Paul J. Boyle
and Robin Flowerdew

Improving Distance Estimates between Areal Units in Migration Models

There are many methods of modeling migrant flows within a set of areal units, but it is common in most to incorporate some measure of distance as an explanatory variable. These distances are effectively meant to represent the typical distance between pairs of areas that would be traveled by potential migrants. They are usually calculated between population-weighted centroids derived for each zone. It is argued here that this method of calculating distance is biased and that the zonal system used will influence the final model parameters that are intended to describe the underlying migration process. The distances between nearby zones will be particularly poorly specified using this approach, but other problems arise which relate to the shape of the zones and the position of the zones in relation to each other. This paper describes an alternative method of calculating these distances which reduces this bias. It is shown that the resulting models fit the data far more satisfactorily and that the residuals from models incorporating this approach are significantly different from those identified from models that use the standard method of specifying distance.

INTRODUCTION

A large literature has developed that discusses the most appropriate methods for modeling migration matrices. Following the early work of Zipf (1946) and Stewart (1948), multiple regression models evolved using the basic gravity model variables of origin and destination size and the distance between these places to predict migration. It is expected that the flows will be positively related to the population size of both the origin and destination and negatively related to the distance between them. A range of alternative socioeconomic variables may be incorporated to improve these types of model.

The work carried out here was supported by funding from the ESRC (grant number H 507 25 5131). The authors thank Jianfa Shen for his help with some of the calculations, Tony Gatrell for his valuable insights into this work, and the anonymous referees for their useful comments. The Census data has been provided through the Manchester Computer Centre with the support of the ESRC/JISC initiative. Results are published with the permission of Her Majesty's Stationery Office (Crown Copyright).

Paul J. Boyle is lecturer in geography, University of Leeds. Robin Flowerdew is senior lecturer in geography at Lancaster University.

Geographical Analysis, Vol. 29, No. 2 (April 1997) © 1997 The Ohio State University Press
Submitted: 3/3/95. Revised version accepted: 8/29/96.

More recently, the use of ordinary least squares (OLS) regression in modeling counts of migrants has been criticized, primarily because it is based on the continuous normal distribution. Flows of migrants are non-negative whole numbers and consequently a discrete probability distribution is appropriate; if each case is assumed to be independent the Poisson distribution should be used (Flowerdew 1991). Poisson regression models can be fitted as generalized linear models in packages such as GLIM and the goodness of fit can be assessed using the deviance; comparing the model deviance with the null model deviance (the most basic model, which assumes each individual flow has the mean size) it is possible to calculate a pseudo r^2, or G^2, value which can be compared between models. Additionally, Baxter (1982) has shown that the widely used family of spatial interaction models based on entropy maximizing (Wilson 1970) are special cases of Poisson regression models.

Considerable attention has been given for many years to the role of distance in migration (for example, Olsson 1965). Migration is generally assumed to reduce with distance because it increases the generalized costs of moving, or because it reduces the amount of information about destinations available to potential migrants. Distance may be measured in a number of ways, including Euclidean distance, road distance, time distance, or cognitive distance. Euclidean distance is most commonly adopted, mainly because it is relatively simple to compute given the coordinates of centroids for each of the origin and destination areas. Use of more sophisticated distance measures appears to make little difference to model fit (Olsson 1965, p. 58), except in cases where Euclidean distance involves crossing a barrier such as an estuary or bay. Debate has also evolved over how distance decay should be modeled (Taylor 1975). In the early work of Stewart (1948) and Zipf (1946), which was based on Newton's original gravity model, migration was assumed to decline with the square of distance. More recently two forms of distance decay function have dominated the literature, the power and the negative exponential functions. Various issues influence the adoption of either of these functions, but it generally seems that the negative exponential function is more effective when dealing with short-distance interaction, such as migration within urban centers, and the power function is more effective for describing longer-distance flows, such as migration between urban areas (Fotheringham and O'Kelly 1989).

DISTANCE CALCULATIONS

Data in migration studies usually refer to zones (states, counties, etc.) that have spatial extent, but distances must be computed between points, usually the centroids of the zones. These distances should be representative of the separation between zones, but error is inevitably introduced by this process. Arbia (1989) refers to this problem of incorrectly representing geographical information, necessitated by the need to perform some statistical procedure, as "model error." Clearly intercentroid distances are affected by the scale at which zones have been defined and by how the zonal boundaries have been drawn. The modifiable areal unit problem (Openshaw 1984) is therefore relevant. The issue of distance measurement for spatial zones has been discussed in some detail by writers on the location/allocation problem (reviewed by Current and Schilling 1987); however their concern is to design an optimal network of facilities despite this source of error, whereas the objective here is to investigate how far the error affects our interpretation of the effect of distance on migration flows.

