<table>
<thead>
<tr>
<th>Title</th>
<th>Developing a Continuous Space Representation of a Simulated Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Cullinan, John</td>
</tr>
<tr>
<td>Publication Date</td>
<td>2010-09-03</td>
</tr>
<tr>
<td>Publisher</td>
<td>Taylor &amp; Francis</td>
</tr>
<tr>
<td>Link to publisher's version</td>
<td><a href="http://www.informaworld.com/10.1080/17421772.2010.493954">http://www.informaworld.com/10.1080/17421772.2010.493954</a></td>
</tr>
<tr>
<td>Item record</td>
<td><a href="http://hdl.handle.net/10379/1551">http://hdl.handle.net/10379/1551</a></td>
</tr>
</tbody>
</table>
Developing a Continuous Space Representation of a Simulated Population

John Cullinan

Abstract

Spatial microsimulation models typically match census of population data with survey data in order to simulate synthetic populations of individuals and households within small-scale geographic areas. For most spatial microsimulation applications this level of spatial precision is satisfactory. For others, more precise information on the location of simulated units may be required. To this end this paper develops a continuous space representation of a simulated population. It presents a statistical matching approach for assigning simulated households from a spatial microsimulation model to unique spatially-referenced residential locations. The allocation is based on a random assignment after splitting the simulated households into two groups: those predicted to reside in apartments and those predicted to reside in houses. The resulting ‘geohouseholds’ have a range of potential applications in economic and spatial analysis.

\(^a\) Department of Economics, National University of Ireland, Galway, Ireland.
1. Introduction

Spatial microsimulation models match census of population data with nationally-representative survey data in order to simulate synthetic populations of individuals and households within small-scale geographic or administrative areas. The aim is to simulate a virtual population with personal and socio-economic characteristics that are as close as possible to the actual population, at a small-area level. The simulated population can then be used to analyse the spatial impact of policies. This level of spatial disaggregation is generally sufficient for standard spatial microsimulation applications. However, in order to use such techniques in other contexts, such as modelling accessibility or destination choices, more precise information in relation to the residential location of simulated individuals may be required. To this end, this paper develops a continuous space representation of a simulated population. More specifically, it presents a statistical matching approach for assigning households from a spatial microsimulation model for Ireland (SMILE) to unique, spatially-referenced, residential locations using GeoDirectory, a buildings database. The resulting spatially-referenced simulated households are termed ‘geohouseholds’.

The geohouseholds developed in this paper can be used in a number of different economic and spatial modelling contexts\(^1\). For example, they can be used to estimate the number and socio-demographic characteristics of individuals residing within an estimated catchment area of a public amenity (e.g. recreation site), public institution (e.g. hospital or school), commercial enterprise (e.g. shopping centre) or any specified geographic location, when combined with GIS-based network analysis (see Cullinan, 2008; Cullinan et al., 2008). Furthermore, geohouseholds can also be used to provide more accurate estimates of travel distances for simulated individuals to such facilities. They also extend naturally to the consideration of accessibility issues for a variety of services, since according to Skov-
Petersen (2001), “the location of the resources relative to the users, the transport-system, and
the way space and distances influences the potential usage of facilities are central issues when
modelling accessibility and mobility”. In this context, Cullinan (2008) used geohouseholds
in an environmental valuation framework to model the actual and potential demand for
recreation at existing and prospective forest recreation sites respectively, as well as their
associated total access/amenity values.

The geohouseholds presented in this paper can be considered a new sub-electoral division
(ED) level simulation-based geography for Ireland. The approach presented is similar in a
number of respects to one previously considered by Ballas and Clarke (2003), who discussed
the possibility of developing a point dataset of houses, containing housing type attributes,
using remote sensing techniques. This dataset could then be linked to spatially
disaggregated microsimulated households at ED level, in order to further disaggregate the
simulated population within each ED. Their idea was essentially to join two separate
databases on the basis of variables they have in common, with the linkage between the two
datasets to be achieved using statistical matching techniques. Despite discussing the process
of matching remotely sensed data and simulated households, Ballas and Clarke (2003) did not
actually undertake such a modelling exercise. Indeed, to the best of my knowledge, there
have not been any previous attempts to undertake such a match until now, where the
allocation is based on a random assignment after splitting the simulated households into those
predicted to reside in apartments and those predicted to reside in houses.

The remainder of this paper is organized as follows: Section 2 provides a general discussion
of spatial microsimulation and details of the SMILE model used in this paper, while Section 3
presents an assessment of the compatibility of the SMILE base data and the GeoDirectory
database for undertaking the statistical match. Section 4 sets out the statistical matching
approach used for undertaking the allocation process and since this represents the first time that a continuous space representation of a simulated population has been developed, a detailed technical account and description of the process is presented. Section 5 provides a brief discussion of some applications of geohouseholds to date and Section 6 presents some concluding remarks.

2. Spatial Microsimulation and SMILE

2.1 Spatial Microsimulation

Microsimulation is defined as a methodology concerned with the creation of large-scale simulated population micro-level datasets in order to describe economic and social events by modelling the behaviour of individual agents (Ballas et al., 2006). It generally involves linking two or more datasets, where they have at least one common variable. Microsimulation models are typically large-scale datasets of the attributes of micro units (individuals, households or firms) and include three main types: static, dynamic and spatial. Static models examine micro units at a point in time, while dynamic microsimulation models project the population in a base year forward through time by simulating transitions such as fertility and mortality, at the individual level. According to Ballas and Clarke (2003), “microsimulation as a technique dates back to the work of Orcutt (1957) and Orcutt et al. (1961) who studied household dynamics in a micro-analytic framework”.

Spatial microsimulation (SM) models, also known as geographical microsimulation models, simulate ‘virtual’ or ‘synthetic’ populations of individuals (usually within households) in given geographical areas, so that the characteristics of these populations are as close as possible to their ‘real-world’ counterparts (Ballas et al., 2006). In other words, the goal of spatial microsimulation is to ‘construct’ small area population microdata. Once
‘constructed’, the principal aim of SM models is to consider both the socio-economic and spatial impact of policy decisions. Their defining characteristic, distinguishing them from other types of microsimulation models, is the geographic dimension whereby units within the SM model are spatially referenced to a specific (usually small-scale) geographic area.

