Detecting Ethnic Residential Clusters Using the Optimisation Partitioning Method

Seong-Yun Hong[1]

[1]School of Environment
The University of Auckland
Private Bag 92019
Auckland, New Zealand
Email: yun.hong@auckland.ac.nz

1. Introduction

To investigate the causes and consequences of ethnic residential clusters, it is essential to employ a reliable and valid method for locating them. In the past, researchers were able to rely on their eyes to distinguish the boundaries of ethnic enclaves, largely because minority populations were often segregated in small areas whose borders were fairly clear. Contemporary ethnic residential neighbourhoods, however, tend to be located across several suburbs and their exact boundaries are not as apparent as traditional inner-city ghettos. The delineation of such *ethnoburb*-like clusters has become somewhat vague and arbitrary, as there is often no clear-cut point on the continuum of population density that draws a line between clusters and non-clusters. This consequently necessitates the use of a more objective method to detect ethnic residential clusters in modern cities.

One possible approach is the application of local statistics (or scan statistics) such as the local G* statistic (Getis and Ord 1992; Ord and Getis 1995). However, given that most existing methods were essentially developed to detect unusual (and unexpected) concentrations of events under the hypothesis of randomness, these may not be the most appropriate approaches to identify the extent of ethnic residential clustering. In addition, these methods typically require specification of the shape and size of clusters before their application, which is also a significant limitation. Although there are a few algorithms designed to detect arbitrarily-shaped clusters (see, for example, Patil and Taillie 2004; Tango and Takahashi 2005), they are not feasible when the sizes of the clusters are large.

The approach proposed here is a variation of the optimisation partitioning method, which performs reasonably well when the number of clusters and their approximate locations are known (Everitt et al. 2001). As will be demonstrated, the clustering procedure is simple and straightforward, and it directly addresses the question of where a particular cluster ends.

2. Clustering Procedures

Suppose that the study region is divided into n census tracts, $\Omega = \{x_1, x_2, x_3 \ldots x_n\}$, and the aim is to identify a particular number of groups whose data values are distinctively larger than those of the remaining census tracts. Each group consists of geographically continuous census tracts, $A_i = \{x_{i1} \ldots x_{in}\}$, where $i = 1 \ldots g$, and they do not overlap with one another. For convenience, let A_0 denote a set of residuals that are not included in A_i, so $\Omega = A_0 \cup A_1 \cup \ldots \cup A_g$. The fundamental idea behind this method is that the quality of a given clustering can be represented by numerical indices and the best possible subsets can be found by optimising the index values. Among a variety of numeri-

cal clustering measures suggested by statisticians since the 1960s, this study employs one of the simple criteria, the within-group sum of absolute deviations, whose minimum value indicates the best solution:

$$w = \sum_{i=0}^{g} \sum_{j=1}^{n_i} a_{ij} \left| \bar{b}_i - b_{ij} \right| \quad (1)$$

where n_i is the number of census tracts in A_i, a_{ij} is the weight of the corresponding census tract, and b_{ij} is the data value of interest. In the examples in the following sections, for instance, the size of the census tracts is used as the weight and the population density is used as the data value. $\bar{b}_i$ refers to the weighted mean of the data values in A_i, which indicates the overall population density of the cluster in this case.

Perhaps the most straightforward way to identify a set of A_i that minimises w is to investigate all possible combinations and then choose the best one. When n is small and the spatial structure is simple, this approach is feasible and even guarantees the identification of a global optimum (if there is one). Not surprisingly however, it rapidly becomes impractical as n increase due to its computationally expensive nature, even with today's computing power (Everitt et al. 2001).

The *hill-climbing* algorithm is an alternative that overcomes the limitation of examining all possible combinations, and the below outlines the clustering procedures implemented in the present study:

1. Choose a census tract that is suspected to be part of a cluster and calculate its total within-group absolute deviations, w.
2. Combine the current set of a cluster with its neighbouring census tracts in all possible configurations and compare the w values.
3. Replace the set with the one that minimises w and repeat (b) and (c) until no further improvements can be made.
4. Repeat the procedures above for each cluster of interest.