The problem is stated clearly in the context of migration modeling by Gordon:

> A more general problem for analysts of interregional flows stems from the use of distances between population centroids of large regions to approximate the actual distance traveled by migrants between pairs of regions. On average, these movement lengths will almost always be less than the distance between centroids and, depending on the distribution of population, may be very much shorter. A better approximation can be obtained on the basis of simulated flows between smaller units, if an estimate of the "true" distance decay function for migrants is available. (Gordon 1975, p. 161)

Despite this statement, Gordon gives little detail of how to make this approximation, although he states that it improves the interregional migration model formulated by Weeden (1973) on which he is commenting. Another relevant study was undertaken by Plane (1984) who, in attempting to derive "inferred distances" on the basis of migration flows, experimented with the effects of random displacement of zonal centroids, finding that it made little difference.

The fullest examination of this question in the geographical literature is Webber's (1980) theoretical analysis of aggregation in spatial interaction models. He shows that calculation of distances between zones in an origin-constrained spatial interaction model based on population-weighted centroids leads to biased estimates of mean distances traveled, and to differences in results dependent on the zonal system used. He proposes calculating the proportion of moves from zone k that go to zone l as a weighted average of the proportion of moves from zone i that go to zone j for all i in k and all j in l; the weights are based on the proportion of the population of k residing in i. This removes the bias in estimates of mean distances traveled. The use of migration-weighted distances adopted in this paper for a different form of spatial interaction model is similar in spirit to Webber's results.

It is important to be clear about what the distance variable in migration studies is actually intended to measure. In a model of spatial interaction the empirically observed flows are related to the distance between the respective areal units and this is a measure of the "average" distance that potential migrants in these places are apart. Thus, a population-weighted centroid is more appropriate than the geometrical centroid. However, to assume that migrants from place i to place j have the same distribution as the general population of i, or that those who move will end up evenly distributed across j, is likely to be incorrect. The very nature of the distance decay commonly observed in migration flows means that, other factors being equal, those living in that part of place i which is close to place j will be more likely to migrate to j than those living in parts of place i which are more distant from place j. Obviously, the population-weighted centroid of i is the average location of the entire population resident in i, and not the average location of potential migrants from i to j. Measuring the distance from i to j between their respective population-weighted centroids will therefore introduce bias into the model. It will be a more important source of bias for zones that are close to each other and effectively negligible for distant pairs of zones.

The spatial structure of the areal units may also introduce bias. Contiguous pairs of zones that have long common boundaries compared to their areas are likely to experience more migration than those which have a short common boundary compared to their areas, simply because there will be more short-distance moves that happen to cross the boundary. Similarly, zones that are elongated rather than compact are liable to have more short-distance, cross-border migrants. The size of the zone is also important as the larger the zone, the more inaccurate the population-weighted centroid will be as a measure of the average migration location for in- and out-migrants to particular destinations. These ideas are demonstrated by Rogerson (1990), using ideas originally developed by the eighteenth-century scientist Buffon. He calculates the relationship between migration distance and the probability of crossing a border under various assumptions about region size and shape and the nature of the distance decay function. In some zonal systems, one zone may completely encircle another. This makes the estimation of distance between population-weighted centroids even more of a problem as they may be located very close to each other. Indeed, the population-weighted centroid of the outer area may well be located within the inner area. Less extreme than this "doughnut" problem is the "croissant" problem where one zone, curved in shape, partly surrounds another.

This paper considers the problem of defining the distances between the fifty-five counties[1] of England and Wales for use in migration modeling. A method for estimating the "average migration distance" that migrants are likely to have moved between these zones is proposed and the resulting model fit, parameters, and residuals are examined.

THE STUDY AREA AND DATA

The problem of measuring distance would be simplified if individual-level flows were measured between point locations, rather than areas, but in the analysis of migration the data are generally retrieved from sources, such as censuses, that provide data aggregated into areal units. In this study, the data were extracted from the Special Migration Statistics Set Two (SMS II) derived from the 1981 Census and held at the Manchester Computer Centre (MCC). The SMS II provide a 100 percent count of the total number of migrants, disaggregated by sex, moving between the wards of England and Wales and the similarly sized postcode sectors of Scotland. These zones usually contain a few thousand people. It is possible to extract flow data for any aggregation of wards that aggregate neatly into Local Authority districts which themselves aggregate neatly into counties. In this theoretical study the data were extracted at the county level for England and Wales (Figure 1).

According to these data there were 4,230,417 people whose address at the time of the 1981 Census was different from their address one year previously. Of these, only 1,062,670 moved between the fifty-five counties, emphasizing the strength of the distance decay effect on migration within England and Wales. Of the 2,970 potential intercounty flows only one (from Powys to the Isle of Wight) was zero. The remaining flows ranged between 52,054, from Inner to Outer London to 2 from the Isle of Wight to Mid Glamorgan, and from Northumberland to the Isle of Wight. There was considerable variation in the populations of these fifty-five counties with a maximum of 4,182,980 in Inner London and a minimum of 108,128 in the Welsh county of Mid Glamorgan. In general those counties with the largest populations tend to have the smallest areas.

POPULATION-WEIGHTED CENTROID MODEL

A variety of methods exist for calculating areal centroids, but in migration modeling it has been assumed that population-weighted centroids are most appropriate. These are related to the distribution of the population within each area

---

[1] Some of the fifty-five areas are not strictly counties. Rather, they are the units corresponding to the files used for making the 1981 Census Enumeration District data available to the academic community. The former metropolitan county of Greater London is divided into Inner and Outer London; the other former metropolitan counties and the shire counties make up the remaining units. The term "county" is used here to refer to these units.

