Despatialising geodemographic propensities.

Martin Callingham

School of Geography, Birkbeck College, University of London
Malet Street, London WC1E 7HX
Tel: +44 (0)20 7631 6473 +44 (0) 20 7439 4377
Email: m.callingham@geography.bbk.ac.uk

1. Summary

This paper shows that the standard way of calculating market propensities for geodemographic system can give incorrect results due to the neglecting of the spatial effect.

2. Background

Geodemographics are a way of classifying small areal units primarily on their census counts across a number of variables (Webber & Craig, 1978).

Classifications of this type found commercial use in the early 1980's when it was discovered that general behaviour, including market propensities, could be ascribed to each areal type (Bermingham, McDonald & Baker, 1979).

The standard way of attributing propensities to geodemographic classifications is by obtaining a profile of users (across a gedemographics system) and comparing it to the national profile (Harris, Sleight & Webber, 2005).

3. Purpose

This paper shows that the standard calculation of market propensities for geodemographic types using customer postcodes collected in a retail outlet is fundamentally flawed.

The problem lies in neglecting the two facts:

- geodemographic types spatially auto correlate
- distance is a powerful factor affecting propensity to visit a store.

The standard method of measuring propensities ignores the distance effect and therefore inadvertently ascribes higher propensities to geodemographic types that are on average nearer.

4. Data used

The geodemographic system used in this work is that which was produced by the Office of National Statistics and is known as the Output Area Classification or OAC for short (Vickers & Rees, 2006). The classification is a threefold hierarchical classification; this work is based on the 7 Super Groups which have been given loosely descriptive names as shown in Table 1.

Table 1. Names for OAC Super Groups

Super Group	Name
1	Blue Collar Workers
2	City Living
3	Countryside
4	Prospering Suburbs
5	Constrained by Circumstances
6	Typical Traits
7	Multicultural

The data used in this work is 48,109 customer postcodes of a restaurant chain which has 26 outlets that are nationally distributed. The customer postcodes were collected in store and verified against the Office of National Statistics Post Code Directory and the OAC code and grid reference (present on the Directory) attached.

5. Preliminary Analysis

The numbers of customers of each type were counted and percentaged on the total. These percentages were then indexed on the comparable national percentages of people of each geodemographic type. This is the standard method of calculating propensities from customer postcode files. The propensities are shown in Figure 1; an index of 1 means a propensity is equal to the national average.