Although there is a vast amount of literature devoted to the selection of an initial set, such complicated mathematical calculations are not required in the present context because, as mentioned earlier, the approximate locations of ethnic clusters are often clearly given in the map. This clustering algorithm provides convincing results when applied to hypothetical data sets in the following section as well as for the Korean population data in Section 3.2.

3. Examples

3.1 Hypothetical data

In this section, the optimisation clustering method is applied to three hypothetical data sets shown in Figure 1. The results are then compared to those from the local G* statistic to demonstrate the advantages and limitations of the proposed approach.

The first two data sets were generated from an exponential distribution with λ = 0.005, arranged in basic grids of 10-by-10 metres, and contain a spherical- and a linear-shaped cluster, respectively. To illustrate how the performance of the present algorithm is affected by the initial configurations, four different starting points were chosen for each data set: two inside the cluster and two outside. The first two clustering outcomes in Figure 2 and Figure 3 demonstrate that when the starting points were located well within the clusters of interest, the nearby cells with high data values were effec-

tively grouped as intended and the *w* values were reasonably minimised. However, when the starting points were placed outside of the clusters (i.e., Figure 2 (c) – (d) and Figure 3 (c) – (d)), the algorithm seems to be trapped by small local variations in the data values, failing to recognise the obvious clusters. These results emphasise the point that the choice of where to begin is crucial for the reliability and accuracy of the method and that the use of a priori information can greatly enhance the performance of the algorithm.

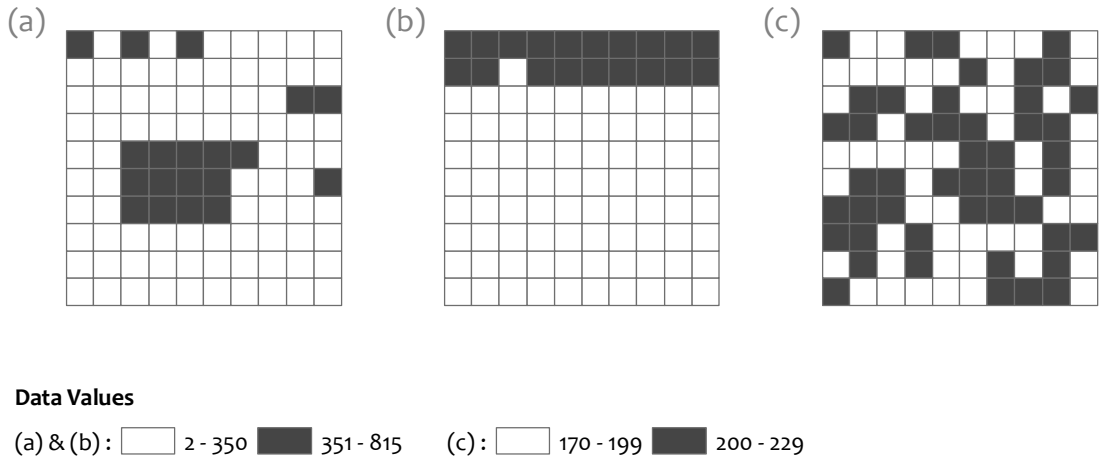

Figure 1. Hypothetical data sets with different spatial patterns

Figure 2 (e) and Figure 3 (e) display the extent of the most likely clusters found using the local G* statistic along with the centres of other statistically significant clusters (i.e. $z \geq 1.96$). The results were generally analogous to those obtained using the optimisation clustering method, but they included some units containing very low data values. The degree of such false detection was more severe in the second data set in which the cluster has a linear shape, implying that this approach might be less reliable in determining the extent of non-spherical clusters compared to the proposed approach.

