7, 8, 12 and 13 are not global optima but local optima, albeit with fits close to those of the global optima.

It emerges that when using the optimum-cluster searching methods presented here, the optimum *K* is sensitive to changes in the catalogue. To deputize the *K*-means algorithm, to the extent that it is considered to have the same nuanced appreciation of earthquake behaviour as a working seismologist, would be a dangerous approach. *K*-means is simply designed to partition a set of data (commonly point data, though we have extended that to line segments here) in such a manner as to minimize the total within cluster sum-of-squares, for a given set of initial centroids.

Although a powerful tool in partitioning a set of spatially distributed data, the *K*-means algorithm is a largely unintelligent algorithm, unable to take into consideration the needs of the output. Clusters that contain a very small number of points, even singletons, are tolerated within the *K*-means algorithm, and also in many of the indices to identify optimum *K*. The Krzanowski and Lai index is no exception. If using these partitions to define clusters of similar seismicity (and eventually uniform source zones), clusters or zones containing a very small number of events may not be suitable for the purposes of seismic hazard analysis. Parameters of earthquake behaviour [e.g. Gutenberg & Richter (1944) *a*- and *b*-value, or $M_{\max}$] for such zones will be subject to enormous uncertainty, or may be biased by a single event. Therefore, an optimum *K* defined by the procedure described previously, does not necessarily translate directly into the optimum source model for use in seismic hazard analysis.

Where the data contains compact well-separated clusters, a consistent and global optimum may be found and may be robust. For the distribution of hypocentres and/or ruptures in the Aegean, this is not the case. Spatial clusters of earthquakes in the Aegean are usually not well-separated. Furthermore, earthquakes have also occurred in regions well away from clear tectonic margins and large-scale active fault structures. Their inclusion into the data set means that they are subject to partition too. This can result in isolated earthquakes, presumably with ill-determined physical features, being attributed to clusters in such a manner that may appear unrealistic to a person well-acquainted with the seismotectonics of the region in question. When attempting to delineate uniform source zones, this has the impact of expanding such a zone over a much greater area than is appropriate to adequately model the seismic hazard in the region. However, in lower seismicity regions it may be the case that seismicity is well-distributed spatially, and hence attribution to an active seismogenic structure may not be possible. In these circumstances where information regarding the seismotectonics of a region may not be comprehensive, cluster analysis may prove a substantially more robust method of delineating zones with similar seismic properties.

The use of the perpendicular bisector of neighbouring zones as a tool for delineation of zone boundaries is simple and effective, but does not necessarily take into consideration the shape of faults in each zone. When partitioning a set of evenly distributed point data this may not necessarily be an issue. When partitioning ruptures, however, it may be desirable to delineate zones in accordance with the orientation of rupturing. This can, of course, be done by manually using the partitions as a guide for zone shape. If opting to automate the process, the following method can be implemented as an alternative.

For each cluster, a linear set of discrete, evenly spaced points is defined along the length of all the ruptures in the cluster (a spacing of 200 m is used here). The set of points is then smoothed across a grid of 0.1° × 0.1° spacing, using the spatial seismicity smoothing method (Frankel 1995; Stirling *et al.* 2002). The smoothed number of points in each cell (Ns_i) is calculated by:

$$Ns_i = \frac{\sum_j \left[N_j \exp\left(-d_j^2/c_i^2\right)\right]}{\sum_j \left[\exp\left(-d_j^2/c_i^2\right)\right]}, \tag{11}$$

where N_j is the number of rupture points in grid cell *j*, d_j is the distance between the centre of the current cell and the centre of cell *j*, c_i is the correlation distance (assumed here to be a constant 50 km). The correlation distance can be varied in accordance with the length of observed faulting within the cluster being considered if deemed necessary.