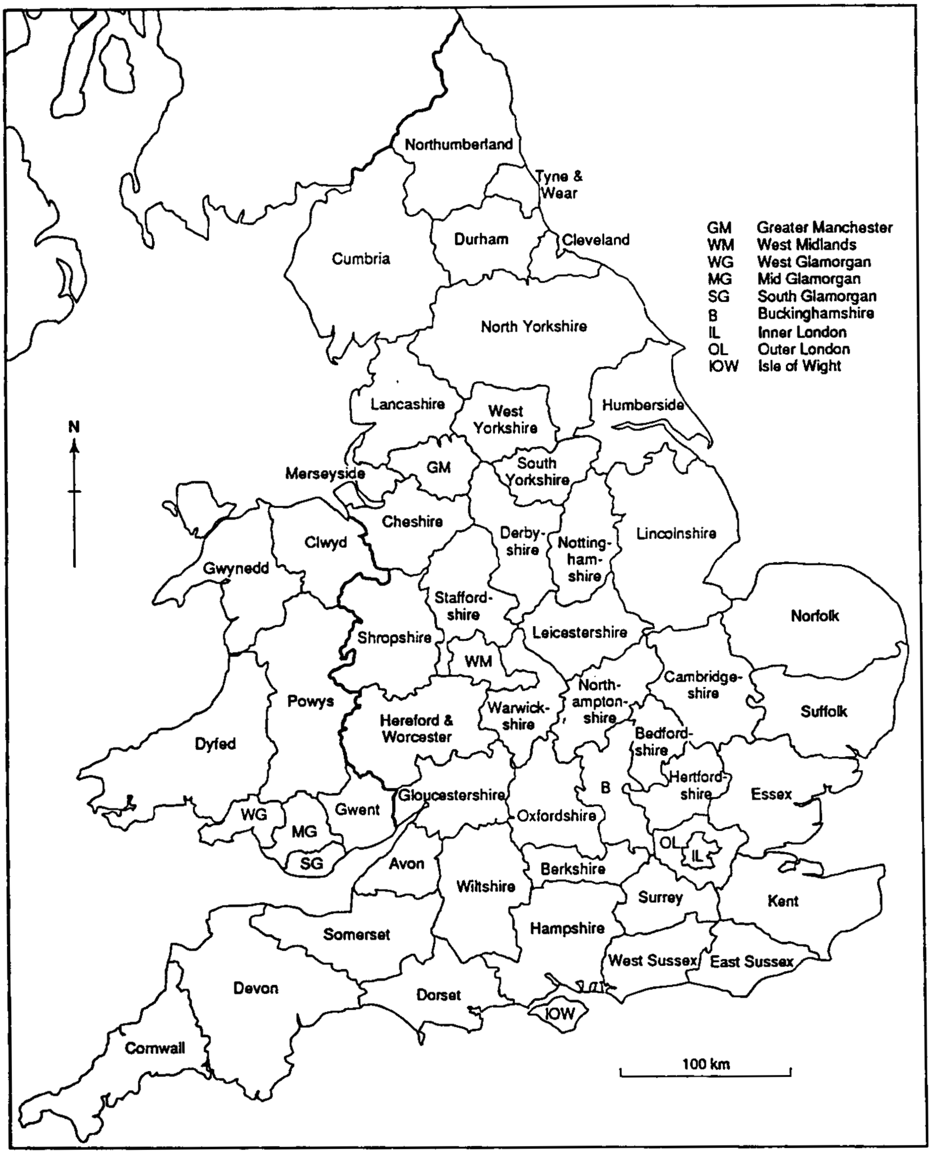

FIG. 1. "Counties" in England and Wales

and have been commonly adopted for the calculation of distance (for example, Flowerdew and Boyle 1992; Flowerdew and Salt 1979; Gordon 1988). The 1981 British Census provides areal centroids for the 130,000 Enumeration Districts (EDs) in Britain which were manually derived using Ordnance Survey maps. Ward centroids are calculated as the population-weighted average of those ED centroids that fall within a particular ward; wards are neat aggregations of EDs. The x and y coordinates of the ward centroid are calculated as

$$wx = \sum x_i \cdot p_i / p \tag{1}$$

$$wy = \sum y_i \cdot p_i / p \tag{2}$$

where x_i, y_i, and p_i are the eastings, northings, and population of each ED within a ward, and p is the total ward population. This method can be used to derive centroids for areas that are aggregations of wards, and this procedure was used to calculate population-weighted centroids for the fifty-five counties and the 403 districts that aggregate neatly into these counties. The Euclidean distance between these centroids is then calculated using:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

where x_i, x_j, y_i, and y_j are the eastings and northings of the centroids in areas i and j and d_{ij} is the distance between them. Some of these distances were altered where physical barriers, such as river estuaries, were thought to make Euclidean distance inappropriate. A Fortran program was written to recalculate these distances for those pairs of areas that were misrepresented by straight-line Euclidean distances because of an intervening water body (such as between Cornwall and West Glamorgan).