In contrast to the first two data sets, the third example follows a Poisson distribution with $\lambda = 200$, and there are no apparent clusters present. As with the previous examples, four different starting points were chosen for the optimisation clustering method. However, since there are no notable concentrations of data values, the four points were randomly selected (Figure 4 (a) – (d)). The results highlighted on the grids, together with the box plots, indicate that the algorithm performed poorly in this case – it only captured small, insignificant local variations in the data set. The local G* method, by contrast, successfully revealed that there is no subtle clustering of the high values (Figure 4 (e)).

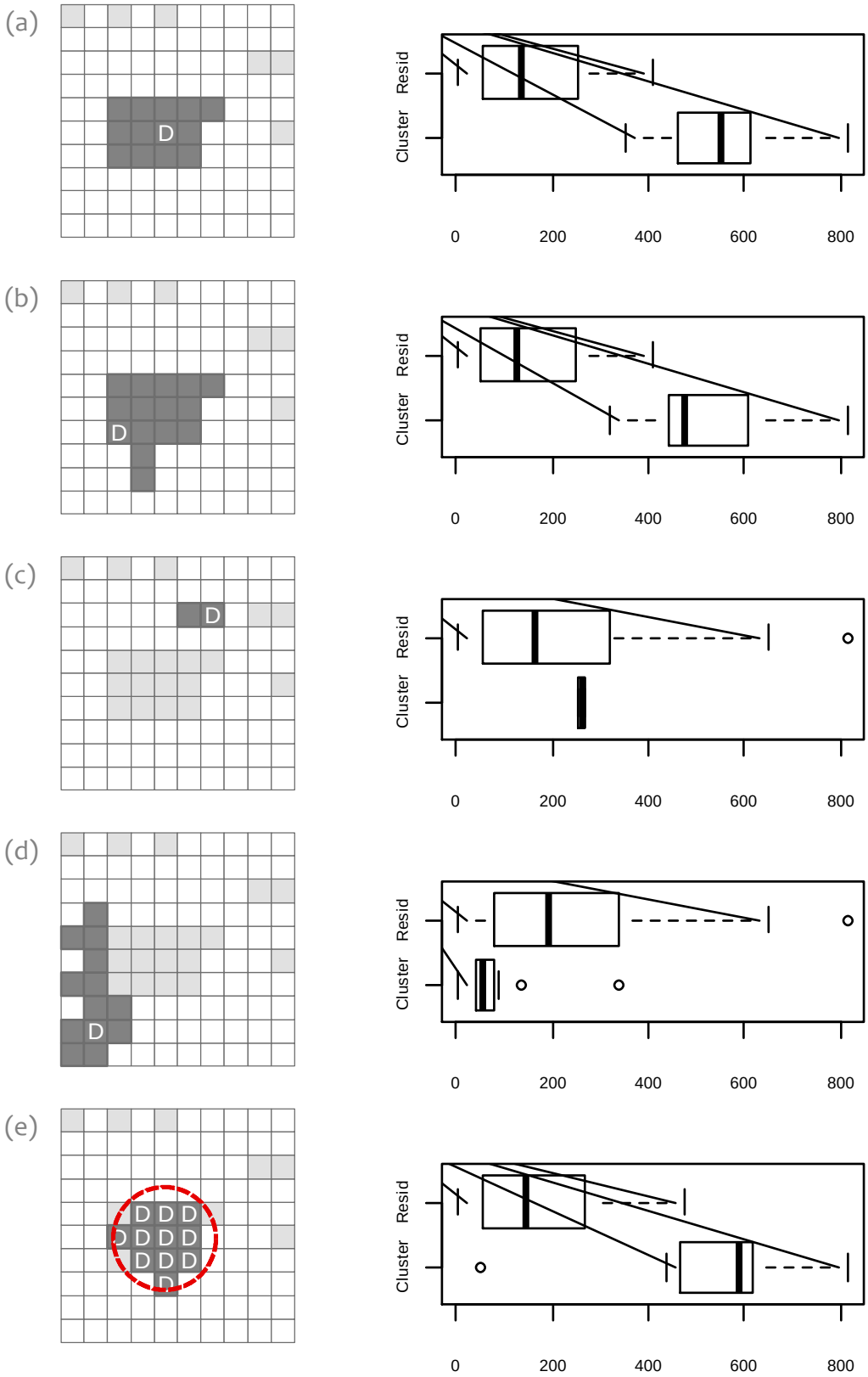

Figure 2. Clustering results for the first data set: (a) – (d) using the proposed clustering algorithm with different starting points; (e) using the local G statistic

