Fuzzy Geographically Weighted Clustering

G. A. Mason[1], R. D. Jacobson[2]

[1] University of Calgary,
Dept. of Geography
2500 University Drive NW
Calgary, AB, T2N 1N4
Telephone: +1 403 210 9723
Fax: +1 403 282 6561
Email: g.mason@ucalgary.ca

[2] University of Calgary,
Dept. of Geography
2500 University Drive NW
Calgary, AB, T2N 1N4
Telephone: +1 403 220 6192
Fax: +1 403 282 6561
Email: dan.jacobson@ucalgary.ca

1. Introduction

Geodemographic analysis has been described as "the analysis of spatially referenced geodemographic and lifestyle data" (See and Openshaw, 2001, p.269) It is widely used in the public and private sectors for the planning and provision of products and services. Geodemographic analysis often uses clustering techniques which are used to classify the geodemographic data into groups, making the data more manageable for analysis purposes. Clustering identifies a number of geodemographic groups (clusters), each group having a particular geodemographic profile. Each geographical area under consideration is then assigned to a group based on its similarity to the group profile.

Fuzzy clustering offers a method of clustering that uses the principles of fuzzy logic to calculate a membership value for each subject in each of the groups. So rather than assigning a geographical area to a single group, each area is allocated a membership value in each of the groups (clusters), thus helping to overcome the issues of ecological fallacy. The fuzzy clustering algorithm typically used in geodemographic analysis is Bezdek's fuzzy c-means clustering algorithm known as FCM (Bezdek et. al., 1984). Fuzzy geodemographic analysis using FCM has been investigated by Feng and Flowerdew (1998, 1999), and See (1999), but has received scant attention since - an exception being the recent investigation by one of the authors (Mason, 2006).

This paper proposes a modification to the fuzzy clustering algorithm to incorporate geographical effects, suitable for geodemographic analysis.

2. Fuzzy Clustering with Neighbourhood effects

Openshaw (1998, p104) pointed out that "geodemographics is simple minded, in that it assumes that residential area type 27 is the same and behaves in the same way wherever it happens to be located". He then asked "But what happens if Type 27 areas respond differently depending on their map location?"

In order to overcome this shortcoming, and to incorporate geographical effects into geodemographic analysis, Feng and Flowerdew (1998) proposed an extension to the fuzzy clustering technique, which provides for the *ex post facto* adjustment of the cluster membership

values based on “neighbourhood effects”. The neighbourhood effects incorporate geography into the model. The neighbourhood effects formula adjusts the cluster membership as shown in equation 1.

$$\mu'_i = \alpha \mu_i + \beta . \frac{1}{A} \sum_j^n w_{ij} \mu_j \quad (1)$$

where μ_i' is the new cluster membership of area i
μ_i is the old cluster membership of area i

$$\alpha + \beta = 1 \quad (2)$$

where α and β are scaling variables to affect the proportion of the original membership vs. the weighted (calculated) membership

$$w_{ij} = \frac{p_{ij}^{\ b}}{d_{ij}^{a}} \quad (3)$$

where p is the length of the common boundary between i and j;
d is the distance between i and j ;
a and b are user definable parameters

A is a factor to scale the "sum" term to the range 0 to 1.

3. Fuzzy Geographically Weighted Clustering

This research continues on the theme of incorporating geography into geodemographic analysis, by positing that the incorporation of neighbourhood effects into the clustering algorithm (not just an *ex post facto* manipulation of the cluster memberships) will have an influence on the cluster centre values, thus creating “geographically aware” clusters. Additionally, it addresses two shortcomings of the neighbourhood effect model as proposed by Feng and Flowerdew (1998). This research borrows from the principles of geographical spatial interaction (Birkin and Clarke, 1991), by incorporating a basic spatial interaction model into the weighting of the memberships.

A diagrammatic overview of the concept is presented in Figure 1.

3.1 Fuzzy Geographically Weighted Clustering Algorithm

A previous investigation by one of the authors (Mason, 2006) identified that the cluster centres are not only sensitive to changes in variable values but are also sensitive to the amount of fuzziness applied via a fuzzy exponent configuration variable. This observation led to the hypothesis that the application of the neighbourhood effects as presented by Feng and Flowerdew (1998), if included in the fuzzy clustering algorithm, would also affect the cluster centres, making them “geographically aware”.