The resulting distances in the county-level migration system ranged from 582.3 kilometers between Cornwall and Northumberland to only 1.34 kilometers between Inner and Outer London. The latter of these pairs of counties is an example of the "doughnut" problem as Inner London is entirely encompassed by Outer London and the resulting population-weighted centroids are located close to each other in Inner London.

The flows between the fifty-five counties were modeled using a gravity model based on the Poisson distribution where the explanatory variables were the origin and destination populations and distance was measured between population-weighted centroids. The Poisson regression model my be defined as

$$Y_i = \exp\left(\sum_i \beta_i X_i\right) + \varepsilon_i \tag{4}$$

where the random variable Y_i is assumed to have a Poisson distribution whose parameter is logarithmically linked to a linear combination of the explanatory variables (Lovett and Flowerdew 1989). The gravity model may be expressed as

$$M_{ij} = \exp(\beta_0 + \beta_1 \ln P_i + \beta_2 \ln P_j + \beta_3 \ln d_{ij}) + \varepsilon_i \tag{5}$$

where

M_{ij} = the migrants moving between i and j,
P_i = the total population in i,
P_j = the total population in j,
d_{ij} = the distance between i and j.

The fit of each model may be assessed using the deviance measure which is calculated as

$$D = 2\left(\sum_i \sum_j M_{ij} \ln(M_{ij}/\hat{M}_{ij})\right) \tag{6}$$

where $\hat{M}_{ij}$ = the expected migrants moving between i and j.

The model deviances can be compared to the null model deviance to estimate the fit of each model. The null model is the simplest model where the estimated

TABLE 1
Model Parameters and Deviances

	Deviance	Degrees of freedom	Constant	ln P_i	ln P_j	ln d_{ij}	Contiguity parameter
Model 1	534,675	2,966	−6.250	0.7061	0.4685	−0.795	
Model 2	357,691	2,966	−6.541	0.7893	0.5601	−1.248	
Model 3	355,068	2,966	−9.994	0.7863	0.5361	−0.502	1.195
Model 4	319,124	2,966	−9.009	0.8272	0.5900	−0.957	0.695

Model 1 = Gravity model with population-weighted centroid distances
Model 2 = Gravity model with migration-weighted distances
Model 3 = Gravity model with population-weighted centroid distances and a contiguity dummy variable
Model 4 = Gravity model with migration-weighted distances and a contiguity dummy variable

values of each observed flow are given as the average flow size within the entire system. Standardized residuals from these models can be calculated from

$$SR = (M_{ij} - \hat{M}_{ij})/\sqrt{\hat{M}_{ij}}. \tag{7}$$

The gravity model was fitted to the 2,970 intercounty flows and the resulting deviance of 534,675 was a reduction of 77.6 percent from the null model deviance of 2,385,862 (Table 1). The deviance reduction suggests that the model explains a large proportion of the variation, but it does not fit the data. In those cases where the migration matrix is not excessively sparse (Boyle and Flowerdew 1993) the chi-squared test for goodness of fit is a reasonable measure of the success of the model; for the model to be regarded as providing a good fit to the data, the deviance should be not much in excess of model degrees of freedom (slightly less than the sample size if only a few parameters are fitted). This model used distances calculated between each of the fifty-five population-weighted centroids; the distance parameter shows quite a strong distance decay in the migration of people between the counties of England and Wales (Table 1). The population parameters suggest that migration is not proportional to population size, particularly at the destination; destinations with smaller populations attract proportionally more migrants than destinations with large populations.

MIGRATION-WEIGHTED CENTROID MODEL

It has been suggested above that the distances calculated between population-weighted centroids are inappropriate estimates of the average distances moved by migrants between a pair of areal units. The distances will tend to overestimate the average distances moved and the bias will be most severe for those pairs of areas that are close to, or contiguous with, each other. This study addresses this problem by implementing an innovative method of estimating these average migration distances.

As suggested above the results from this model are biased, however, as the distances used between each pair of counties do not reflect the probable average migration distance likely to be moved between them. In order to circumvent this problem estimates of the average intercounty migration distances moved are required. This can be done by breaking down the single flow from county *i* to county *j* into estimated flows between their constituent parts. The estimation procedure utilizes the county-level gravity model parameters to estimate the flows between a larger set of zones that aggregate neatly into the county boundaries, in this case, the 403 Local Authority districts (the distances between the districts are derived from population-weighted centroids). For the purpose of this theoretical study, it is assumed that we know the populations of the districts but not the migration between them.

The migration between districts is estimated on the basis of the model derived at the county level:

$$\hat{M}_{AB} = \exp(\beta_0 + \beta_1 \ln P_A + \beta_2 \ln P_B + \beta_3 \ln d_{AB}). \tag{8}$$

The migration-weighted distance (MWD_{ij}) between each pair of counties is then estimated as

$$MWD_{ij} = \left(\sum_{A \in i} \sum_{B \in j} \hat{M}_{AB} d_{AB} \right) / M_{ij} \tag{9}$$

where

$\hat{M}_{AB}$ = estimated migration between district A and district B,
d_{AB} = distance between district A and district B.