Spatial microsimulation models provide an alternative to undertaking spatial analysis using other datasets that have little or no spatial referencing. For example, the Living in Ireland (LII) survey, a nationally representative longitudinal survey undertaken between 1994 and 2001, contains a spatial variable disaggregated into 12 separate categories: the five cities in Ireland, a category for Dublin County, an ‘open countryside’ category, as well as five categories for towns and villages of different sizes. However, for many applications this relatively ‘coarse’ level of spatial referencing is often not sufficient to fully assess spatial patterns, variations or factors, and frequently implies a need for spatial microsimulation. According to Ballas et al. (2006), “one of the major advantages of microsimulation is that it can be a substitute for conducting detailed surveys to produce survey data ….. at the small area level”. They also state that “the power of spatial microsimulation modelling frameworks lies in their ability to estimate policy-relevant variables at the small area level for which published data do not exist”. In recent years there has been an increase in the popularity of spatial microsimulation modelling which can be partly attributed to the increased availability of spatial data from national censuses, as well as new methods for generating microdata from aggregated tabulations (Williamson, 2002). Despite this, it is worth stressing that the outputs from spatial microsimulation models are model estimates, “the reliability of which depends on a range of factors” (Ballas et al., 2005).

There are four general ‘procedures’ usually associated with spatial microsimulation (Ballas and Clarke, 2003). The first involves the construction of a microdata set, generally from
nationally representative survey data such as the LII surveys. The second procedure then involves sampling from this dataset in order to develop or ‘create’ a micro-level simulated population of individuals within small-scale geographic areas, who match known data for those areas. Once ‘created’, static what-if simulations are considered in the third procedure e.g. what are the impacts of alternative policy scenarios on the simulated population? Finally, the fourth procedure involves dynamic modelling, in order to update the micro-level dataset and conduct future-oriented what-if simulations.

A key feature of spatial microsimulation models is that they provide individual-level data, as opposed to aggregate-level data, and this data is spatially disaggregated at a small-scale geographic level. This has a number of potentially important benefits in a number of different contexts and there have been a large number of applications of spatial microsimulation models to date in a range of areas e.g. impacts of labour market changes (Ballas et al., 2006a), local impacts of national policies (Ballas and Clarke, 2001), health inequalities (Ballas et al., 2006), health care needs (NATSEM, 2004), urban and regional analysis (Birkin and Clarke, 1988; Hooimeijer, 1996; Wegener and Spiekermann, 1996), service delivery (King, McLellan and Lloyd, 2002), crime (Kongmuang, 2004), aggregation of environmental benefit values (Cullinan, 2008; Hynes et al., 2008), and traffic flows (Veldhuisen et al., 2000).

2.2 Simulation Model of the Irish Local Economy

The Simulation Model of the Irish Local Economy, or SMILE, is a static, dynamic population, spatial microsimulation model for Ireland and is used in this paper. It was developed jointly by Teagasc (the Irish agriculture and food development authority), the School of Geography at the University of Leeds and the Department of Economics at NUI, Galway – see Ballas et al. (2005a) for a description of the initial development of the SMILE
model. The static process creates simulated base population data and assigns census attributes to individuals using a combinatorial optimisation statistical matching approach. More specifically, SMILE uses survey and census data to develop spatially disaggregated population microdata estimates for each small-area electoral division in Ireland. Living in Ireland data from 2000 is used as the survey microdata, while small area population statistics (SAPS) from the Irish Census of Population in 2002 are used as the geographically-referenced population data. On completion of the matching process, the static component of SMILE is a simulated population of households and individuals from the LII sample representing the target population in each ED by age, gender, level of education, whether the household is a farm or non-farm household, as well as by the total numbers of individuals and households in the ED. Alignment processes are then undertaken to improve the accuracy of selected key variables and the spatial representativeness of the simulated population. For example, the sum of SMILE’s income variables for each county are adjusted such that they are equal to data from the national accounts. The model has been used in a number of applications to date, including analysing the implications of CAP reforms (Ballas et al., 2006a) and modelling access to GP services (Morrissey et al., 2008).

More generally, spatial microsimulation techniques involve merging spatially disaggregated census of population data with nationally-representative survey data using statistical matching techniques. The process implies using one of a number of different possible algorithms, which include iterative proportional fitting (IPF) and combinatorial optimisation techniques such as deterministic reweighting, probabilistic reweighting and simulated annealing (SA) – see Williamson et al. (1998) and Ballas et al. (2005a). While IPF techniques use estimated ‘joint-probability distributions’ in order to create synthetic data at the small area level, combinatorial optimisation techniques reweight existing microdata samples in order to fit small area population statistics. In order to generate small area
microdata, SMILE uses simulated annealing. Simulated annealing has its origins in thermodynamics and was first suggested as an optimisation method by Kirkpatrick et al. (1983). It has since been analysed and applied to a variety of problems by Dowsland (1993), Laarhoven and Aarts (1987), Openshaw and Rao (1995), Williamson et al. (1998) and Ballas et al. (2007).

Simulated annealing generally requires two types of data: sample population microdata and geographically-referenced small area population or census statistics. The process followed in SMILE involves selecting a series of SAPS tables that describe the population of the geographic areas of interest (the EDs) and then finding records from the microdata survey that best match these tables. Kelly (2004) summarises the approach as follows: it “works by reading in SAPS tables and [LII] data, selecting [LII] households at random to populate the SAPS tables, applying a simulated annealing algorithm to find the best fitting set of households and saving the set of [LII] households that best fits the SAPS tables”.

The idea behind the approach is that a survey dataset such as the Living in Ireland can essentially be reweighted to ‘populate’ the small-scale geographic areas in order to match the profile of the populations of the areas on the basis of the matching variables. The survey dataset provides a detailed record for a sample of households and their occupants. The aim of simulated annealing is to sample from the individuals within this survey to find a set of households that best matches the population of the small area being considered, as described in the SAPS from the Census of Population.