The fuzzy c-means algorithm (Bezdek et al., 1984), is an iterative algorithm that re-calculates the cluster centres and the associated membership values on each iteration, until optimality is

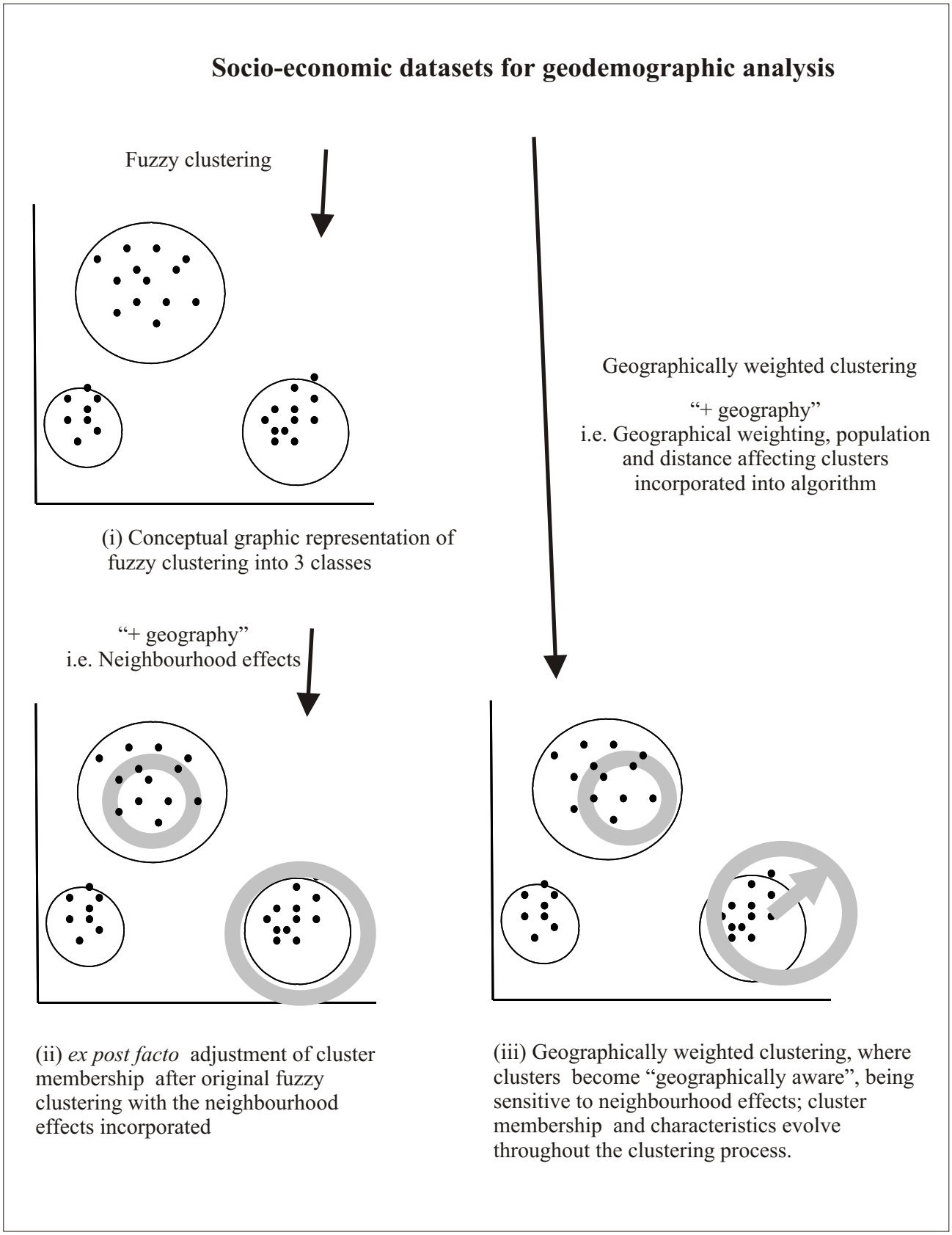

Figure 1: Conceptual overview of Geographically Weighted Clustering

achieved. The proposed modification to the algorithm adds an additional step to each iteration, that applies a weighting to the membership values using the technique noted in 3.2 below, subsequent to the "standard" calculation of the membership values as prescribed by Bezdek.

3.2 Spatial Interaction Effects

Feng and Flowerdew's neighbourhood effects have some limitations. First they ignore the effects of areas which have no common boundary (See, 1999). Second, they exclude the effects of population - a key geodemographic consideration. To overcome these limitations, a modified version of the cluster membership adjustment is proposed that incorporates a spatial interaction effect model similar to that discussed by Birkin and Clarke (1991). This proposed model calculates the influence of one area upon another as the product of the populations of the areas. A distance decay effect is implemented in the divisor. This effect is implemented through the weighting factor as described in equation 6.

The adjusted cluster membership for the fuzzy geographically weighted clustering algorithm, which is calculated in each iteration of the fuzzy clustering algorithm, is shown in equation 4:

$$\mu'_i = \alpha \mu_i + \beta . \frac{1}{A} \sum_j^n w_{ij} \mu_j \tag{4}$$

where μ_i' is the new cluster membership of area i
μ_i is the old cluster membership of area i

$$\alpha + \beta = 1 \tag{5}$$

where α and β are scaling variables to affect the proportion of the original membership vs. the weighted (calculated) membership

$$w_{ij} = \frac{(m_i m_j)^b}{d_{ij}^a} \tag{6}$$

where m_i, m_j are the population of areas i and j respectively
d_{ij} is the distance between i and j
and a and b are user definable parameters

and A is a factor to scale the "sum" term, and is calculated across all clusters, ensuring that the sum of the memberships for a given area for all clusters is equal to one.