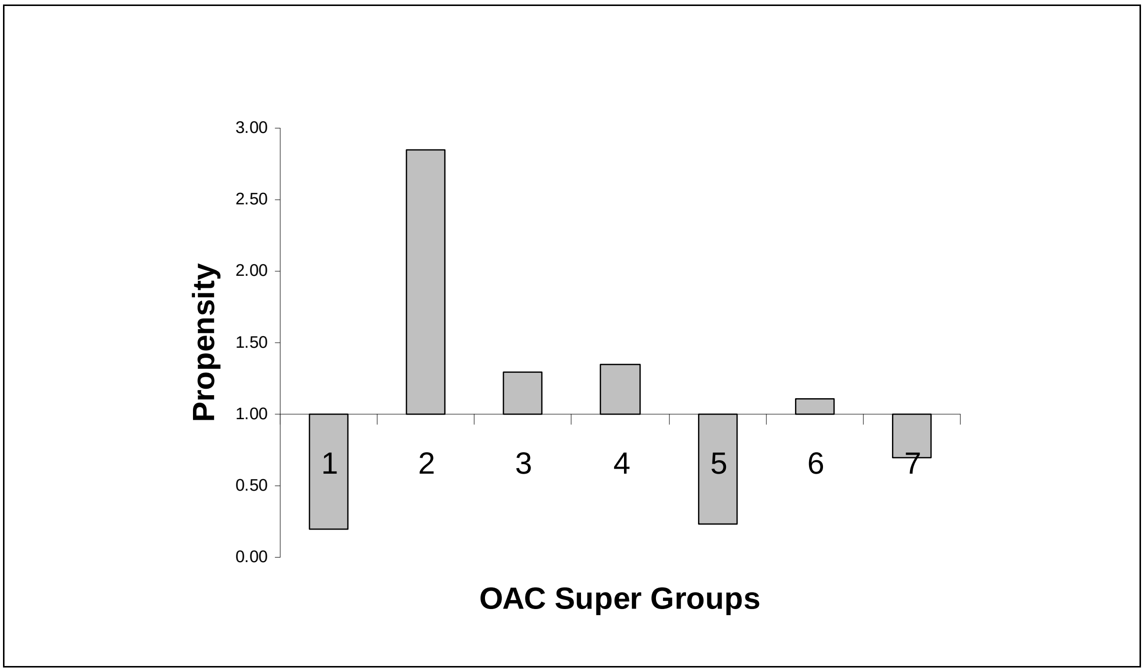

Figure 1. Chart of relative propensity to visit a restaurant by OAC Super Group

There are big differences in the propensity of people to visit this type of restaurant from different OAC Super Groups. It varies from almost 3.0 times the national average (Super Group 2 - *City Living*) to virtually 0 (Super Groups 1 - *Blue Collar Workers* and 5 - *Constrained by Circumstances*).

6. The influence of distance from the restaurant

Each customer was allocated to its nearest restaurant and its distance calculated and categorised into a distance band. This enabled a count of customers in each distance band to be made.

The catchment areas for each restaurant were similarly delineated and the numbers of residents in each distance band summed from 2001 census data.

The probability of a person living in a given distance band visiting the restaurant was calculated by dividing the customer counts by their equivalent population counts. As the customer count came from a sample of customers, the calculated set of probabilities were relative ones: for convenience these numbers were multiplied by 100. A plot of the variation of relative probability to visit is shown in Figure 2 and clearly indicates the marked effect of distance.