Once a grid of smoothed rupture points is defined a two dimensional Gaussian function is then fit to the points:

$$f(x, y) = Ae^{-\left[a(x-\bar{x})^2+2b(x-\bar{x})(y-\bar{y})+c(y-\bar{y})^2\right]} \tag{12}$$

$$a = \frac{\cos^2\theta}{2\sigma_x^2} + \frac{\sin^2\theta}{2\sigma_y^2}, \quad b = -\frac{\sin 2\theta}{4\sigma_x^2} + \frac{\sin 2\theta}{4\sigma_y^2} \text{ and}$$

$$c = \frac{\sin^2\theta}{2\sigma_x^2} + \frac{\cos^2\theta}{2\sigma_y^2},$$

where $\bar{x}$ and $\bar{y}$ are the mean points, with standard deviations of σ_x and σ_y respectively; *A* the amplitude of the function and θ the angle of rotation relative to the longitude axis. This process is then repeated for all the clusters until a functional surface is formed across the region.

For the $K = 30$ rupture partition the function surface of eq. (12) is shown in Fig. 15. The alignment of the Gaussian function for each cluster often aligns closely with the ruptures, where the length of the longest rupture is large compared to the distribution of ruptures within the cluster. Where the ruptures are well-dispersed within a cluster the Gaussian function appears more circular.

To translate this surface into seismogenic source zones the boundary of each zone is delineated by the location of equal function value between a cluster and its neighbours. This can be done either by inspection or via an algorithm. Here we opt to partition the grid of evenly spaced points, as used in the tessellation method (0.02° × 0.02°), and assign each point to a cluster according to which cluster gives a higher value of the smoothed function for the point. For the $K = 30$ example this is shown in Fig. 16, with the partition of the rupture catalogue superimposed over it.

The zone model shown in Fig. 16 clearly bears a significant resemblance to zones delineated using the tessellation method for the $K = 30$ rupture partition in Fig. 9(D). Whilst the mosaic pattern is not quite so uniform, the general shape of the zones is very similar. This would suggest that although the tessellation method of delineating uniform zones from the partitions does not explicitly account for the physical effects of faulting, it may serve as a reasonable approximation. Furthermore, the alternative method shown here requires the input of a greater number of parameters. These include the resolution of the smoothing grid and the correlation distance. It is not obvious that the additional uncertainty introduced from these parameters in the alternative method, and the detach-

