Automated Boundary Creation: Atomic Small Areas in Ireland

A Stewart Fotheringham, Peter F Foley, Martin Charlton

National Centre for Geocomputation, National University of Ireland Maynooth

Abstract. This paper describes the creation of a set of small-areas for the reporting of census data in the Republic of Ireland. The current areas used for reporting the results of the quinquennial population censuses are known as Electoral Divisions; they are large compared with similar reporting areas in Northern Ireland, they have widely varying populations and considerable internal social heterogeneity which makes them unsuitable for a wide variety of planning tasks. We describe an automated method of creating a suitable census geography which uses existing digital map and gazetteer data. We describe its structure and operation, validation and its application to nationwide. The areas have a prescribed minimum size, are designed to be consistently small, nest into the existing ED geography, cover the whole country, are constrained by natural boundaries, use streets as their unifying feature, and are reasonably homogenous.

Keywords: census, electoral divisions, gazetteer, spatial patterns

1 Introduction

In many countries census data is reported for a set of zones whose extent covers the entire country. Such zoning systems are usually hierarchical in nature with the upper tiers of the hierarchy being administrative or governmental units. In the United Kingdom the areal units used to collect census information (Enumeration Districts) are different from those used to report it (Output Areas) although this was not the case until the 2001 Census – there are 116895 Enumeration Districts in England and Wales with 175434 Output Areas. The Output Areas have a minimum size of 40

households and 100 residents, although the recommended size is 125 households. Northern Ireland is part of the United Kingdom, with 5022 Output areas and 2591 Enumeration Districts with an average size of 260 households or 650 residents (ONS 2007).

The situation in the Republic of Ireland is slightly different. The areas used for the collection of the quinquennial census are known as Enumeration Areas, but the areas used for reporting the data are known as Electoral Divisions (EDs). EDs nest into adminstrative counties. There are 3440 EDs with an average size of 1145 residents or 376 households. EDs very greatly in size; in the published Small Area Population Statistics the largest has 24400 residents and the smallest 55 (or 7859 and 17 households). One ED is split between two counties; 32 EDs have low populations (<55) and their data is amalgamated with an adjacent ED. The number of EDs in the SAPS is 3409. The average household size across the Republic is 3.04 residents per household, although there is wide variation across the EDs from 1.81 to 12.60 residents per household. Not only do the EDs vary wildly in size, there is considerable internal social heterogeneity. The ED boundaries have been stable for many years, and the recent boom in the Irish economy (the 'Celtic Tiger') has generated a rapid demand for new housing. The result is that increasingly EDs are an unsuitable spatial unit for attempting to understand local social changes, particularly in urban areas, and especially for targetted spatial delivery of policies designed to alleviate social problems.

A related issue is in recent efforts to consider social patterns in both Northern Ireland and the Republic. The spatial granularity of the census reporting units in the two countries inhibits reliable joint analysis of census and other information due to the different scales involved. There are notable border effects when regression modelling is attempted with ED data in the south and Output Area or Ward level data in the north.

The question arose as to whether a set of atomic small areas (SAs) could be constructed which would become basic collection and reporting units. Such areas would require a consistent definition, would be large enough to avoid any confidentiality issues, be small enough to allow social micrography to be uncovered, and be reasonably similar in size. It would be desirable if they be created from existing data sources, and should be capable of long term maintenance. They should also form a complete partition of the Republic, and they should nest into the existing ED boundaries.

2 Design Issues

Designing a new grography is not a task to be approached lightly or wantonly, and initial thoughts were whether the approach used in the United Kingdom might be employed to create a set of SAs. The underyling spatial building block for the OAs is the unit postcode. The postal and administrative geographies in England and Wales are misaligned. For the 1981 and 1991 Censuses many postcodes straddled Enumeration District boundarie. In 1991 OPCS created a lookup table (OPCS/GROS 1992) which provided population counties for every intersection of postcode and Enumeration District as well as the 'majority' ED for each split postcode: the postcode is a convenient surrogate for a full address and the majority linkage provided a method of assigning postcoded records (whether for households to individuals) to EDs on an 'all-or-nothing' basis. The unit postcodes are also misaligned with larger units in the administrative hierarchy.