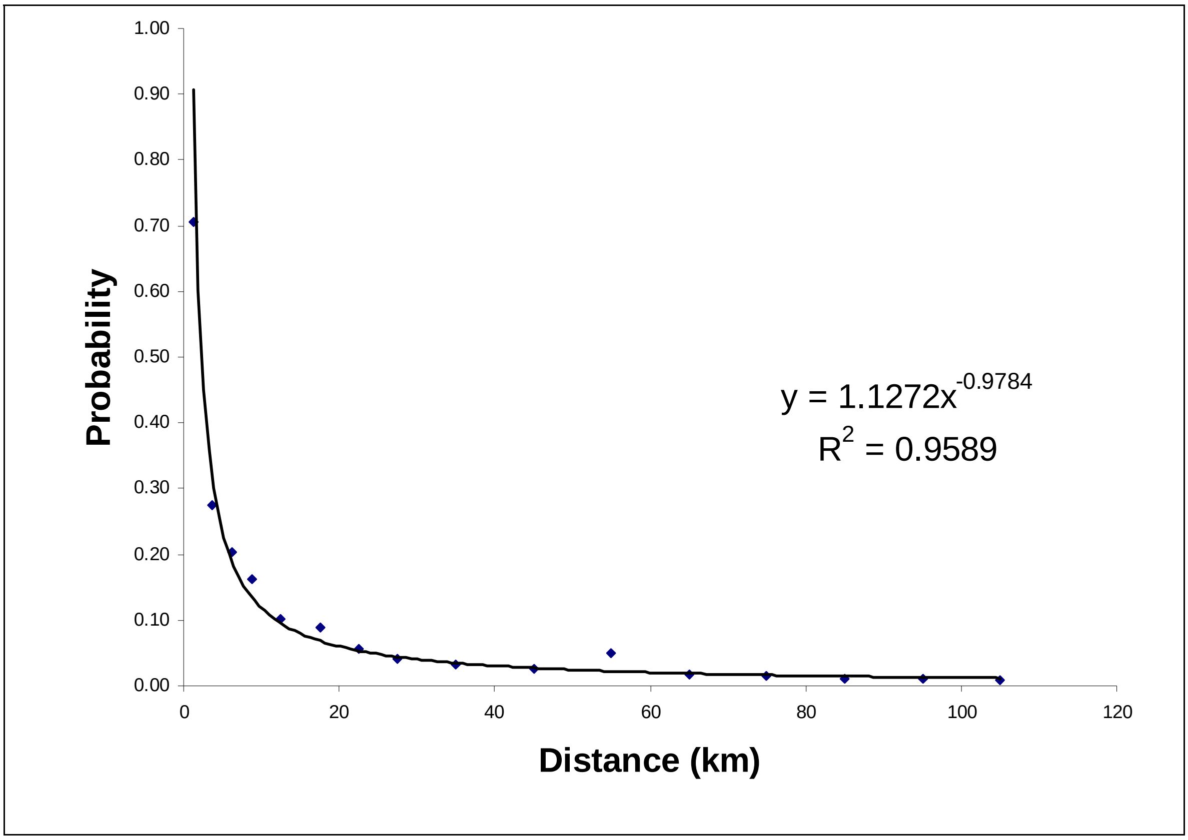

Figure 2. Probability of visiting a restaurant by distance

This indicates that the probability varies approximately inversely with distance. This has to be put in the context of the best geodemographic Super Group having a probability of about three times the average. Clearly the geodemographic and distance effects are both important.

7. The probability of visiting by broad geodemographic type at various distances

The relative probability of visiting the restaurant was recalculated as in Figure 2 but this time separately for each OAC Super Group and the results are shown in Figure 3 for each distance band. The Super Groups are shown in their general order of propensity, so that the first bar within each cluster is Super Group 3, the second is Super Group 4 and so on.