The migration-weighted distance between county i and county j is the average of the distance from each district in i to each district in j, weighted by the estimated number of migrants in each interdistrict flow. Note that the migration-weighted distances are not commutative; MWD_{ij} is not equal to MWD_{ji}. It is then possible to reestimate the intercounty gravity model, incorporating the revised distance estimates. A new set of parameters is produced. These can then be used to provide a second set of estimated interdistrict flows which can then be converted into a second set of intercounty migration-weighted distances. The process is iterative and continues until the deviance for the intercounty model stabilizes. Of course, the modifiable areal unit problem (MAUP) would suggest that using estimates at one scale may be unreliable at another scale, but Amrhein and Flowerdew (1992) suggest that this may be less of a problem in migration models that are correctly specified as Poisson models. Additionally, the results below confirm that this approach is a substantial improvement on the standard method of measuring distance.

COMPARISON OF MODELS

The deviance for the final intercounty model (model 2) is shown in Table 1 and this is clearly a substantial improvement on the deviance from the original county-level model accounting for 84.3 percent, rather than 77.6 percent, of the null deviance. As we would expect from the theoretical arguments stated throughout this paper, the distance parameter also becomes progressively steeper, highlighting the relatively short distances over which migrants tend to move.

The majority of the migration-weighted distances were shorter than the distances calculated from the population-weighted centroids. The relative difference between these two distance measures was expected to be greatest when i and j are close together. Figure 2 plots the distances derived from population-weighted centroids against the migration-weighted distances where the inter-population-weighted centroid distance was less than 100 kilometers. This shows that migration-weighted distances are shorter when the pairs of counties are close. However, the distance from Inner to Outer London (the "doughnut" example) was lengthened from 1.3 kilometers to 13.23 kilometers—the largest increase in distance between two counties. The largest reduction in the distance

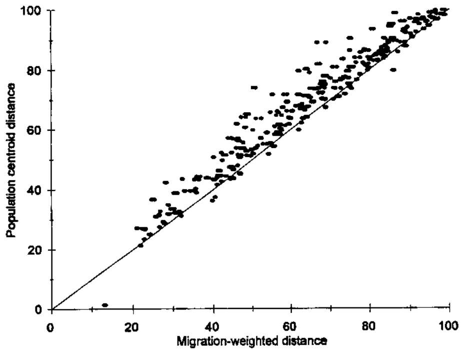

FIG. 2. Distances Calculated from Population-Weighted Centroids and Migration-Weighted Distances Where the Population-Weighted Centroid Distance Is Less than 100 Kilometers

TABLE 2
Observed and Estimated Contiguous and Noncontiguous Flows

	Noncontiguous flows	Contiguous flows
Observed	587,980	474,690
Estimated (model 1)	730,386	332,285
Estimated (model 2)	654,306	408,365

TABLE 3
Observed and Estimated Flows for Models 1 and 2

		Model 2	
		$M_{ij} > \hat{M}_{ij}$	$M_{ij} < \hat{M}_{ij}$
Model 1	$M_{ij} > \hat{M}_{ij}$	683	87
	$M_{ij} < \hat{M}_{ij}$	524	1,676

between two counties was 23.52 kilometers between Powys and Mid Glamorgan. Of the 2,970 migration-weighted intercounty distances, 2,218 were shorter than the distances derived from population-weighted centroids.

The total observed and estimated numbers of migrants between contiguous and noncontiguous counties are shown in Table 2 and the estimated migrant totals are much more similar to the observed totals in model 2. While the estimated number of migrants moving between contiguous counties was 142,405 smaller than the observed total in model 1, it was only 66,325 smaller in model 2. It is not surprising that part of the contiguity effect is still apparent, because the correction based on interdistrict flows cannot be expected to adjust for the presence of very short distance flows that happen to cross county boundaries.

Table 3 provides a breakdown of the 2,970 flows based on whether the estimated number of migrants was larger or smaller than the observed number in models 1 and 2. The off-diagonal totals are those where the estimated number of migrants from the two models were in the opposite direction and as many as 611 fell in this combined category. This shows that a considerable number of flows may be misinterpreted if the researcher is interested in identifying particular flows that are larger than expected. Table 4 shows that the flows over distances less than 50 kilometers and between 150 and 299 kilometers were pre-

TABLE 4
Observed and Estimated Flows over Different Distances

	< 50	50–149	150–299	> 300
Observed	310,065	431,615	247,207	73,783
Estimated (model 1)	253,410	400,359	326,244	82,658
Estimated (model 2)	294,200	479,132	242,415	46,923