The SA algorithm used in SMILE proceeds by first selecting an initial random sample of records from the LII microdata until there are sufficient households and individuals represented in the selected sample i.e. where the numbers of households and individuals are equal to the SAPS population totals in the ED being considered. These units are then used to
create summary tables containing the same categories of variables as in the selected SAPS tables. The tables for the simulated population and the SAPS are then compared and the total absolute error ($e$) between cell sizes in the two tables is calculated. Once $e$ is calculated for the first simulation, a number of records $T$ in the simulated set are randomly selected and replaced with other records from the LII, once again chosen randomly from the universe of records in the LII microdata. The value of $e$ is then recalculated for the new simulation and the difference or change in error ($\Delta e$) is calculated.

In order to decide if the changes should be accepted or rejected, a decision-making rule is required and SMILE uses the following. If the change in the sum of absolute errors is less than or equal to zero (i.e. if $\Delta e \leq 0$), then this suggests an improvement in the match and the changes are accepted. If $\Delta e$ is positive, then $\exp(-\frac{\Delta e}{T})$ is estimated and compared to a random number between 0 and 1. If the value of $\exp(-\frac{\Delta e}{T})$ is greater than the random number, then the changes are accepted. If not, the changes are rejected and reversed. Thus, the SA algorithm also allows for sub-optimal changes to occur.

As the simulation progresses, the value of $T$ decreases i.e. the number of records selected for swapping at one time is reduced. According to Kelly (2004), “this process allows a faster rate of improvement in the error term than selecting one record at a time”. There are a maximum of 5,000 iterations in the SA process for SMILE and there is also a ‘restart’ method whereby the SA process begins again with a new initial sample of records. This provides a more in-depth search of the record space by allowing a greater number of household combinations to be explored. Finally, the simulation is deemed to be complete when the value of $e$ is less than some predefined target level $v$. The SA process for SMILE is
summarised in Figure 1 and provides a full simulated population for Ireland, at ED level, with associated personal, socio-economic and other characteristics\textsuperscript{vi}.

[Insert Figure 1 about here]

Overall then, SMILE simulates both household- and individual-level data. In total, 1.29 million households are simulated and the SMILE static process is designed to be consistent with the numbers of households in the 2002 SAPS data, both in total and by ED\textsuperscript{vii}. The number of simulated individuals in SMILE is 3.92 million. However, while each household unit is assigned to a specific ED, this represents the extent of our knowledge of the location of each unit. This has important implications for using spatial microsimulation in a number of contexts. Indeed, Ballas and Clarke (2003) note that “one difficulty at present with spatial microsimulation models is that most of the probabilities are calculated from known distributions (provided by data sources such as the Census of Population) at the small area level (e.g. the ED level in the UK). It is not possible to know precisely whereabouts within a small area such as the ED a particular household … is actually located”.

This issue is illustrated in Figure 2, which shows a selection of EDs in and around Galway City in the West of Ireland, as well as their geographic centroids. Smaller EDs are located within the city and larger EDs in rural areas of the county. In terms of the level of spatial referencing within SMILE, while we know the ED in which each simulated household is located, we do not know where within the ED the simulated unit is based. And since EDs are not uniform in size or shape, nor in terms of population size or distribution, this can present a problem for using SMILE to undertake certain types of spatial analyses\textsuperscript{viii}.

[Insert Figure 2 about here]
To see this consider Figure 3, which adds georeferenced data on the road network and the spatial coordinates of Renville Forest Park (RFP) to Figure 2. Also included is a 16km network-based service (catchment) area representing the geographic extent of the market for recreation at RFP. The edges of this area can be thought of in terms of the choke price for recreation at the site i.e. anyone residing inside the area is a potential visitor, anyone outside is not (the extent of this area was estimated using on-site visitor survey data – see Cullinan (2008) for more details). The problem associated with using SMILE for considering the catchment population of RFP is apparent from Figure 3. There is not a close overlap between the estimated service area and the EDs for which we have simulated population data (or indeed Census of Population data). Therefore, it is not clear which simulated SMILE units reside within the catchment area, which has important implications for using SMILE when conducting spatial analysis in relation to Renville Forest Park. What is required is a more spatially-disaggregated simulated population in order to identify and model the catchment population of the site.

[Insert Figure 3 about here]

3. **Assessing the Compatibility of SMILE and GeoDirectory**

To help address such difficulties, this paper links simulation and location data to develop geohouseholds and uses two principal datasets in the statistical matching process. These are the SMILE base dataset discussed above and GeoDirectory. GeoDirectory is a geodatabase of all buildings in the Republic of Ireland and is developed by An Post, Ireland’s national postal service provider, and Ordnance Survey Ireland (OSi). Its purpose is to create a definitive reference directory of addresses in Ireland and to assign them accurate postal and geographic addresses (OSi & An Post, 2006). Each of the approximately 2 million building
records contained in GeoDirectory (as of July 2007) include a standardised postal address (i.e. one that conforms to a uniform An Post predetermined standard), details about whether the building is used for commercial or residential purposes and the spatial (x,y) coordinates or geocodes which accurately locate the centre point of each building to within one metre. GeoDirectory also identifies buildings that are derelict, vacant, invalid and/or under construction and also distinguishes between residential buildings that are houses and apartments. The data and information contained within GeoDirectory facilitates a range of analyses. For example, Garvey et al. (2008) use GeoDirectory to derive proxy measures for hinterland populations for towns in Ireland in an investigation of the impact of network connectivity on their economic performance. Kalogirou and Foley (2006) use the database to help develop hospital accessibility measures, while Cullinan et al. (2008) use it in conjunction with GIS techniques to derive catchment area population indicators.

GeoDirectory data for County Galway from 2004 is used in this paper as it represents the closest available year to the 2002 Census of Population data that is used in the SMILE static matching process. Buildings classified as commercial only or of unknown usage are excluded from the analysis, as are any residential buildings classified as derelict, vacant, invalid and/or under construction. In total, GeoDirectory contains 87,618 residential addresses for County Galway in 2004. These residences are linked to 80,372 separate residential buildings with associated spatial coordinates.