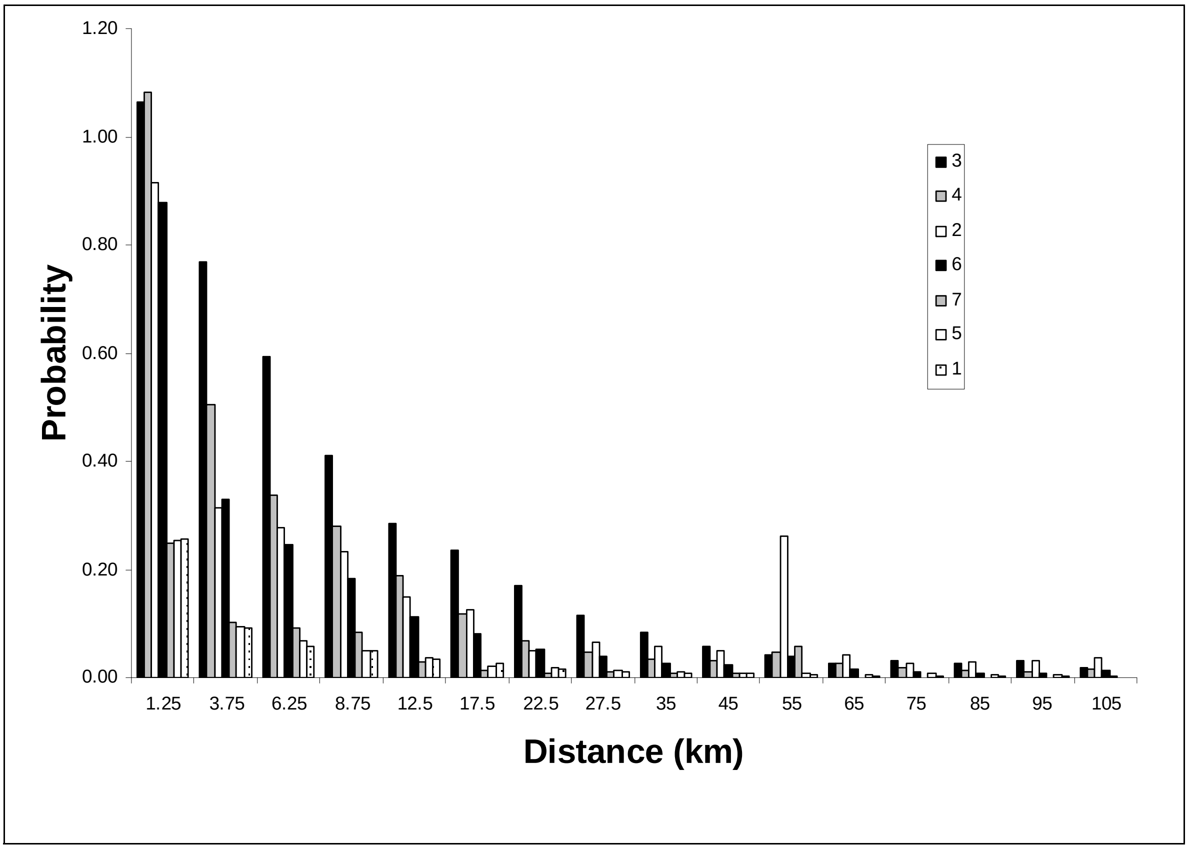

Figure 3. Probability of visiting by OAC Super Groups and distance

As would be expected, the probability for each Super Group is different and these differences are broadly replicated at each distance band.

8. Producing relative probabilities that are despatialised

An estimate of the relative propensity of people in the different OAC Super Groups to visit the restaurant can be obtained from the data shown in Figure 3 by indexing it on the distance effect shown in Figure 2 for each distance band. The best estimate of a despatialised index is the mean across the distance bands. This is shown in Figure 4 – note that the order of Super Groups is different from Figure 3, but the same as in Figure 1.

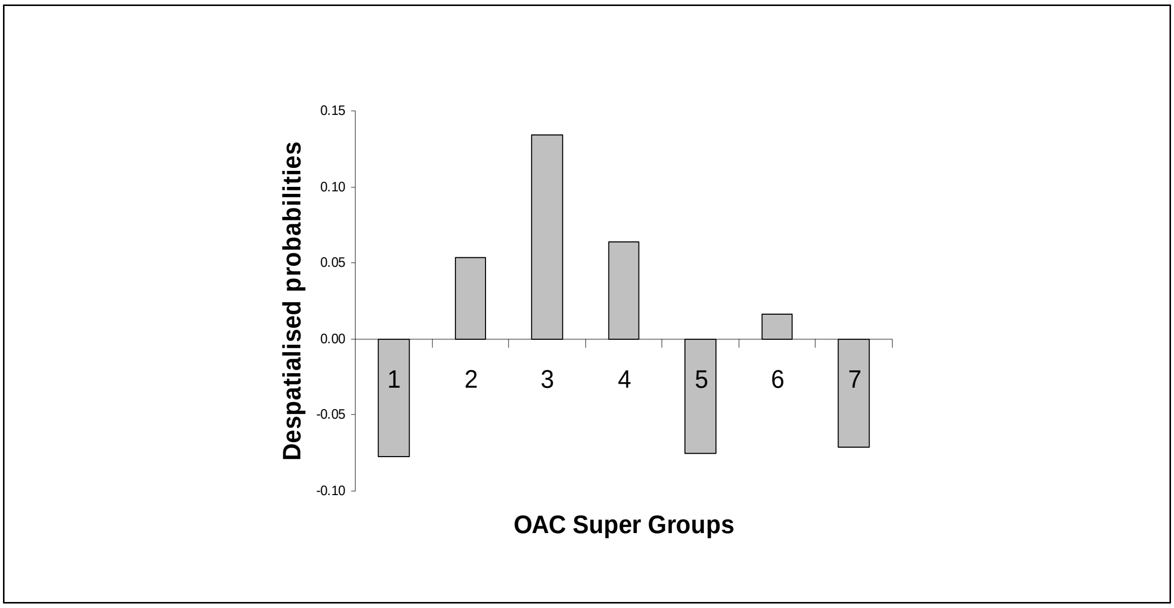

Figure 4. Despatialised indices

The comparable figures calculated using the standard method are shown in Figure 1. It is clear that there are some marked differences, the most noticeable of which is that *City Living* (OAC 2) people are shown to have the greatest probability of visiting the restaurant when estimated by the standard method whilst it is shown to be lower than *Countryside* (OAC 3) when calculated in a despatialised way. The use of the standard propensity figures would therefore lead to new restaurant locations being selected that would not be optimum.