4. Results

The Fuzzy Geographically Weighted Clustering (FGWC) algorithm was run against a test dataset of socio-economic demographic variables, using a=1, b=1, and β values from 0.0 to 0.5. A seven cluster scenario with fuzzy exponent of 1.2 was used. The aim was to determine how the cluster centres and cluster membership values change based on the application of the new spatial interaction weighting factor in the algorithm.

4.1 Nearest Cluster Distance

Taking the cluster centre for the β =0.0 scenario as the base (this is the equivalent of the normal fuzzy clustering) a calculation was made of the multidimensional euclidean distance (based on standardized variable values) to the nearest cluster of each of the β=0.1, β=0.2, β=0.3, β=0.4 and β=0.5 scenarios. This distance measure is often used as a cluster similarity measure and is used here to show how far the cluster centre has moved as a result of the spatial membership weighting.

	α=0.9 β=0.1	α=0.8 β=0.2	α=0.7 β=0.3	α=0.6 β=0.4	α=0.5 β=0.5
c1 ; α=1.0	0.074	0.091	0.099	0.161	0.217
c2 ; α=1.0	0.033	0.040	0.133	0.116	0.179
c3 ; α=1.0	0.026	0.037	0.085	0.113	0.182
c4 ; α=1.0	0.178	0.170	0.068	0.180	0.199
c5 ; α=1.0	0.742	1.040	1.241	1.445	1.621
c6 ; α=1.0	0.154	0.177	0.190	0.305	0.355
c7 ; α=1.0	0.536	0.291	0.277	0.406	0.560

Table 1: Distance of nearest cluster centre for β=0.0 to the other β scenario's, for each of the seven clusters

We can see in Table 1 that as a general trend, for increasing β (decreasing α), the further away the nearest cluster is, indicating a greater difference in the characteristics defined by that cluster centre. An illustration of the movement of the cluster centre across changing β is shown in Figure 2.

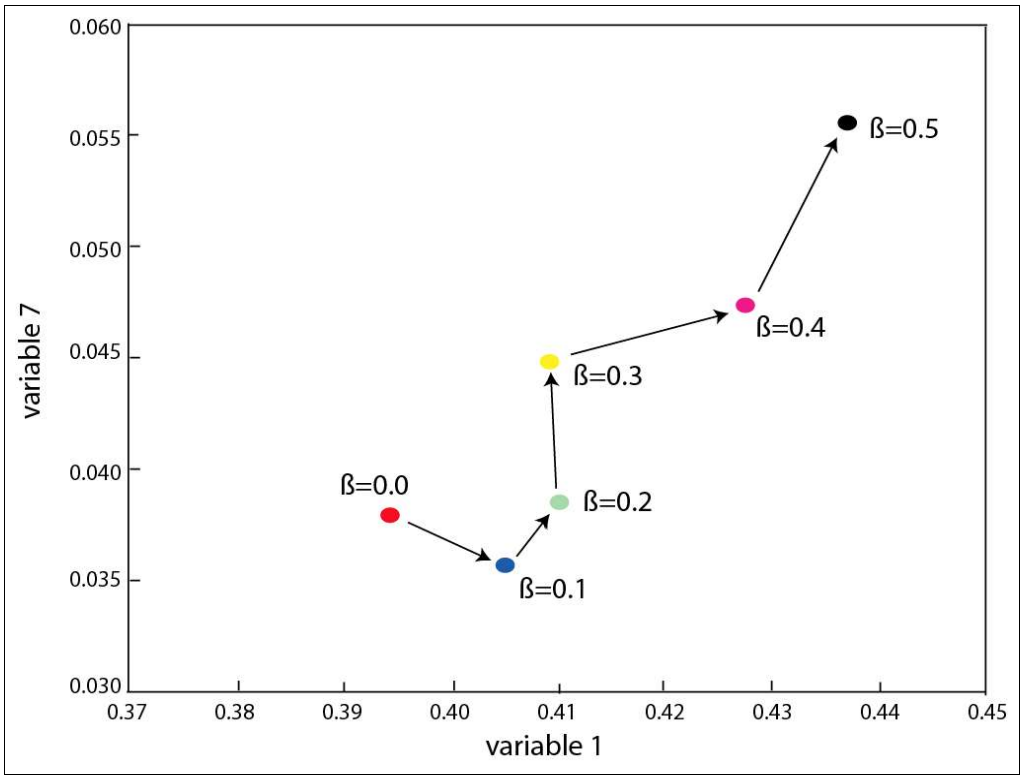

Figure 2: The trajectory of a cluster centre across increasing β for variables 1 and 7.