In comparison, SMILE contains 66,306 private households in total for County Galway. This includes 21,053 households in Galway City (22 EDs in total) and 45,253 households in Galway County (216 EDs). Thus, there are clearly sufficient residential locations in total for all simulated households in the county. The surplus residences in GeoDirectory are likely to represent general excess capacity in some areas, second and/or holiday homes, as well as new...
builds between 2002 and 2004. Every household unit in SMILE and residential building in GeoDirectory is assigned to an electoral division and Table 1 presents some summary statistics in relation to the distribution of simulated household units and residential buildings across EDs. The mean and median number of SMILE households per ED is less than the corresponding average numbers of GeoDirectory residential locations.

A closer examination of the numbers by ED revealed two types of anomalies. First, six electoral divisions contain a greater number of SMILE household units than GeoDirectory residential locations, though in all cases the differences are relatively small, implying a total deficit of 98 dwellings. The second anomaly arises because household or population estimates are not available in SMILE for 4 electoral divisions in County Galway. Therefore, households cannot be simulated for these EDs in SMILE. The implications of these anomalies and how they were dealt with in the matching process are discussed in Cullinan (2008). For the remaining EDs, the number of GeoDirectory residential locations exceeds the number of SMILE households, implying sufficient residential locations for the matching/allocation process in the vast majority of EDs.

To illustrate this, the difference in the number of GeoDirectory residential buildings per ED and the number of SMILE households per ED are compared graphically in Figure 4 for County Galway. Interestingly, the spatial distribution of these differences suggests that most of the excess residential buildings are in the west of the county. This area, much of which is known as Connemara, is a popular tourist area and contains a large number of holiday or second homes. However, large differences are also evident in the more urban centres of Galway City, Tuam, Athenry and Loughrea. This could be a result, in part, of high levels of residential development in these areas between 2002 and 2004, or a pattern of general excess.
supply in more urban areas. Nonetheless, despite these differences, there is a high correlation between the number of households per ED within SMILE and the number of GeoDirectory residential locations per ED, with an estimated correlation coefficient of 0.98. Along with the aggregate numbers in the two datasets, I take this as evidence of a broad consistency between the two datasets for the statistical matching process.

[Insert Figure 4 about here]

4. Matching SMILE Households and GeoDirectory Locations

4.1 Statistical Matching

Statistical matching, also known as data fusion, involves linking or joining data from two or more datasets or samples, each of which contains some data items that are absent from the other. Typically, the information being matched comes from different databases containing similar records and the match is made on the basis of similar characteristics. The statistical matching or data fusion process joins the two surveys by using a matching algorithm or statistical test of similarity. In general, the ‘donor’ survey contains the data which is to be transferred to the second or ‘recipient’ survey (Baker et al., 1997). One of the first examples of statistical matching was Okner (1972), while later examples include Radner (1981), Paass (1986) and Baker et al. (1997).

According to D’Orazio et al. (2006), statistical matching aims to integrate two (or more) datasets where (i) the different datasets contain information on a set of common variables, as well as variables that are not jointly observed, and (ii) the units observed in the data sets are different. Sutherland et al. (2002) simplifies the process as follows: suppose sample A has two variables (X, Y) while sample B contains variables (X, Z). The variable X is common to
both datasets, and under certain assumptions, allows us to create a new dataset C containing the three variables (X, Y, Z) by merging the records of A and B with close values of X. According to Sutherland et al. (2002), “this is legitimate if Y and Z are related to each other only through X, i.e. if, conditionally on X, Y and Z are independent”.

4.2 The Matching Variable and Approach

The aim here is to match SMILE households with GeoDirectory locations. Merging the two datasets depends on the existence of a suitable matching variable, whereby household characteristics and spatial coordinates are independent, conditional on the matching variable. Ballas and Clarke (2003) discussed the idea of joining a simulation database and a location database on the basis of some common variable, stating that it might be possible to build an ‘index of similarity’ between a simulated household and a remotely sensed house. According to the paper, such an ‘index’ could comprise information in relation to the type of dwelling, number of rooms, property size etc. Close values in the index could then be used to match the records from the two datasets using statistical matching techniques. Alternatively, the matching variable could represent a simpler measure of residential building or housing type.

Given this choice of matching variable, define \( S = [ED, H, X] \) as the SMILE dataset containing information on each simulated household’s electoral division of residence \( ED \), housing index/type \( H \), as well as \( X \), a vector of household characteristics. Let \( G = [ED, H, C] \) represent the GeoDirectory dataset containing information on residential buildings including \( ED \), the electoral division of each dwelling, housing index/type \( H \) and the geographic coordinates of the building \( C \). Using statistical matching techniques, it is possible to combine the spatial microsimulation model \( S \) and the GeoDirectory dataset \( G \) in order to derive a new matched dataset \( M = [ED, H, X, C] \) which includes data on the electoral division, housing type, characteristics and geographic coordinates of the simulated
household. Thus, each record of $S$, the microsimulation dataset, is assigned to one record of $G$, the GeoDirectory dataset, in order to obtain $M$. This is illustrated in Figure 5.

![Insert Figure 5 about here]

While SMILE does include detailed information in relation to dwelling type and number of rooms for some observations, this data is unfortunately missing for a number of simulated households within the database. GeoDirectory, on the other hand, distinguishes only between residential locations that are houses or apartments, though this information is available for every observation within the dataset. This lack of consistency between the potential matching variable in the two datasets complicates the matching process. Thus, the approach followed instead is to match or allocate simulated households in each ED to GeoDirectory residential locations in the same ED, based on the probability that a SMILE household unit lives in an apartment, as opposed to a house. This implies that the matching variable $H$ in GeoDirectory is an apartment/house indicator variable, while for SMILE the variable $H$ is an estimated probability.

Overall, the approach involves first estimating this probability for each simulated household, followed by the numbers of apartments and houses in each ED using GeoDirectory. Of the 87,618 GeoDirectory residential buildings in Galway as of 2004, 78,638 (89.7%) were identified as houses and 8,981 (10.3%) as apartments. A significant variation in the number of apartments as a proportion of all dwellings per ED is evident, ranging from 0% to 58.7%, with the highest proportions in urban EDs as would be expected. Simulated households that are most likely to be apartment-dwellers are assigned to GeoDirectory apartments in the appropriate ED, while the remaining SMILE household units are assigned to GeoDirectory houses in the same ED. For both dwelling types, this allocation is undertaken in a random
manner. To assign the correct numbers of SMILE households to apartments and houses requires an alignment process, which is discussed subsequently.