9. Examination of the spatial effect

It has been hypothesised that the differences between the two methods arise because the mean distance of the different OAC Super Groups from the sites is not the same. In particular, the results would suggest that the *City Living* people would be closer to the sites than average whilst the *Countryside* people would be further away than average.
To test this, the mean distance of the population of OAC Super Group was calculated up to 105km. The results are shown in Table 2.

Table 2. Mean distance of OAC population within catchments up to 105 km

Super Group	Name	Dist (km)
1	Blue Collar Workers	37.92
2	City Living	19.3
3	Countryside	40.35
4	Prospering Suburbs	31.16
5	Constrained by Circumstances	34.48
6	Typical traits	31.07
7	Multicultural	14.73
All		30.61

There was a big variation of mean OAC distances and, as expected, the *City Living* people were closer than average to the restaurants and the *Countryside* people further away. It is worth remarking on the fact that the mean distance for all Super Groups was 30.61 km although the radius of calculation was 105 km. This reflects that the more distant output areas were more likely to be in countryside with a much lower population density (the restaurants were always located in urban areas)

10. Checking more formally the relationship between the two sets of probabilities

In principle it should be possible to use the distances in Table 2 to explain the differences between the indices produced by the two methods.

The standard indices were weighted by multiplying them by their mean distance and the resulting estimates of the despatialised indices adjusted to sum to the same total as the standard indices. Table 3 shows the data and the effect of applying the distance weights.

Table 3. Comparison of the indices from the two methods and conversion from one to the other by distance weighting.

Super Group	Name	Standard Index	Despatialised Index	Mean distance Km	Modelled despatialised
1	Blue Collar Workers	0.19	0.26	37.92	0.27
2	City Living	2.85	2.07	19.30	2.03
3	Countryside	1.30	2.33	40.35	1.93
4	Prospering Suburbs	1.35	1.38	31.16	1.55
5	Constrained by Circumstances	0.24	0.32	34.48	0.30
6	Typical traits	1.10	0.99	31.07	1.26
7	Multicultural	0.70	0.28	14.73	0.38
	R squared		0.64		0.96

This shows that a reasonable approximation of the despatialised (and arguably better) indices can be obtained from those calculated in the standard way simply by weighting directly by the mean distance each OAC population is away from the site.

The R-squared of the standard indices to the despatialised indices is .64 and for the modelled despatialised data and the actual despatialised data is 0.96. This suggests that the differences between the two sets of indices are at least in part due a distance effect. The relationship between the despatialised and modelled despatialised data is shown in Figure 5.