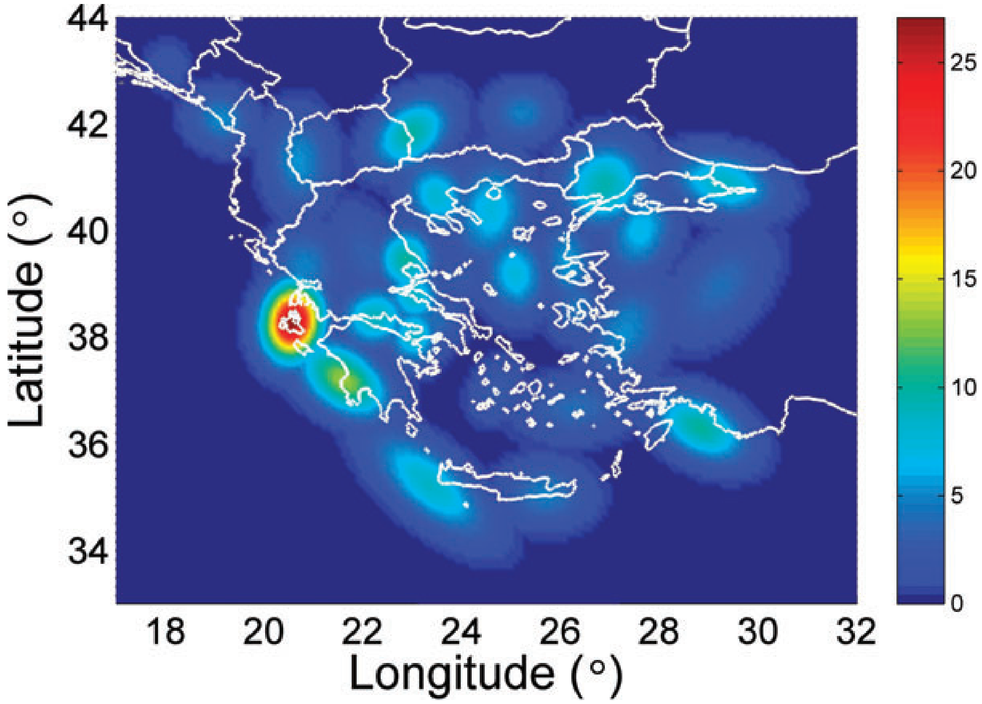

Figure 15. Functional surface of eq. (12) fit to the $K = 30$ partition of Aegean ruptures. See text for description of the method.

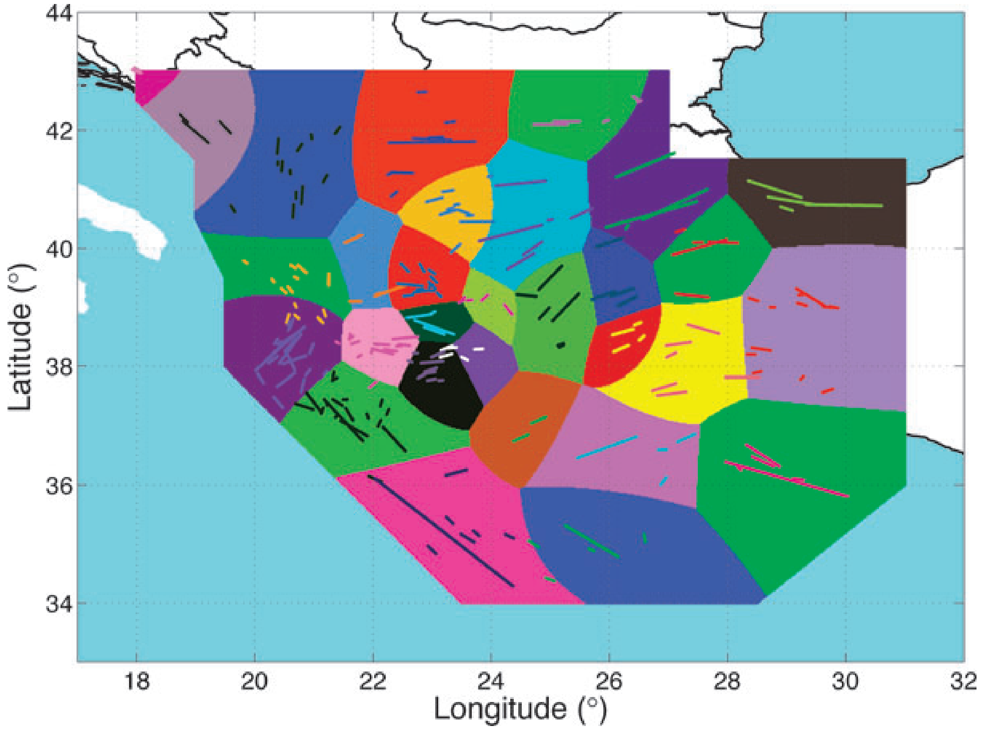

Figure 16. Seismic source zones for the $K = 30$ partition created by splitting neighbouring clusters along lines of equal fit of the function in eq. (12). The partitioned ruptures are superimposed on top of the zones.

ment from the observed spatial distribution of seismicity arising from Gaussian smoothing, justifies the additional complexity.

It is clear that source models produced by way of K-means analyses should not be used without some attempt to relate the partition to the observed tectonics of a region. Although objectivity is the aim of this procedure, there is still a strong case for analysing the zones in the manner presented here and accepting or rejecting particular models depending on whether they are recognized to be a reasonable approximation of the tectonics of the region. Where there may be multiple source models of similar virtue, stochastic seismic hazard methods, such as the Monte Carlo methods presented here or those of Cramer *et al.* (1996) and Smith (2003), can be used to investigate the impact that epistemic uncertainty in the source model has on the overall uncertainty within the seismic hazard analysis.