The allocation process presented in this paper could be improved given more detailed information within the buildings database on dwelling type, value and on the number of rooms per residence. This would allow an index of similarity to be constructed, as suggested in Ballas and Clarke (2003). Matching on the basis of the house/apartment distinction implicitly assumes that the characteristics of each household and its spatial coordinates are independent, within each ED, conditional on whether the residence is a house or apartment. This may or may not be true and Sutherland et al. (2002) note that this assumption can rarely be checked in practice. The implicit assumption would, I believe, be more likely to hold in the case where $H$ contained a richer set of information in relation to the type, value and size of residence. Overall however, given the datasets available at this point in time, the house/apartment distinction represents the best criterion on which to base the match.

### 4.3 Event Prediction

The starting point in the allocation process involves modelling whether a simulated SMILE household is a likely apartment-dweller or house-dweller on the basis of a number of explanatory variables, such that:

$$A_i = f(X^h_i, X^{head}_i, \eta_i)$$

where $A_i$ is an indicator variable taking a value of one if household $i$ resides in an apartment and a value of zero otherwise, $X^h_i$ is a vector of explanatory variables relating to the household and $X^{head}_i$ is a vector of variables relating to the head of household $i$. $\eta_i$ is a stochastic error term. The variables used to estimate [1] are listed in Table 2 and include the
number of persons in the household, the net monthly income of the household, as well as the age and gender of the head of household.

[Insert Table 2 about here]

Given the binary nature of the dependent variable, a logit model was estimated via the method of maximum likelihood defining $P_{i}^{Apt} = P(A_i = 1)$ as the probability that household $i$ resides in an apartment. Under the assumptions of the logit model, $P_{i}^{Apt} = \Lambda(X_i' \beta)$ where $\Lambda(.)$ represents the logistic cumulative distribution function such that $\Lambda(X_i' \beta) = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$, $\beta$ is a vector of parameters and $X_i$ includes $X^H$ and $X^{Hol}$ for household $i$. Thus, the log likelihood function is defined as:

$$\ln L = \sum_{i \in N} \ln F(X_i, \beta) + \sum_{i \in N} \ln \{1 - F(X_i, \beta)\}$$

$$= \sum_{i \in N} \ln \Lambda(X_i' \beta) + \sum_{i \in N} \ln \{1 - \Lambda(X_i' \beta)\}$$

[2]

where $N$ is the set of all observations $i$ such that $A_i \neq 0$. Maximisation of [2] provides $\hat{\beta}$, unbiased estimates of the model coefficients $\beta$. It can easily be shown that

$$\ln \left( \frac{P_{i}^{Apt}}{1 - P_{i}^{Apt}} \right) = \logit(P_{i}^{Apt}) = X_i' \beta$$

so that $\hat{P}_{i}^{Apt}$ can be estimated for each household using $\hat{\beta}$ and appropriate values of $X_i$.

The econometric model was estimated using Living in Ireland survey data for 2000 and results are presented in Table 3. They suggest that the probability a household resides in an apartment is negatively related to the number of persons in the household, as well as the income of the household. The age of the head of household is also found to be negatively
related to $P^{Apt}$, though the estimated relationship is non-linear. Overall, the model appears to
fit the data well on the basis of the goodness-of-fit measures presented. Furthermore, the
estimated coefficients are generally of the expected sign. Thus, the estimated model is
considered appropriate for predicting the probability of apartment-dwelling for households
within SMILE.

[Insert Table 3 about here]

The process followed here is an example of ‘event prediction’ in a spatial microsimulation
model, whereby the event under consideration relates to whether a particular household unit
is resident in an apartment or a house. The objective of ‘alignment’ in microsimulation
models is to ensure that the aggregates of such event predictions or outputs match external
‘control’ totals (O’Donoghue et al. 2009). Since the output for each individual in a discrete
choice model is a probability, a decision rule is necessary in order to use the econometric
model for predictive purposes. In essence, a calibration mechanism is required to produce the
correct number of events.

Using alignment requires that predictions need to be ordered, so that the top $x\%$ of
households can be selected for the ‘event’. This process generally entails the addition of
some randomly distributed term to the econometric model’s predictions, since not all of the
variation in the dependent variable is accounted for by the explanatory variables. One
approach is to take a random number from the logistic distribution $\varepsilon_i$, which is then added to
the model’s linear prediction. The inverse logit $\hat{P}_{i}^{Apt} = \logit^{-1}(X_i\hat{\beta} + \varepsilon_i)$ is then calculated
and used to rank households, where:
\[
\text{logit}^{-1}(X_i \hat{\beta} + \varepsilon_i) = \frac{\exp(X_i \hat{\beta} + \varepsilon_i)}{1 + \exp(X_i \hat{\beta} + \varepsilon_i)} \tag{3}
\]

Once ranked on the basis of \( \hat{P}_i^{\text{Apt}} \), the top \( x \% \) of households can then be selected. So, for example, if the control or quota says that 20\% of households reside in an apartment (the event), then the top 20\% of simulated households ranked on the basis of \( \hat{P}_i^{\text{Apt}} \) are classified as residing in apartments. The remaining 80\% are therefore classified as residing in houses.

An alternative to this alignment approach is to instead take a random number drawn from a uniform distribution \( u_i \), which is then subtracted from the inverse logit without the stochastic term, to produce:

\[
\hat{P}_i^{\text{Apt}} - u_i = \text{logit}^{-1}(X_i \hat{\beta}) - u_i = \frac{\exp(X_i \hat{\beta})}{1 + \exp(X_i \hat{\beta})} - u_i \tag{4}
\]

The values of \( \hat{P}_i^{\text{Apt}} - u_i \) can then be used to rank households in the same manner as discussed above. The difference between the two methods is that the first utilises the error structure of the logit model so that individuals are ranked by \( \text{logit}^{-1}(X_i \hat{\beta} + \varepsilon_i) \), while the second ranks by \( \text{logit}^{-1}(X_i \hat{\beta}) - u_i \). However, the two methods produce different rankings.