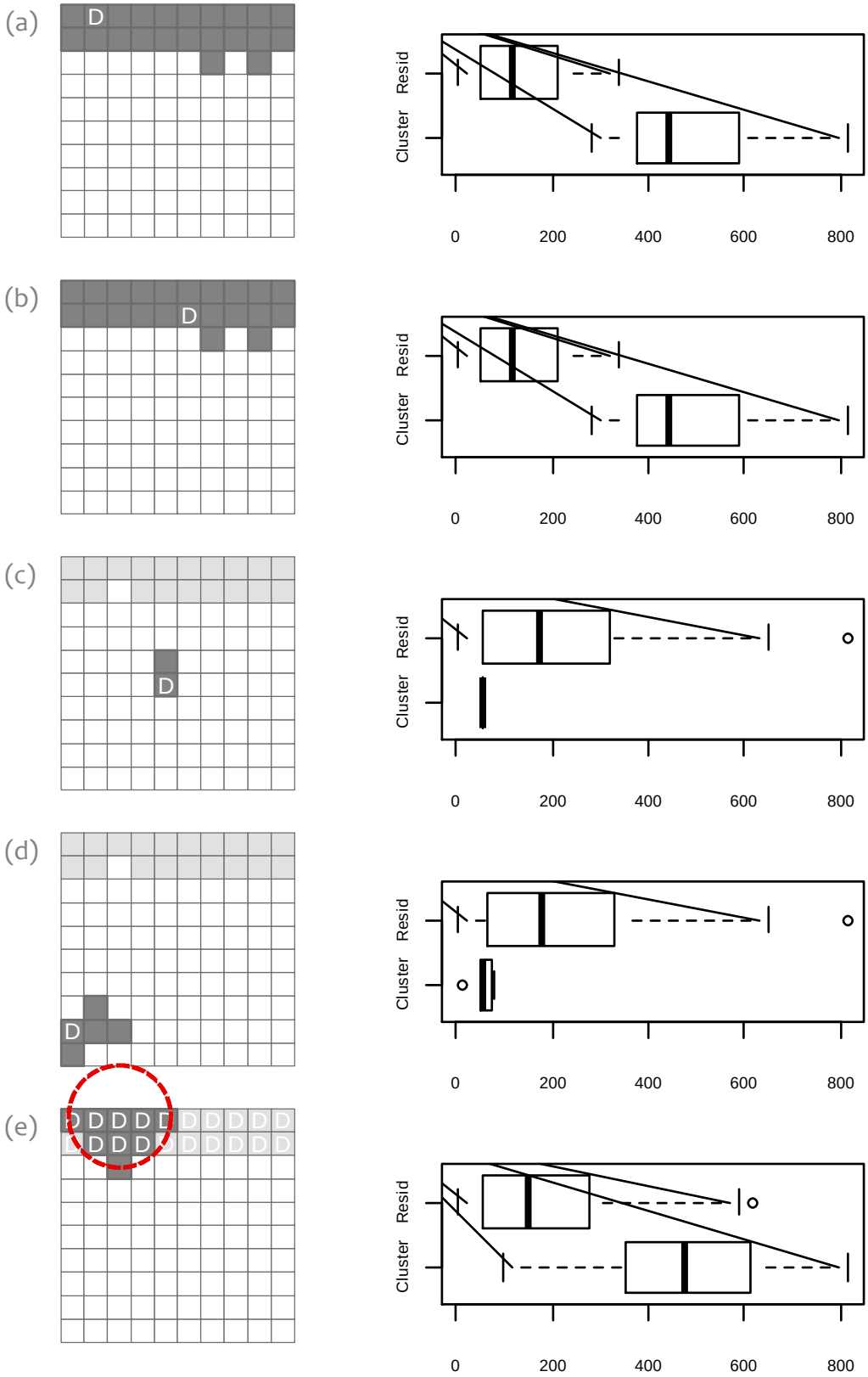

Figure 3. Clustering results for the second data set: (a) – (d) using the proposed clustering algorithm with different starting points; (e) using the local G statistic