4.2 Moran's I

In the consideration of the effects of the FGWC algorithm on spatial autocorrelation, it was hypothesized that spatial autocorrelation of the cluster membership values would increase with increasing β, since the weighting factor incorporates the influence of local areas. To test this, Moran's I was calculated for each cluster in each scenario to identify how the spatial auto-correlation changes with increasing β. For each β, the totals across all clusters were also calculated as an indicator of the total spatial autocorrelation.

	β=0.0	β=0.1	β=0.2	β=0.3	β=0.4	β=0.5
c1	0.2217	0.2867	0.1541	0.3456	0.2502	0.3408
c2	0.3110	0.2690	0.3124	0.4170	0.3472	0.4010
c3	0.2583	0.2290	0.4166	0.3187	0.4018	0.4368
c4	0.3851	0.1795	0.2996	0.2176	0.2122	0.1905
c5	0.1234	0.0881	0.2352	0.4469	0.2609	0.5521
c6	0.1071	0.3485	0.1883	0.1843	0.3792	0.4438
c7	0.1419	0.3632	0.3612	0.2125	0.5109	0.2539
TOTAL	1.5485	1.7640	1.9674	2.1426	2.3624	2.6189

Table 2: Moran's I calculated for each set of cluster memberships, across varying β

As can be seen in Table 2, spatial autocorrelation increases for increasing β.

4.3 Standard Deviation of Membership Values

The standard deviations of the membership values for each cluster (labelled c1 through c7) was calculated across the values of β. The results are shown in Table 3.

This shows that overall the memberships are getting less distinct as the spatial interaction effect of the surrounding areas weighs more heavily on the resultant membership values. We also can see that the maximum membership value is decreasing with β.

	β = 0.0	β = 0.1	β = 0.2	β = 0.3	β = 0.4	β = 0.5
c1	0.3390	0.2997	0.2441	0.2432	0.1875	0.1172
c2	0.3095	0.2524	0.2565	0.2474	0.1991	0.1570
c3	0.3418	0.2988	0.2457	0.2262	0.2120	0.1869
c4	0.3430	0.2222	0.2687	0.2196	0.1406	0.0962
c5	0.1899	0.2861	0.2623	0.2096	0.1658	0.1629
c6	0.3111	0.2935	0.1955	0.1928	0.2217	0.1907
c7	0.2441	0.2799	0.2657	0.1588	0.1945	0.1554
max mem	1.0000	0.9588	0.9221	0.8799	0.8326	0.7659

Table 3: Standard deviations and maximum membership values of memberships across β

5. Conclusion

The Fuzzy Geographically Weighted Clustering algorithm offers a geographically aware alternative to a regular clustering algorithm which provides the capability to apply population and distance effects into a geodemographic cluster analysis. The resultant cluster centres are different than the unweighted scenario reflecting the application of the spatial interaction effects. The spatial autocorrelation of the cluster membership values increase as the spatial interaction

weighting is increased, and the membership values indicate an increasing homogenization of the clusters, as measured by the standard deviation of the membership values. There is considerable potential for further research into this technique and for the application to real world scenario's.

6. References

Bezdek, J.C., R. Ehrlich, et al. (1984) FCM: the fuzzy c-means clustering algorithm, *Computers and GeoSciences* 10: 191-203

Birkin, M and G. P. Clarke (1991) Spatial Interaction in Geography, *Geography Review*, 4(5), pp 16-24

Feng, Z. and R. Flowerdew (1998), Fuzzy Geodemographics: a contribution from fuzzy clustering methods. In S. Carver (ed) *Innovations in GIS 5*. London: Taylor & Francis.

Feng, Z. and R. Flowerdew (1999) The use of fuzzy classification to improve geodemographic targeting, in Gittings, B (ed.) *Innovations in GIS, Vol 6*. London: Taylor & Francis.

Mason, G. (2006), A Methodological Investigation of Fuzzy Geodemographic Clustering and Visualization, *Unpublished Thesis, University of Leeds*, [online] http://homepages.ucalgary.ca/~gamason/MSc/MScThesis.html

Openshaw, S. (1998), Towards the Marketing System that Thinks, *Institute of Direct Marketing Lecture*, [online] http://www.geog.leeds.ac.uk/presentations/98-8/tsld104.htm

See, L. (1999) Geographical applications of fuzzy logic and fuzzy hybrid techniques, *PhD dissertation*, School of Geography, University of Leeds

See L. and S. Openshaw (2001). Fuzzy geodemographic targeting. In G. Clarke, M. Madden (Eds.) *Regional Science in Business*. Berlin:Springer. pp 269-282.