To see why, assume that a logistic random number \( \varepsilon_i \) is generated using the uniform random number \( u_i \), such that:

\[
\varepsilon_i = -\text{logit}(u_i) = -\ln\left(\frac{u_i}{1-u_i}\right) \tag{5}
\]
or

\[
\hat{u}_i = \frac{\exp(-\varepsilon_i)}{1 + \exp(-\varepsilon_i)} \tag{6}
\]

As a result,

\[
\hat{P}_i^{Apt} - u_i = \logit^{-1}(X_i \hat{\beta}) - u_i
= \frac{\exp(X_i \hat{\beta})}{1 + \exp(X_i \hat{\beta})} - \frac{\exp(-\varepsilon_i)}{1 + \exp(-\varepsilon_i)}
\]

\[
\not= \frac{\exp(X_i \hat{\beta} + \varepsilon_i)}{1 + \exp(X_i \hat{\beta} + \varepsilon_i)}
= \logit^{-1}(X_i \hat{\beta} + \varepsilon_i)
= \hat{P}_i^{Apt*}
\tag{7}
\]

Thus, the rank produced by the two methods will not be the same. As a result, even given the same underlying random variation, both methods will result in different sets of households being selected\textsuperscript{xi}.

A reservation concerning the second method outlined above is that different household units can have different ranking value ranges\textsuperscript{xii}. This is because the random number \(u_i \in [0,1]\) is subtracted from the deterministically predicted \(\hat{P}_i^{Apt}\). Thus, if \(\hat{P}_i^{Apt}\) is found to be small, say equal to 0.1, then the range of possible ranking values will be given by \([-0.9, 0.1]\). On the other hand, if \(\hat{P}_i^{Apt}\) is found to be large, say 0.9, then the range will be \([-0.1, 0.9]\).

A preferred situation is one in which the range of possible ranking values will be the same for all households (O’Donoghue \textit{et al.}, 2009). Since the first approach draws a logistic random variable, it provides a ranking variable \(\hat{P}_i^{Apt*} = \logit^{-1}(X_i \hat{\beta} + \varepsilon_i)\) that takes values in the range
[0,1], irrespective of the value of $\hat{P}_{t}^{\text{Apt}}$. Thus, the first method is preferred here and is the approach followed in calculating the probability that a simulated household resides in an apartment.

4.4 The Alignment Process

To calculate the probability that a SMILE household resides in an apartment, $\hat{P}_{t}^{\text{Apt}}$ is predicted using the estimated coefficients in Table 3, the appropriate values of $Hsize$, $Hhy$, $Age$ and $Sex$ for that household using the SMILE static model and a random number from the logistic distribution $\epsilon_i$. Once $\hat{P}_{t}^{\text{Apt}}$ is calculated for each simulated household, the next step involves an alignment process to ensure that the final numbers of occupied apartments, both in total and within EDs, correspond to a set of external control totals. These totals relate to the numbers of households residing in apartments and in houses per ED.

In order to derive the external control totals, the number of apartments in each electoral division was estimated by running a query in the GeoDirectory database. The number of apartments in an ED gives an upper bound on the number of SMILE households per ED that can be allocated to apartments in GeoDirectory. Since the number of GeoDirectory residential locations generally exceeds the number of SMILE households per ED, it is assumed for the allocation process that equal proportions of houses and apartments are unoccupied within each ED. These proportions are calculated from a comparison of GeoDirectory buildings and SMILE households at ED level and applied separately to the total number of apartments per ED. Once the numbers of ‘occupied’ apartments per ED are estimated, these act as quotas for the numbers of SMILE household units per ED to be assigned to apartments.
Given these external control totals, SMILE households are then ranked in decreasing order on the basis of $\hat{P}_{i}^{Apn*}$. Those households with the highest values of $\hat{P}_{i}^{Apn*}$ are chosen such that the number of SMILE households to be assigned to apartments is equal to the number of apartments in GeoDirectory designated as occupied residential locations. These SMILE households are then allocated in a random manner to precise GeoDirectory apartment locations, using a specially constructed algorithm. Similarly, SMILE households predicted to reside in houses are assigned randomly to the residential houses in GeoDirectory in the same ED. The algorithm has a range of quotas relating to the numbers of GeoDirectory apartments and houses per ED that are occupied, as well as the number of SMILE households per ED. It proceeds by linking units to locations in a random manner subject to the quotas.

Overall, the process involves choosing those household units within SMILE that have the greatest likelihood or probability of residing in an apartment and assigning these, in a random fashion, to spatial coordinates which record the precise location of actual apartments in the GeoDirectory database. Furthermore, those households that are least likely to live in apartments are assigned, again in a random manner for each ED, to residential houses. The resultant spatially-referenced SMILE households are termed geohouseholds and represent the first continuous space representation of a simulated population.

5. Applications

Ballas and Clarke (2003) outlined a number of potential benefits of matching remotely sensed data (such as GeoDirectory) and microsimulation outputs. The principal benefit, they state, is that such a representational framework could potentially allow for a more efficient modelling of spatial interactions between households and agents and other subsystems such as major firms, retailers, financial institutions and schools. One could easily extend this list
to include other facilities such as outdoor recreation facilities (see below), or indeed any public amenity or institution. More specifically, Ballas and Clarke (2003) state that conventional models that aim at modelling the interactions between different zones can be refined using such an approach in order to model the spatial interaction between individual agents. This, they claim, would allow car traffic flows, travel to work flows and shopping trips to be realistically represented and modelled more efficiently\textsuperscript{xiv}. The range of potential applications could also be reasonably extended to incorporate links to cellular automata or agent-based modelling frameworks, where the goal is to model the interactions of multiple agents.