TABLE 5
Ten Highest Contributors to the Deviance from Model 1

Origin	Destination	Observed	Estimated	Deviance
Outer London	Inner London	30,930	72,696	30,669
Outer London	Essex	12,535	3,113	16,077
Outer London	Surrey	11,745	3,801	10,614
Outer London	Kent	9,294	2,648	10,047
West Midlands	Staffordshire	8,061	2,173	9,356
Surrey	Hampshire	4,665	930.3	7,573
West Midlands	Hereford and Worcester	6,146	1,840	6,215
Cornwall	Devon	2,242	317.2	4,919
Tyne and Wear	Northumberland	3,864	967.4	4,909
West Yorkshire	North Yorkshire	4,646	1,348	4,902

TABLE 6
Ten Highest Contributors to the Deviance from Model 2

Origin	Destination	Observed	Estimated	Deviance
Inner London	Outer London	52,054	32,481	9,954
Hampshire	Devon	2,219	338.9	4,579
Devon	Hampshire	1,945	303.8	3,940
Outer London	Essex	12,535	6,882	3,727
Cornwall	Devon	2,242	439.6	3,701
Devon	Cornwall	2,432	549.9	3,467
Surrey	Hampshire	4,665	1,789	3,191
West Yorkshire	Manchester	1,356	4,362	2,843
Manchester	Merseyside	1,813	5,063	2,776
Norfolk	Suffolk	2,449	679.2	2,742

dicted better by model 2 than model 1, while the prediction of the flows between 50 and 149 kilometers and over 300 kilometers were predicted better by model 1.

The major contributors to the deviance from model 1 are provided in Table 5. It is evident that all ten of these flows were between contiguous counties and the decentralizing flows from Outer London and the West Midlands were much larger than predicted, while the flow from Outer London into Inner London was much smaller than predicted (not surprisingly given the small distance value derived from the population-weighted centroids). As many as seventy-seven of the top one hundred contributors to the deviance were between contiguous counties. In contrast, Table 6 provides the major contributors to the deviance from model 2. The flow from Outer to Inner London remained poorly predicted, although the estimate was a significant improvement from model 1. One of the decentralizing flows from Outer London remained in the list, but the flows out of the West Midlands conurbation were no longer included. Seven of the flows did not appear in Table 5. Flows involving Devon, Corn-

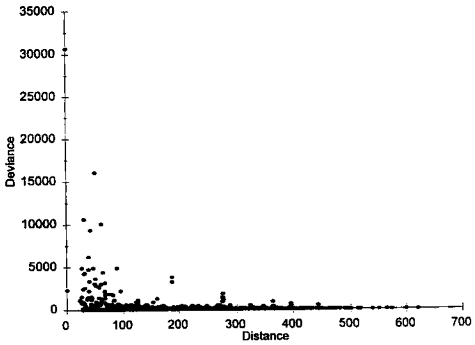

FIG. 3. Flow Deviances from Model 1 and Population-Weighted Distance

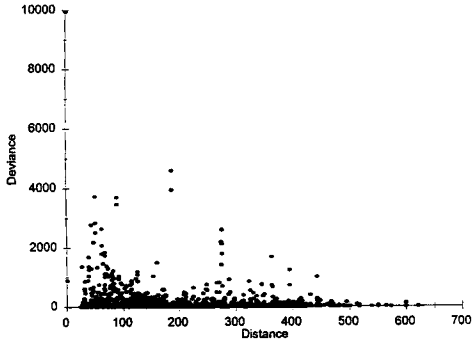

FIG. 4. Flow Deviances from Model 2 and Population-Weighted Distance

wall, and Hampshire now appear to be unusually large and it is possible that the movement of armed forces personnel may be influential in these flows (Boyle 1995). The flows between Manchester and Merseyside and West Yorkshire and Manchester were much smaller than expected. Eight of these top ten flows and only forty-five of the top one hundred were between contiguous counties.

It is possible to divide the overall model deviance into contributions from each individual flow; those flows with high contributions to the deviance are the ones that the model fits least well, and deviance contributions can be regarded as model diagnostics in a similar way to residuals. Figures 3 and 4[2] plot contributions to deviance against the population-weighted distances. The high deviances from model 1 are predominantly short distance flows with a small number of very poorly estimated flows. On the other hand, the highest deviances from model 2 were less dominated by the flows over short distances. Even so, large deviance values are obtained when the absolute difference between M_{ij} and $\hat{M}_{ij}$ is large. This will tend to occur when M_{ij} is large and this is usually the case when the areas are close together.

An alternative method is to calculate standardized residuals that are positive when the observed flow is larger than the estimated flow and negative otherwise. The five largest positive and negative residuals are shown in Table 7 and

---

[2] Note that the y axis on Figures 3 and 4 and Figures 5 and 6 are not identical. This has been done to emphasize the distribution of the data along the x axis.