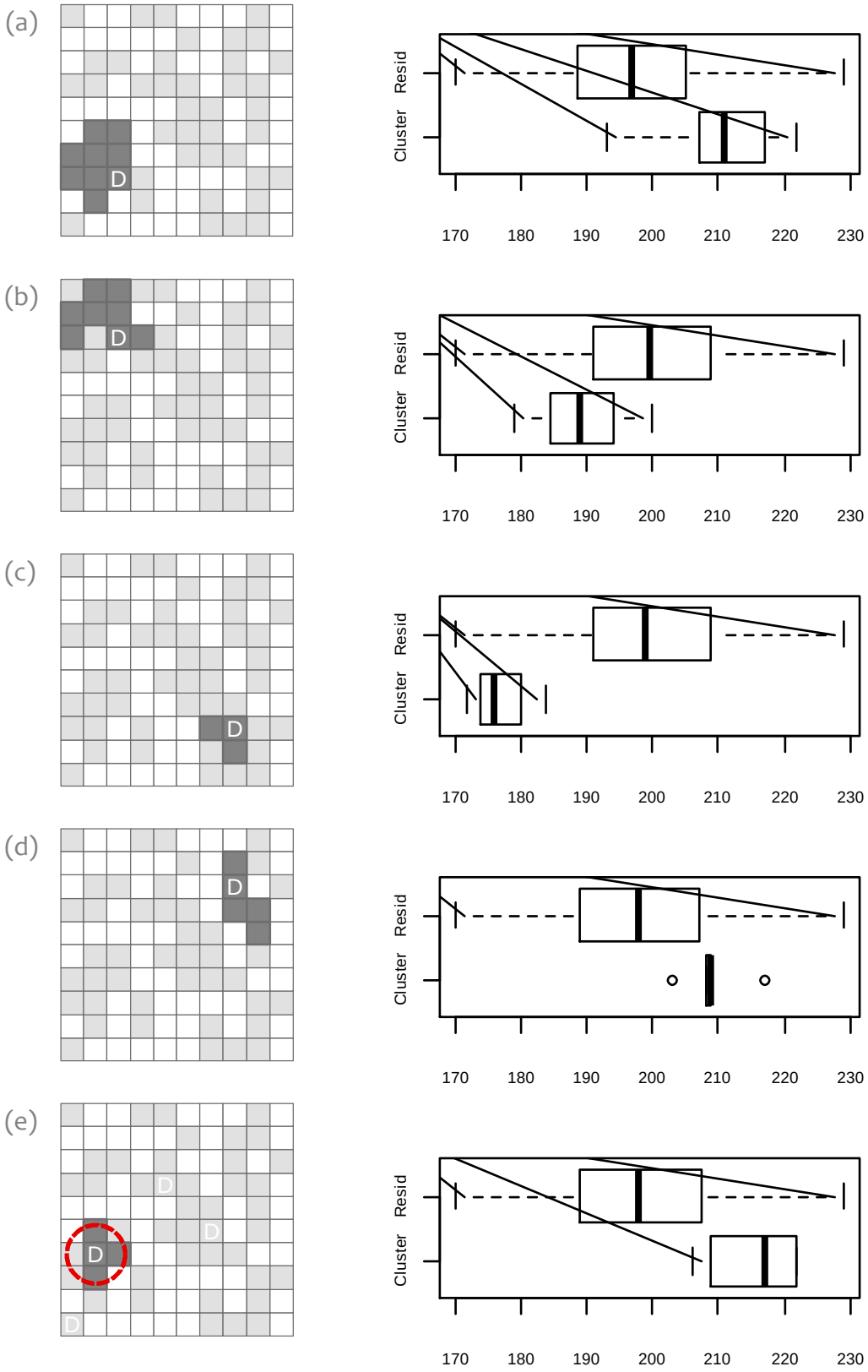

Figure 4. Clustering results for the third data set: (a) – (d) using the proposed clustering algorithm with different starting points; (e) using the local G statistic

3.2 Korean residential clusters in Auckland, New Zealand

In this section, the proposed algorithm is applied to Korean population data in the 2006 New Zealand Census. The geographic distribution of Koreans in Auckland, the largest urban area in the country, is similar to that of the typical *ethnoburb* described by Li (1997): they are relatively clustered in certain parts of the city, but not isolated from the other population groups. They occupy large areas but with low population density (see Table 1 and Figure 5).

Table 1. Global measures of segregation for Koreans in Auckland, 2006

	Value
Spatial dissimilarity index ($\widetilde{D}$)	0.465
Spatial exposure index ($\tilde{P}^*$)	
Korean to European	0.644
Korean to Maori	0.198
Korean to Pacific peoples	0.080
Korean to other Asian	0.426

The choropleth map of Korean population density in Figure 5 provides a visual impression that Koreans are mainly clustered in North Shore City and Auckland Central. Most high- and medium-density census tracts (i.e., black and dark grey areas) are concentrated in the northern part of the region and around the CBD areas. In addition to these evident agglomerations, several tracts with medium Korean population density are also concentrated in the eastern and western parts of the map, which can possibly be classified as small clusters as well.

The circles on the map indicate the locations of possible clusters. For each cluster, up to three census tracts were chosen as starting points for the clustering algorithm. The selected tracts either have distinctively higher population density compared to others or are located approximately at the centre of each cluster. The clustering algorithm was applied to all possible starting configurations (i.e., nine different sets) and one that minimised the overall within-group sum deviation, *w*, was chosen. The box plot on the bottom right of Figure 5 shows that the proposed algorithm effectively identified all high- and medium-density census tracts and grouped them as clusters.

4. Summary

In general, the optimisation clustering method is the same as other local or scan statistics in that it attempts to identify a set of geographically close observations that have high (or low, depending on the context) values compared to the rest of the data. What distinguishes this algorithm from others is that it does not require defining ‘close’ or ‘high’ prior to its application, providing a significant advantage over other approaches. It is important to note, however, that the validity of the clustering results is radically affected by the initial configuration because an invalid classification from an earlier stage cannot be corrected later. Nonetheless, the examples in the previous sections illustrate that this method generates convincing outcomes when the starting points are well chosen. Thus, the proposed approach can be useful in determining the extent of *ethnoburb*-like residential clusters where the approximate locations of clusters are often apparent but their exact boundaries are not.