The question arises as to where these partitions fit within the field of existing source models in the Aegean region. Comparison has been made, in this paper, with the Papaioannou & Papazachos (2000) source model, which is representative of the state of knowledge of seismicity and tectonics at the time of its production. We refrain from suggesting that any of the cluster models here should be given preferential consideration over existing source models derived by other authors. What is demonstrated here are two approaches that increase the objectivity of the source delineation process, and can be analysed with respect to the observed seismicity of the Aegean. This may, however, increase the range of models that could be considered within an analysis of the epistemic uncertainty of the source zone. We also recognize that, as with traditional approaches to source zoning, these partitions may change as new information becomes available.

What is presented in this paper serves as an initial example of the application of K-means to seismic source delineation where sources are visualized as line ruptures rather than simply point sources. Many further elaborations could be made to this method depending on the tectonic information available. A different hierarchical cluster analysis approach to seismotectonic zonation has been applied to Iran by Zamani & Hashemi (2004). This uses Ward's (1963) method of cluster analysis to separate regions of 'similarity', based on 25 geological and geophysical parameters for each zone. They too do not resolve the issue of the optimum number of clusters (in their case leaving it to the user).

There is enormous potential for extending the K-means method and including more geological and geophysical information into the analysis. Possible further developments that would be beneficial are indicated below. Some of these relate to the improvement of the performance of the K-means algorithm, others relate to the incorporation of additional geophysical properties of the seismotectonics. Due to the extensive variety of computing problems to which K-means can be applied, it is reasonable to expect that more modifications to the algorithm will be developed in the future. Some modifications may prove useful in this application, especially those designed to find the global optimum partition and the optimum number of clusters. There may very well be modifications and improvements to both the K-means procedure and the indices of partition quality already in the published literature that could help make this application more robust. Particular facets of the seismicity problem that could be developed include:

(1). The use of stochastic optimization procedures designed to find the 'universal optimum' partition.

(2). Incorporation of uncertainty in the hypocentral location and magnitude into the weighting scheme.

(3). The use of alternative clustering techniques to distinguish earthquakes that cluster around a physical structure from spatially homogenous background seismicity.

(4). Incorporating more information about the dimension and geometry of fault segments into the cluster analysis.

(5). Defining other characteristics of faulting in a region into the K-means procedure, for example, slip rate, stress drop, recurrence intervals of large earthquakes, etc.

9 CONCLUSIONS

The K-means cluster analysis method presented here is a novel approach to delineating seismic source zones when the principal source of seismotectonic information available is the hypocentral distribution. It is also a way of developing objectivity in the seismic source zonation process. The Line K-means algorithm and the weighting methodology in the point K-means algorithm both demonstrate attempts to avoid treating hypocentres as a homogeneous set of points, which is the traditional assumption of K-means. The success of the Line K-means algorithm in partitioning a set of line segments is encouraging. This variation of the algorithm could have applications beyond the seismic hazard example shown here, e.g. rock physics, image processing etc.

The current application to the Aegean demonstrates how this can be used as a foundation to seismic hazard analysis and attempts to address some of the inherent shortcomings of the conventional PSHA procedure. Though no single partition can be considered definitive, source models containing 20–30 zones emerge as the most appropriate for the Aegean. In particular, the 29-cluster model of shallow seismicity performed well for both line segment and point clustering, which suggests it is the most apposite for modelling Aegean seismicity. Models within the range stated represent a compromise between detail in identifying different seismotectonic regimes, and the requirement of having a sufficient number of earthquakes in each zone from which to determine the parameters of seismicity relevant for seismic hazard analysis and their uncertainties.