TABLE 7
Highest Positive and Negative Standardized Residuals from Model 1

Origin	*Destination*	*Observed*	*Estimated*	*Residual*
Outer London	Essex	12,535	3,113	168.9
Outer London	Kent	9,294	2,648	129.2
Outer London	Surrey	11,745	3,801	128.9
West Midlands	Staffordshire	8,061	2,173	126.3
Surrey	Hampshire	4,665	930.3	122.4
West Midlands	Derbyshire	516	1,380	−23.3
South Yorkshire	Manchester	492	1,517	−26.3
Manchester	South Yorkshire	503	1,787	−30.4
Inner London	Outer London	52,054	63,876	−46.7
Outer London	Inner London	30,930	72,696	−154.9

TABLE 8
Highest Positive and Negative Standardized Residuals from Model 2

Origin	Destination	Observed	Estimated	Residual
Inner London	Outer London	52,054	32,481	108.6
Hampshire	Devon	2,219	338.9	102.1
Devon	Hampshire	1,945	303.8	94.2
Cornwall	Devon	2,242	439.6	86.0
Cornwall	Hampshire	959	101.1	85.3
South Yorkshire	Manchester	492	2,292	−37.6
Manchester	South Yorkshire	503	2,666	−41.9
Manchester	West Yorkshire	1,690	4,665	−43.6
West Yorkshire	Manchester	1,356	4,362	−45.5
Manchester	Merseyside	1,813	5,063	−45.7

seven of the flows were between contiguous counties. The large positive residuals tended to be in the south of England, while the three negative residuals that were not contiguous were flows within the North and Midlands of England. Table 8 shows the residuals from model 2 and five of these were between contiguous counties; only three of these residuals appeared in Table 7. There was a general tendency for the negative residuals to be flows in the north and northwest of England, while the positive residuals continued to be dominated by flows involving Cornwall, Devon, and Hampshire. It is noticeable that the estimated flow between Inner London and Outer London is much smaller than the observed flow. As with the major contributors to the deviance, the largest positive and negative residuals in model 2 were considerably different from those in model 1.

Figure 5 plots the standardized residuals from model 1 against distance. A number of short-distance flows were particularly high positive residuals, while the flow from Outer to Inner London stands out as a high negative residual. Figure 6 provides a similar plot for the results from model 2 and, although short distance flows continued to be identified as high residuals, the spread of high positive and negative residuals in relation to the distance moved was much more even.

INCLUDING CONTIGUITY DUMMY VARIABLES

It is common to introduce a contiguity dummy variable into such models which Weeden (1973) and Jun and Chang (1986) suggest is justified if it is as-

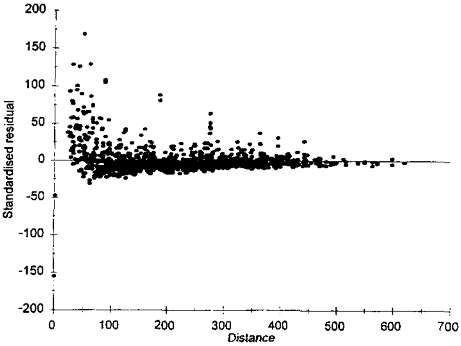

FIG. 5. Standardized Residuals from Model 1 and Population-Weighted Distance

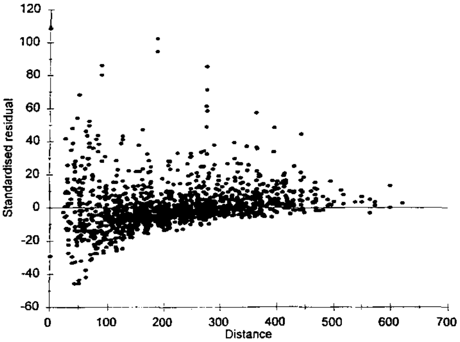

FIG. 6. Standardized Residuals from Model 2 and Population-Weighted Distance

sumed that flows between noncontiguous counties will be primarily for employment reasons while flows between contiguous areas may be more likely to be housing related. It has the effect of making the observed and estimated total flow between all pairs of contiguous areas equal and it will inevitably improve the model fit. It could be argued that much of the bias introduced by using distances measured between population-weighted centroids will be accounted for using such a dummy variable. In Table 1 the parameter estimates and deviances are provided for models that include a contiguity dummy variable. Model 3, which includes distances calculated from population-weighted centroids and a contiguity dummy variable, was a significant improvement over model 1 and the deviance was slightly lower than that for model 2. The distance decay parameter became very small once the contiguity dummy variable was included. Model 4, which included migration-weighted distances and a contiguity dummy variable, was the best model of all, and the relatively large reduction in the deviance from model 3 shows that the migration-weighted distances made important improvements to the estimates of both contiguous and noncontiguous flows. The distance decay parameter remained quite large in this case.

Table 9 shows that as many as 385 of the 2,970 residuals changed sign between models 3 and 4. The ability of the two models to estimate the number of migrants moving over different distances (Table 10) once again showed that the model using migration-weighted distances (model 4) provided more accurate estimates of short-distance flows and flows between 150 and 299 kilometers. A

TABLE 9
Observed and Estimated Flows for Models 3 and 4

		Model 4	
		$M_{ij} > \hat{M}_{ij}$	$M_{ij} < \hat{M}_{ij}$
Model 3	$M_{ij} > \hat{M}_{ij}$	880	75
	$M_{ij} < \hat{M}_{ij}$	310	1,705

TABLE 10
Observed and Estimated Flows over Different Distances

	<50	50–149	150–299	>300
Observed	310,065	431,615	247,207	73,783
Estimated (model 3)	291,648	407,109	281,923	81,991
Estimated (model 4)	302,745	458,573	245,927	55,424

number of differences between the major contributors to the deviance and standardized residuals were identified between models 3 and 4 as was the case for models 1 and 2. For those interested in identifying the major unusual flows among the fifty-five counties in England and Wales, the results from these two models would provide very different interpretations.