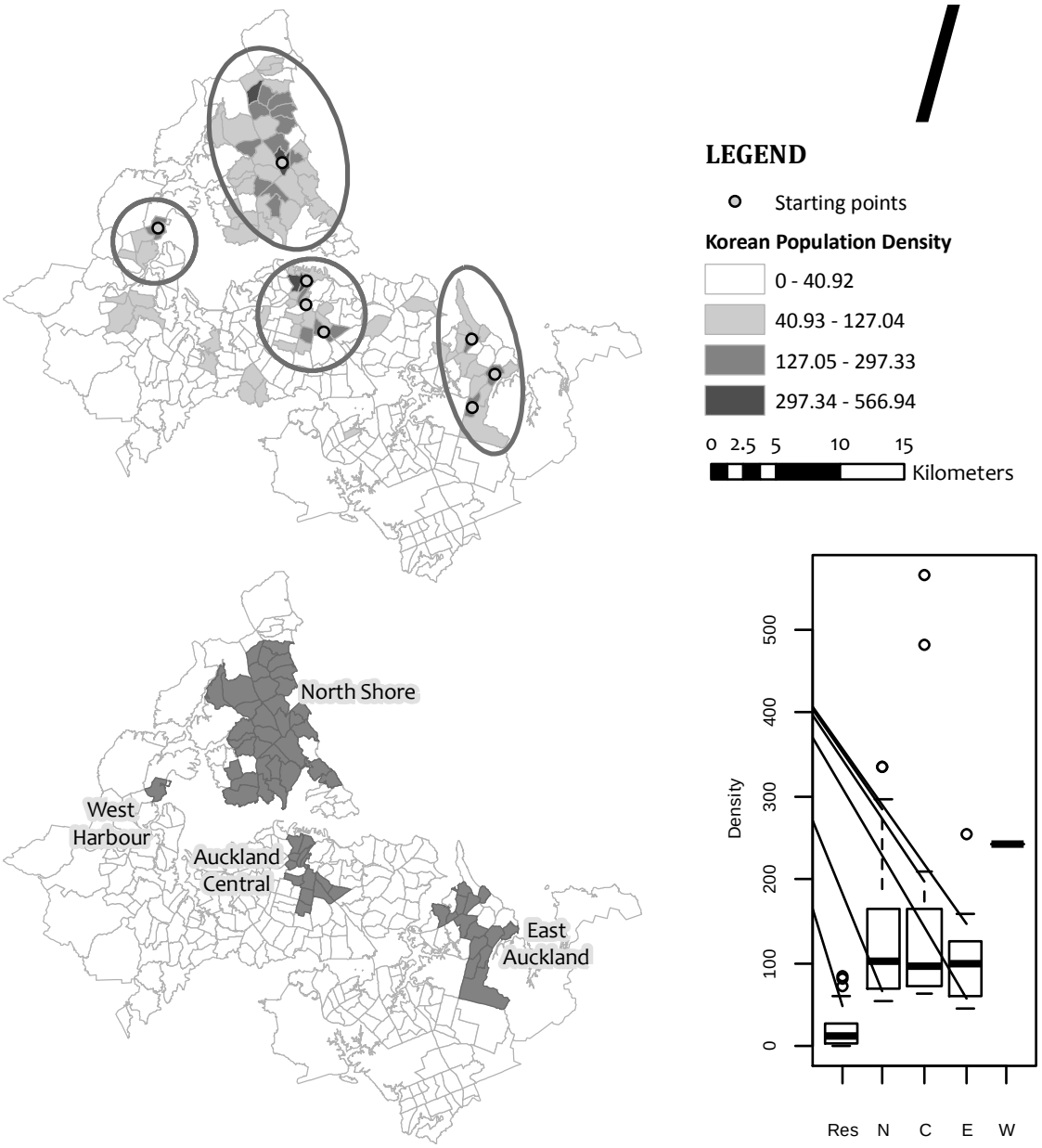

Figure 5. Clustering results for the Korean population in Auckland, 2006

References

Everitt B, Landau S and Leese M, 2001, *Cluster analysis* (4th ed.). London: Arnold; Oxford University Press.

Getis A and Ord JK, 1992, The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24: 189-206.

Ord JK and Getis A, 1995, Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, 27: 286-306.

Patil GP and Taillie C, 2004, Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11(2): 183-197.

Tango T and Takahashi K, 2005, A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4(11).

Li W, 1997, Ethnoburb versus Chinatown : Two Types of Urban Ethnic Communities in Los Angeles. Colloque "*les problèmes culturels des grandes villes*", http://www.cybergeo.eu/index1018.html.