Several problems with the K-means method of zoning have also been highlighted. Inherent in this methodology (and other clustering techniques) are the problems of identifying the 'global' optimum partition and the most appropriate number of clusters. Attempts to resolve these problems include ensemble analyses, use of cluster quality indices that are uncorrelated to increasing K, and stochastic seismic hazard analysis. While the ensemble analysis may be sufficient to identify a near-optimum partition, it has not been possible to define a single value of K that is insensitive to the catalogue used. It is recommended that several partitions with different values of K be compared and reviewed by inspection, and if necessary incorporated into the uncertainty of the seismic hazard analysis by way of a logic tree or Monte Carlo method.

An implementation of the K-means algorithm to a catalogue of earthquakes for Aegea has been provided here. It demonstrates how a seismic source model may be created when only an earthquake catalogue is available. It is also useful for assessing the spatial variation in seismicity, and as a first order approximation of the boundaries between different seismotectonic regimes. Several recommendations on potential further development of the algorithm, and its relevance in the context of partitioning seismic sources, have also been made. It is hoped that this method will be developed further and prove a useful tool in the creation of seismic source models, particularly in parts of the world where seismicity is poorly constrained by the existing knowledge of local faults.

ACKNOWLEDGMENTS

We would like to thank Dr Christos Papaioannou for his insight into existing source models of the Aegean region. We are also grateful to two anonymous reviewers, whose comments helped improve this manuscript. This work was supported by a University of East Anglia (Faculty of Science) studentship.

REFERENCES

Abrahamson, N., 2006. Seismic hazard assessment: problems with current practice and future developments, in *First European*

Conference on Earthquake Engineering and Seismology, Geneva, Switzerland.

Allen, M.P., Evans, G.T., Frenkel, D., & Mulder, B.M., 1993. Hard convex body fluids, *Adv. Chem. Phys.,* **84,** 1–166.

Arias, A., 1970. A measure of earthquake intensity, in *Seismic Design for Nuclear Power Plants*, pp. 438–483, ed. Hansen, R.J., MIT Press, Cambrdige, Massachusetts.

Barani, S., Spallarossa, D., Bazzurro, P., & Eva, C., 2007. Sensitivity analysis of seismic hazard for Western Liguria (North Western Italy): a first attempt towards the understanding and quantification of hazard uncertianty, *Tectonophysics,* **435,** 13–35.

Barka, A., 1999. The 17 August 1999 Izmit earthquake, *Science,* **285,** 1858–1859.

Barka, A. *et al*., 2002. The surface rupture and slip distribution of the 17 August 1999 Izmit earthquake (M 7.4), North Anatolian Fault, *Bull. seism. Soc. Am.,* **92,** 43–60.

Beauval, C., Scotti, O., & Bonilla, F., 2006. The role of seismicity models in probabilistic seismic hazard estimation: comparison of a zoning and a smoothing approach, *Geophys. J. Int.,* **165,** 584–595.

Bender, B., & Perkins, D.M., 1987. SEISRISK III: a computer program for seismic hazard evaluation, in *USGS Open File Report*, USGS, Menlo Park.

Bradley, P.S., & Fayyad, U.M., 1998. Refining initial points for *K*-means clustering, in *Proceedings of the 15th International Conference on Machine Learning*, pp. 91–99, J. Morgan Kauffman, San Francisco.

Burton, P.W., Xu, Y., Qin, C., Tselentis, G.A., & Sokos, E., 2004. A catalogue of seismicity in Greece and the adjacent areas for the twentieth century, *Tectonophysics,* **390,** 117–127.

Calinski, R.B., & Harabasz, J., 1974. A dendrite method for cluster analysis, *Commun. Stat.,* **3,** 1–27.

Clarke, P.J. *et al*., 1997. Geodetic estimate of seismic hazard in the Gulf of Korinthos, *Geophys. Res. Lett.,* **24,** 1303–1306.

Cornell, C.A., 1968. Engineering seismic risk analysis, *Bull. seism. Soc. Am.,* **58,** 1583–1606.