To date, the geohouseholds developed in this paper have been used in a recreation demand modelling context to model the actual and potential demand for outdoor recreation activities at existing and prospective forest sites respectively, as well as their associated total access values (Cullinan, 2008). They have also been used to estimate the number and socio-demographic characteristics of individuals residing within estimated catchment areas of a number of recreation sites and to provide more accurate estimates of travel distances for simulated individuals to the site (Cullinan, 2008; Cullinan et al., 2008). Measurements of travel distances and times represent key inputs in travel cost modelling of recreation demand, as well as in other areas.

To illustrate the benefits of geohouseholds, consider Figure 6. The map shows the distribution of geohouseholds around Renville Forest Park (introduced earlier), and suggests that the geohouseholds provide a better representation of the population within the catchment area of the site than the EDs illustrated in Figure 3.

[Insert Figure 6 about here]
Cullinan (2008) used the geohouseholds developed in this paper to help estimate the total amenity value of Renville Forest Park. In the approach, simulated individuals from SMILE residing in geohouseholds within the catchment area of Renville Forest Park were identified using GIS techniques. An individual-level travel cost model (TCM) of forest recreation demand, adjusted for sampling issues such as truncation and endogenous stratification, was estimated for a sample of visitors to the site. This was then transferred across each individual in the simulated population residing within the estimated catchment area of the site. The SMILE model and GIS-based ‘route analysis’ were used to derive the appropriate values for the explanatory variables in the estimated TCM for each simulated individual. Each simulated individual’s latent demand for visits to Renville Forest Park was then predicted and summed over the population in order to derive an estimate of the total number of visits to the site. This provided a measure of the extent of the market for recreation at the site and was used to estimate the total amenity value of the site\textsuperscript{v}.

The geohouseholds developed in this paper were found to improve the accuracy of visitor number estimates to the forest recreation site. For example, assuming that SMILE households are located at their respective ED centroids as represented in Figure 3 provides an estimate of 24.4 visits per day to the site, compared to an estimate of 31.6 per day using the geohouseholds represented in Figure 6. Thus, when compared to an independent estimate of the average daily number of trips to the site of 35.7 (Brian Cahill, personal communication, 2007), the geohousehold-based estimates were found to provide significantly improved estimates of the total number of trips, in comparison to the centroid-based estimates (Cullinan, 2008). A similar pattern was found when modelling catchment area populations. Using a centroid-based approach, the catchment population of Renville Forest Park was an estimated 53,689 individuals, which was 8,665 (14%) less than the estimate based on geohouseholds. Thus, basing the catchment population estimate on a measure derived using
ED centroids implies a considerably different estimate than one derived using geohouseholds. This is a direct result of the poor overlap between the ED boundaries and the estimated service area. In fact the latter, which is determined by the distances that individuals travel to the site and the road network density in the vicinity of the forest, is unlikely to closely match the ED administrative boundaries. Thus in such circumstances, and where populations are distributed unequally across EDs, an approach based on geohouseholds is likely to provide more accurate estimates for the catchment area populations.

Furthermore, in the context of such applications, it should also be noted that since the geohouseholds derived in this paper are a function of a random allocation process, different sets of geohouseholds will yield different estimates of catchment populations and total trips. However, generating standard errors for these total trip estimates, for example, is an extremely time-consuming task. Preliminary and on-going work conducted on this issue suggests that differences in total trips estimates to Renville Forest Park are not significantly different across different sets of geohouseholds, though no formal statistical tests have yet been undertaken. Indeed, since the SMILE microsimulation base dataset itself is also the outcome of a random process, significant further work remains to be conducted in relation to this issue. This is currently on-going but is beyond the scope of this particular paper.

6. Concluding Remarks

This paper presents a statistical matching approach for assigning households from SMILE, a spatial microsimulation model for Ireland, to unique, spatially-referenced, residential locations. The approach allocates simulated households based on a random assignment after splitting the simulated households into two groups: those predicted to reside in apartments and those predicted to reside in houses. The resulting geohouseholds provide detailed
information in relation to the personal, socio-economic and household characteristics and location of a simulated population. They facilitate a wide range of potential applications in economic and spatial analysis.

A number of caveats associated with the statistical matching approach presented in this paper should be noted. First, the approach assumes conditional independence between the characteristics of each household and its spatial coordinates within each ED and it has not been possible to test this assumption. The matching process could be improved given more detailed information within the buildings database on dwelling type, value and on the number of rooms per residence, allowing an index of similarity to be constructed. Given the datasets available at this point in time however, the house/apartment distinction represents the best criterion on which to base the match. Second, a number of rural EDs did contain any apartments, meaning that the allocation is essentially random for these EDs. This is a weakness in the approach but again can easily be addressed with more comprehensive data. Third, while the SMILE data relates to 2002, the GeoDirectory data used is from 2004 and this has resulted in a small number of anomalies in the matching process. Despite these caveats, the methodological approach set out remains valid and this paper represents the first attempt to develop a continuous space representation of a simulated population.
References


Acknowledgments

The author wishes to thank Stephen Hynes, Cathal O’Donoghue and Eoghan Garvey as well as seminar participants at the Regional Science Association International: British and Irish Section 39th Annual Conference, Limerick, September 2009, the European Association of Environmental and Resource Economists Annual Conference, Amsterdam, June 2009 and the Irish Economic Association Annual Conference, Blarney, April 2009 for useful comments on an earlier version of this paper. Support from Teagasc under the Walsh Fellowship Scheme is gratefully acknowledged.
### Table 1: Number of SMILE Simulated Households and GeoDirectory Residential Locations per ED in County Galway

<table>
<thead>
<tr>
<th></th>
<th>SMILE Households</th>
<th>GeoDirectory Residential Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>281.0</td>
<td>368.1</td>
</tr>
<tr>
<td>Median</td>
<td>160.5</td>
<td>191.0</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>368.9</td>
<td>533.9</td>
</tr>
<tr>
<td>Minimum</td>
<td>30</td>
<td>28</td>
</tr>
<tr>
<td>Maximum</td>
<td>3,608</td>
<td>5,064</td>
</tr>
<tr>
<td>Total</td>
<td>66,306</td>
<td>87,618</td>
</tr>
</tbody>
</table>