CONCLUSION

An original method for determining distances between areal units, for use in modeling spatial interactions, has been discussed and it is argued that this procedure reduces the problems associated with calculating distances from population-weighted centroids. The method gives large improvements in model fit, which result primarily, but not entirely, from the enhanced ability to predict flows over short distances. In general, models using distances calculated from population-weighted centroids underestimate the size of these flows because of the overestimation of the average distance moved by migrants between nearby areas. Interpretations drawn from such models are liable to overemphasize the importance of these short-distance moves. By improving the way the distance decay effect is specified, the impact of substantive factors affecting migration can emerge more clearly.

LITERATURE CITED

Amrhein, C. G., and R. Flowerdew (1992). "The Effect of Data Aggregation on a Poisson Regression Model of Canadian Migration." *Environment and Planning A* 24, 1381–91.

Arbia, G. (1989). "Statistical Effect of Spatial Data Transformations: A Proposed General Framework." In *Accuracy of Spatial Databases*, edited by M. F. Goodchild and S. Gopal, pp. 249–59. London: Taylor and Francis.

Baxter, M. (1982). "Similarities in Methods of Estimating Spatial Interaction Models." *Geographical Analysis* 14, 267–72.

Boyle, P. J. (1995). "Rural In-migration in England and Wales, 1980–81." *Journal of Rural Studies* 11, 65–78.

Boyle, P. J., and R. Flowerdew (1993). "Modelling Sparse Interaction Matrices: Interward Migration and the Underdispersion Problem." *Environment and Planning A* 25, 1201–9.

Current, J. R., and D. A. Schilling (1987). "Elimination of Source A and B Errors in *p*-Median Location Problems." *Geographical Analysis* 19, 95–110.

Flowerdew, R. (1991). "Poisson Regression Modelling of Migration." In *Migration Models: Macro and Micro Approaches*, edited by J. C. H. Stillwell and P. Congdon, pp. 92–112. London: Belhaven.

Flowerdew, R., and P. J. Boyle (1992). "Migration Trends for the West Midlands: Suburbanisation, Counterurbanisation or Rural Depopulation?" In *Migration Processes and Patterns*. Vol. 2: *Population Redistribution in the United Kingdom*, edited by J. C. H. Stillwell, P. Rees, and P. Boden, pp. 44–61. London: Belhaven.

Flowerdew, R., and J. Salt (1979). "Migration between Labour Market Areas in Great Britain, 1970–71." *Regional Studies* 13, 211–31.

Fotheringham, A. S., and M. E. O'Kelly (1989). *Spatial Interaction Models: Formulations and Applications*. Dordrecht: Kluwer.

Gordon, I. (1975). "Employment and Housing Streams in British Inter-regional Migration." *Scottish Journal of Political Economy* 22, 161–77.

——— (1988). "Interdistrict Migration in Great Britain 1980–81: A Multistream Model with a Commuting Option." *Environment and Planning A* 20, 907–24.

Jun, I. S., and H. S. Chang (1986). "Functional Forms and the Relevance of Contiguous Migration in the Study of Migration and Employment Growth." *Annals of Regional Science* 20, 17–27.

Lovett, A., and R. Flowerdew (1989). "Analysis of Count Data Using Poisson Regression." *Professional Geographer* 41, 190–98.

Olsson, G. (1965). *Distance and Human Interaction: A Review and Bibliography*. Philadelphia: Regional Science Research Institute.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography 38. Norwich: Geo Abstracts.

Plane, D. A. (1984). "Migration Space: Doubly Constrained Gravity Model Mapping of Relative Interstate Separation." *Annals of the Association of American Geographers* 74, 244–56.

Rogerson, P. A. (1990). "Buffon's Needle and the Estimation of Migration Distances." *Mathematical Population Studies* 2, 229–38.

Stewart, J. Q. (1948). "Demographic Gravitation: Evidence and Applications." *Sociometry* 11, 31–58.

Taylor, P. J. (1975) *Distance Decay in Spatial Interactions*. Concepts and Techniques in Modern Geography 2. Norwich: Geo Abstracts.

Webber, M. J. (1980). "A Theoretical Analysis of Aggregation in Spatial Interaction Models." *Geographical Analysis* 12, 129–41.

Weeden, R. (1973). "Interregional Migration Models and Their Application to Great Britain." *National Institute of Economic and Social Research Paper* 2. London: Cambridge University Press.

Wilson, A. G. (1970). *Entropy in Urban and Regional Modelling*. London: Pion.

Zipf, G. K. (1946). "The P1P2/D Hypothesis: On Intercity Movement of Persons." *American Sociological Review* 11, 677–86.