Cramer, C.H., Petersen, M.D., & Reichle, M.S., 1996. A Monte Carlo approach in estimating uncertainty for a seismic hazard assessment of Los Angeles, Ventura, and Orange Counties, California, *Bull. seism. Soc. Am.,* **86,** 1681–1691.

Danciu, L., & Tselentis, G.-A., 2007. Engineering ground-motion parameters attenuation relationships for Greece, *Bull. seism. Soc. Am.,* **97,** 162–183.

Du, Q., Faber, V., & Gunzburger, M., 1999. Centroidal Voronoi Tessellations, *SIAM Rev.,* **41,** 637–676.

Ebel, J.E., & Kafka, A.L., 1999. A Monte Carlo approach to seismic hazard analysis, *Bull. seism. Soc. Am.,* **89,** 854–866.

Erdik, M., Biro, Y.A., Onur, T., Sesetyan, K., & Birgoren, G., 1999. Assessment of earthquake hazard in Turkey and neighbouring regions, *Annali di Geofisica,* **42,** 1125–1138.

Feng, Y., & Hamerly, G., 2006. PG-means: learning the number of clusters in data, in *Advances in Neural Information Processing Systems 19*, pp. 393–400, eds Schölkopf, B., Platt, J. and Hormann, T., MIT Press, Cambridge, MA.

Frankel, A., 1995. Mapping seismic hazard in the Central and Eastern United States, *Seismol. Res. Lett.,* **66,** 8–21.

Giardini, D., Wiemer, S., Fah, D., Deichmann, N., 2004. Seismic Hazard Assessment of Switzerland, Swiss Seismological Service, ETH Zurich, pp. 1–88

Grunthal, G., & Wahlström, R., 2001. Sensitivity of parameters for probabilistic seismic hazard analysis using a logic tree approach, *J. Earthq, Eng.,* **5,** 309–328.

Gutenberg, B., & Richter, C.F., 1944. Frequency of earthquakes in California, *Bull. seism. Soc. Am.,* **34,** 1985–1988.

Hamerly, G., & Elkan, C., 2003. Learning the k in k-means, in *Advances in Neural Information Processing Systems 16*, pp. 281–288, eds Thrun, S., Saul, L. K. and Schölkopf, B., MIT Press, Cambridge, MA.

Hartigan, J.A., 1975. *Clustering Algorithms*, John Wiley and Sons, New York.

Hartigan, J.A., & Wong, M.A., 1979. Algorithm AS136: a K-means clustering algorithm, *Appl. Stat.,* **28,** 100–108.

Hatzidimitriou, P.M., Papadimitriou, E.E., Mountrakis, D.M., & Papazachos, B.C., 1985. The seismic parameter b of the frequency-magnitude relation and its association with the geological zones in the area of Greece, *Tectonophysics,* **120,** 141–151.

Jain, A.K., Murty, M.N., & Flynn, P.J., 1999. Data clustering: a review, *ACM Comp. Surv.,* **31,** 264–322.

Jimenez, M.J. *et al*., 2001. Unified seismic hazard modelling throughout the Mediterranean region, *Bollettino di Geofisica Teorica ed Applicata,* **42,** 3–18.

Kaufman, L., & Rousseeuw, P.J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons Inc, USA.

Konstantinou, K.I., Kalogeras, I.S., Melis, N.S., Kourouzidis, M.C., & Stavrakakis, G.N., 2006. The 8 January 2006 earthquake (Mw 6.7) Offshore Kythira Island, Southern Greece: seismological, strong-motion, and macroseismic observations of an intermediate-depth event, *Seismol. Res. Lett.,* **7,** 544–553.

Kramer, S.L., 1996. *Geotechnical Earthquake Engineering*, Prentice Hall, New Jersey.

Krishna, K., & Murty, M.N., 1999. Genetic K-Means Algorithm, *IEEE Trans. Syst., Man Cybernet.–Part B: Cybernet.,* **29,** 433–439.