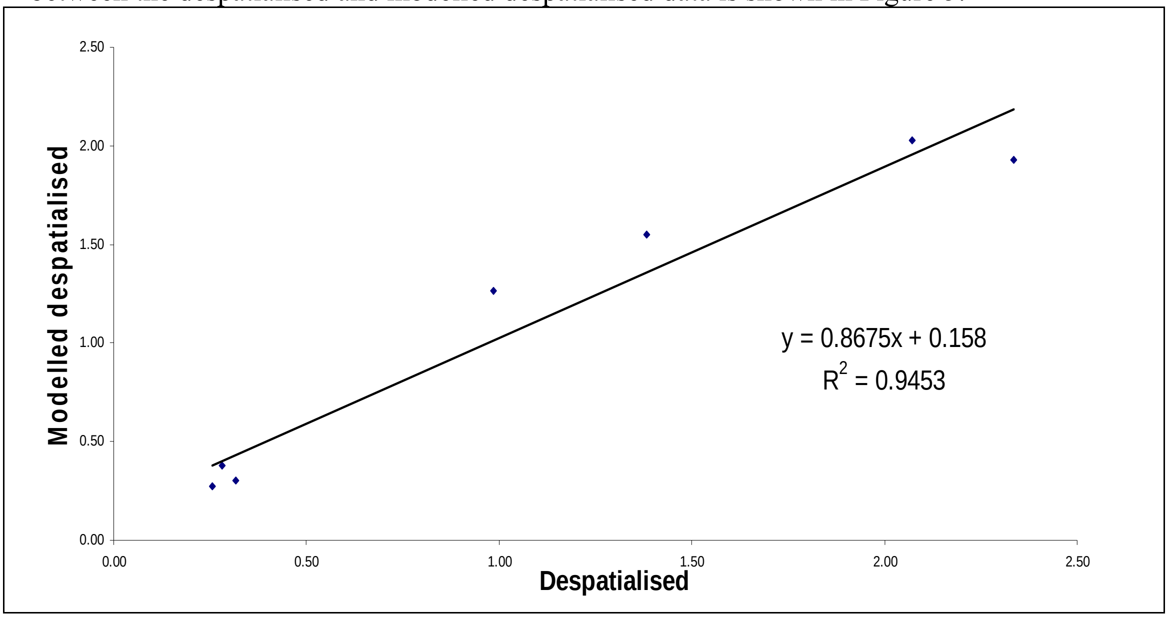

Figure 5. Relationship between the despatialised indices and those modelled from the standard method

11. Discussion

For over thirty years, market propensities have been attributed to geodemographic types using a simple index derived from comparing the percentage of customers in each type to the equivalent percentage in the population. Given a file of customer posycodes, this is a simple and easy to do which and not need specialist GIS knowledge. However, the process neglects the, until recently, unknown fact that geodemographic types spatially autocorrelate (Callingham, 2007). When this new information is coupled with the well known (and common sense) fact that propensity to visit drops off with distance, a flaw is exposed in the way the conventional index method has been calculated as it is now obvious that it favours those geodemographic types that happen to be near the site. Fortunately, it is possible to make corrections to these indices simply by weighting them by the average distance that the geodemographic type's population is from the site as explained in 10 - a figure that is easily calculated.

12. Conclusion

The market propensity indices described are regularly used to help in site selection for all types of retail outlets. Clearly, it is unsatisfactory to use metrics that have inaccuracies in them. In fact what these indices do is to falsely validate the original site selection criteria, though they may of course (if one is lucky) be correct. Therefore, in future these simple metrics should include a further stage of geodemographic distance weighting to remove the bias introduced by auto spatial correlation which appears to be a fundamental property of these systems.

References

Bermingham, J., McDonald, C. & K. Baker. 1979, *The Utility to Market Research of the Classification of Residential Neighbourhoods*. MRS 22nd Annual Conference, Brighton, March 21-23rd 1979.

Harris, R., Sleight, P. & R. Webber. 2005, *Geodemographics, GIS and Neighbourhood Targeting.* (Wiley: London).

Vickers, D. & P. Rees. 2006, *Introducing the Area Classification of Output Areas.* Population Trends 125.

Webber, R. & J. Craig. 1978, *Socio-economic classifications of local authority areas.* Studies in Medical and Population Subjects 35.Office of Population, Censuses and Surveys: London. 2520

Callingham, M. 2007, *Putting geography back into geodemographics*. GIRUK 2007 Conference Maynooth, Ireland.

Office for National Statistics 2005, *Area classification for Output Areas*. [http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/oa/default.asp].

Biography

Martin Callingham is a Visiting Professor at Birkbeck College, University of London where his research interests are areal classification and flow data. He has been involved with geodemographic systems since their inception and was a member of the ONS working party for the creation of their Output Area Classification