*Source: Analysis of SMILE dataset and GeoDirectory data.*

### Table 2: Variable Definitions and Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Sample Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Apartment</em></td>
<td>Indicator variable taking a value of 1 if the household is resident in an apartment</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Explanatory variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Hsize</em></td>
<td>Number of persons in household</td>
<td>3.67</td>
</tr>
<tr>
<td><em>Hhy</em></td>
<td>Net monthly income of household (€)</td>
<td>521.34</td>
</tr>
<tr>
<td><em>Age</em></td>
<td>Age of head of household</td>
<td>52.77</td>
</tr>
<tr>
<td><em>Gender</em></td>
<td>Indicator variable taking a value of 1 if the head of household is female</td>
<td>0.30</td>
</tr>
</tbody>
</table>

### Table 3: Logit Model Estimates – Preferred Model

<table>
<thead>
<tr>
<th>Dep Var: Apartment</th>
<th>Estimated Coefficient</th>
<th>Z-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Hsize</em></td>
<td>-0.3807**</td>
<td>-4.12</td>
</tr>
<tr>
<td><em>Hhy</em></td>
<td>-0.0014**</td>
<td>-2.47</td>
</tr>
<tr>
<td><em>Age</em></td>
<td>-0.2387**</td>
<td>-5.90</td>
</tr>
<tr>
<td><em>Agesq</em></td>
<td>0.0017**</td>
<td>4.16</td>
</tr>
<tr>
<td><em>Gender</em></td>
<td>0.6587**</td>
<td>2.73</td>
</tr>
<tr>
<td>Constant</td>
<td>4.6814**</td>
<td>5.20</td>
</tr>
</tbody>
</table>

N = 4,287 observations  
* significant at 5%; ** significant at 1%  
Pseudo $R^2 = 0.2371$; AIC = 0.157  
Log Likelihood = -330.293  
LR $\chi^2 (5) = 205.25$, Prob $> \chi^2 = 0.0000$
Select SAPS tables to describe population of ED. Select household records from LII households and create working tables.

Compare each working table to each target table. Calculate sum of absolute error ($e$).

Set initial $T$ equal to half the number of households in the ED.

Swap $T$ households.

Calculate change in error ($\Delta e$).

If $\Delta e \leq 0$ or if $\exp(-\Delta e/T) >$ random number (0-1), confirm/save swap and reduce $T$.

If $T = \text{`Stop' T }$ or iterations = “max”, reduce $T$.

If error < target error, save best combination of households and move to next ED.

Source: Adapted from Kelly (2004).
Figure 2: Electoral Divisions in the Vicinity of Galway City, Ireland

Figure 3: Estimated Service Area for Renville Forest Park
Figure 4: Spatial Distribution of the Difference in the Number of GeoDirectory Residential Buildings and SMILE Households in County Galway

![Spatial Distribution Map]

Source: GeoDirectory and data supplied under Licence No. 6155 from OS

Figure 5: Matching SMILE and GeoDirectory

![Matching Venn Diagram]

SMILE (S) GeoDirectory (G)

ED, X H ED, C
Figure 6: Renville Forest Park and Surrounding Geohouseholds

Legend
- Renville Forest Park
- Geohouseholds
- 16km Service Area
- Road Network
- Coastline

Source: GeoDirectory and data supplied under License No. 6155 from OSI
In Ireland, small area population statistics at electoral division level are routinely used to conduct spatio-economic analysis. However, the problems associated with using Irish EDs for spatial analysis are well-known. Indeed, according to Foley et al. (2005), “the current smallest areal unit available at national level, the Electoral Division . . . , is unsatisfactory”. See Cullinan (2008) for more details.

The GeoDirectory data used in this paper is based in part on remotely sensed data.

The second process, known as the dynamic process, ages the population by simulating life cycle characteristics such as demographics, labour market outcomes and migration patterns and is discussed in Kelly (2004a).

The IPF procedure has been used to generate a number of spatial microsimulation models. However, one of the most difficult tasks related to this approach is specifying which variables depend on which others, as well as determining the ‘ordering’ of probabilities (Ballas and Clarke, 2003).

SMILE also contains a number of internal validation measures, including $z$-scores and $z^2$-scores. See Kelly (2004) for a full description.

The process illustrated in Figure 1 is similar to one presented in Ballas et al. (2007) for Micro-MaPPAS, a spatial microsimulation modelling and predictive policy analysis system. Like SMILE, Micro-MaPPAS includes a restart facility to allow for further searching. Unlike SMILE however, it also allows the user to indicate whether particular census variables are more important than others and to weight them accordingly.

The number of households per ED within SMILE ranges from 17 to 7,859, at an average of 376.4 households.

One possibility is to assume that each household resides at the geographic centroid of each ED (red dots in Figure 2) or at some population-weighted centroid.

While all urban EDs contained apartments, a number of rural EDs did not, meaning that the allocation is essentially random for these EDs. This is a weakness in the approach but can easily be addressed with more comprehensive data.

A number of other explanatory variables were considered, including the marital status and employment status of the head of household. The latter was found to be highly correlated with the income variable used and consequently did not add to the overall explanatory power of the model. Variables relating to the region or location type of the household were excluded, since the aim here is to allocate households within a particular ED across that ED. Therefore, location dummies are not required since SMILE households are already assigned to an ED.

As previously stated, however, Ballas and Clarke (2003) did not undertake such a modelling exercise.

The second principal potential benefit of matching locations and simulated households outlined by Ballas and Clarke (2003) relates to the potential for business applications. It would be useful to know, for example, “where particular household distributions are within the ED so that particular streets or blocks of streets could be targeted (say for postal mail shots) rather than every household in the ED”.

Finally, the third benefit relates to the increased potential of remotely sensed data, such as putting “estimations on the types of buildings in terms of housing types and characteristics of their inhabitants. Clearly, it is not possible to categorically say what types of families were in each building. However, it may be possible to give an estimation of the types of families within blocks thus giving very detailed portraits of small areas of our cities”.

Cullinan (2008) also used the same general modelling approach in an environmental value transfer context to estimate the potential amenity value of prospective new ‘policy’ forest recreation sites in Ireland.