Krzanowski, W.J., & Lai, Y.T., 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering, *Biometrics,* **44,** 23–34.

Kuncheva, L.I., & Vetrov, D.P., 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization, *IEEE Trans. Pattern Anal. Mach. Intell.,* **28,** 1798–1808.

Likas, A., Vlassis, N., & Verbeek, J.J., 2003. The global k-means clustering algorithm, *Pattern Recog.,* **36,** 451–461.

Lu, Y., Lu, S., Foyouhi, F., Deng, Y., & Brown, S.J., 2004. Incremental genetic K-means algorithm and its application to gene expression data analysis, *BMC Bioinformatics,* **5,** doi:10.1186/1471-2105-1185-1172.

Makropoulos, K.C., & Burton, P.W., 1983. Seismic risk of circum-pacific earthquakes I: strain-energy release, *Pure Appl. Geophys.,* **121,** 247–267.

McGuire, R.K., 1976. FORTRAN computer program for seismic risk analysis, Open File Report 76-67, USGS, Menlo Park.

Musson, R.M.W., 1999. Probabilistic seismic hazard maps for the North Balkan region, *Annali di Geofisica,* **42,** 1109–1124.

Musson, R.M.W., 2004. Objective validation of source models, in *13th World Conference on Earthquake Engineering*, Vancouver, Canada, Paper No. 2492

Nabney, I.T., 2002. *Netlab: Algorithms for Pattern Recognition*, Springer, London.

Papadopoulos, G.A., Drakatos, G., Papanastassiou, D., Kalogeras, I., & Stavrakakis, G., 2000. Preliminary results about the catastrophic earthquake of 7 September 1999 in Athens, Greece, *Seismol. Res. Lett.,* **71,** 318–329.

Papadopoulos, G.A., Ganas, A., & Plessa, A., 2002. The Skyros earthquake (Mw 6.5) of 26 July 2001 and precursory seismicity patterns in the North Aegean Sea, *Bull. seism. Soc. Am.,* **92,** 1141–1145.

Papadopoulos, G.A., Karastathis, V.K., Ganas, A., Pavlidis, S., Fokaefs, A., & Orfanogiannaki, K., 2003. The Lefkada, Ionian Sea (Greece), Shock (Mw 6.2) of 14 August 2003: evidence for the characteristic earthquake from seismicity and ground failures, *Earth, Planets Space,* **55,** 713–718.

Papaioannou, C.A., & Papazachos, B.C., 2000. Time-independent and time-dependent seismic Hazard in Greece based on seismogenic sources, *Bull. seism. Soc. Am.,* **90,** 22–33.

Papazachos, B.C., 1990. Seismicity of the Aegean and surrounding area, *Tectonophysics,* **178,** 287–308.

Papazachos, B.C., Papaioannou, C.A., Papazachos, C.B., & Savvidis, A.S., 1997. *Atlas of Isoseismal Maps for Strong Shallow Earthquakes in Greece and Surrounding Area (426BC–1995),* 1st edn, P. Ziti and Co., Thessaloniki.

Papazachos, B.C., & Papazachou, C., 1997. *The Earthquakes of Greece*, 1st edn, P. Ziti & Co., Thessaloniki, Greece.

Papazachos, B.C., Papaioannou, C.A., Papazachos, C.B., & Savvidis, A.S., 1999. Rupture zones in the Aegean region, *Tectonophysics,* **308,** 205–221.

Pavlides, S., Papadopoulos, G., & Ganas, A., 2002. The fault that caused the Athens September 1999 Ms = 5.9 earthquake: field observations, *Nat. Hazards,* **27,** 61–84.

Peña, J.M., Lozano, J.A., & Larranaga, P., 1999. An empirical comparison of four initialization methods for the K-Means algorithm, *Pattern Recog. Lett.,* **20,** 1027–1040.

Reasenberg, P., 1985. Second-order moment of central California seismicity, 1969–1982, *J. geophys. Res.,* **90,** 5479–5495.

Reilinger, R. *et al*., 2000. Coseismic and postseismic fault slip for the 17 August 1999, M = 7.5, Izmit, Turkey earthquake, *Science,* **289,** 1.

Reilinger, R. *et al*., 2006. GPS constraints on continental deformation in the Africa-Arabia-Eurasia continental collision zone and implications for the dynamics of plate interactions, *J. geophys. Res.,* **111,** doi:10.1029/2005JB004051.

Shapira, A., 1983. Potential earthquake risk estimations by application of a simulation process, *Tectonophysics,* **95,** 75–89.

Sheng, W., & Liu, X., 2006. A genetic k-medoids clustering algorithm, *J. Heuristics,* **12,** 447–466.

Smith, W.D., 2003. Earthquake Hazard and risk assessment in New Zealand by Monte Carlo methods, *Seismol. Res. Lett.,* **74,** 298–304.

Stirling, M.W., McVerry, G.H., & Berryman, K.R., 2002. A new seismic hazard model for New Zealand, *Bull. seism. Soc. Am.,* **92,** 1878–1903.

Stiros, S.C., 1998. Historical seismicity, paleoseismicity and seismic risk in Western Macedonia, Northern Greece, *J. Geodyn.,* **26,** 271–287.

Tibshirani, R., Walther, G., & Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic, *J. Roy. Stat. Soc. (B),* **63,** 411–423.

Tranos, M.D., Papadimitrou, E.E., & Kilias, A.A., 2003. Thessaloniki-Gerakarou Fault Zone (TGFZ): the western extension of the 1978 Thessaloniki earthquake fault (Northern Greece) and seismic hazard assessment, *J. Struc. Geol.,* **25,** 2109–2123.

Travasarou, T., Bray, J.D., & Abrahamson, N.A., 2003. Empirical attenuation relationship for Arias Intensity, *Earthq. Eng. Struct. Dyn.,* **32,** 1133–1155.

Vamvakaris, D.A., Papazachos, C.B., Karagianni, E.E., Scordilis, E.M., & Hatzidimitriou, P.M., 2006. Small-scale spatial variation of the stress field in the back-arc Aegean area: results from the seismotectonic study of the broader area of Mygdonia basin (N. Greece), *Tectonophysics,* **417,** 249–267.

Ward, J.H., 1963. Hierarchical grouping to optimize and objective function, *J. Am. Stat. Assoc.,* **58,** 236–244.

Welling, M., & Khurihara, K., 2006. Bayesian K-Means as a "Maximization-Expectation" Algorithm, in *Proceedings of the 2006 SIAM Conference on Data Mining, Bethesda, Maryland*, pp. 474–478, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wells, D.L. & Coppersmith, K.J., 1994. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, *Bull. seism. Soc. Am.,* **84,** 974–1002.

Woo, G., 1996. Kernal estimation methods for seismic Hazard area source modelling, *Bull. seism. Soc. Am.,* **86,** 353–362.

Wright, T., Fielding, E., & Parsons, B., 2001. Triggered slip: Observations of the 17 August 1999 Izmit (Turkey) earthquake using radar interferometry, *Geophys. Res. Lett.,* **28,** 1079–1082.

Xie, X.L., & Beni, G., 1991. A validity measure for fuzzy clustering, *IEEE Trans. Pattern Anal. Mach. Intell.,* **13,** 841–847.

Zamani, A., & Hashemi, N., 2004. Computer-based self-organized tectonic zoning: a tentative pattern recognition for Iran, *Comp. Geosci.,* **30,** 705–718.

Zhang, B., Hsu, M., & Dayal, U., 1999. K-harmonic means: a data clustering algorithm, in *Software Technology Library*, pp. 1–26, ed. HPL-1999-124, Hewlitt Packard Research Laboratory, Palo Alto.