The solution for the UK exercise was to create 'postcode polygons' (Martin 1998) which contain those postal delivery points in a single postcode which do not cross a parish or ward boundary. Thiessen polygons were created around individual delivery points (using the ADDRESSPOINT gazetteer from the Ordnance Survey) which were then merged within each postcode/ward intersection. These basic spatial units are then aggregated into larger units with the desired characteristics (100 residents and 40 households) and social homogeneity (Martin 1998). The algorithm which was used to aggregate the postcode polygons is modified version of the Automatic Zoning Procedure (Openshaw 1977). A further design modification allowed for the minimisation of the perimeter^2/area ratio of the resulting polygons, although in the final allocation, the shape criterion has been based on minimising the dispersion of postcodes in each polygon.

This approach creates a problem for the Republic of Ireland as it is one of the few developed countries in the world which does not have a postcode system: this rules out the postcode as a building block. Like the UK, there is a gazetteer of postal delivery points which are geocoded using two projection systems (Irish National Grid and Irish Transverse Mercator) which is known as GeoDirectory (Fahey and Finch 2007). Initial thoughts prompted by a suggestion that a street-based allocation might be possible considered whether groups of delivery points could be created which would match the initial design criteria above.

3 Stage I and II Pilot

Initial design and testing was carried out on two EDs in Northern Kildare, Maynooth and Leixlip. In 2002 Maynooth had a population of 10387 with 3199 households (average size 3.25) and Leixlip had 15154 residents in 4430 households (average size 3.42). Both are commuter towns in North East Kildare about 20 miles west of Dublin, both have urban and rural parts, and Leixlip is home to Intel's microprocessor fabrication facility, a major local employer, and Maynooth is the location of the National University of Ireland Maynooth.

After discussions with the Central Statistics Office, it was decided that the minimum SA size would be fixed at 65 households. There would be no cap on the SA size – this would be determined by the algorithm depending on individual cases. However, as individual geocoded Census records are not available, the residential delivery point would be used as the surrogate for a household: the minimum SA size is 65 residential delivery points.

3.1 Basic Strategy

The initial approach was to join road centrelines into skeletons where the segments forming the centrelines were tagged with the number of residential delivery points nearest to that segment. For each skeleton a list was made of segments which could join it; one segment was chosen from this list and joined to the skeleton. Joining terminated once the property count for the skeleton reached 65. The initial segment was chosen at random from the list of unallocated segments in the ED. The criterion for joining was to chose either (a) the next unallocated segment with the most delivery points, (b) the segment with the fewest delivery points, (c) the longest segment or (d) the shortest segment. These choices correspond to using as attributes the delivery point count or the segment length and with either a 'greedy' or 'abstemious' option.

A FORTRAN program was written to carry out the allocation allowing for the various options. Three outcomes from this process arose. First, a completed skeleton is created with more than 65 households. Second, it is possible for the remaining unallocated segments to have insufficient delivery points to reach the acceptance threshold: these were referred to as 'orphans'. Third, in some cases segments do not join with any other segments: these were called 'singletons'. Orphans were dealt with by redistributing their segments among the already-created skeletons according to the creation criterion (delivery point/segment length and greedy/abstemious). Singletons then had to be merged with neighbouring areas.

The outcome of the program is a list of segments and skeleton codes. The segment which is closest to each delivery point (both residential and commercial) is already known, so the skeleton codes can be transferred to the delivery point locations. Proto-small areas are then formed by creating Thiessen polygons for all the delivery points inside an ED, constrained by the ED boundary, and merging the internal boundaries between those delivery point polygons which have the same skeleton code. Thiessen polygons created for singleton skeletons are merged with the neighbour that shares the longest boundary.

Some GIS operations are required to extract the data from various data sources, integrate it, dump it into a suitable format for the FORTRAN allocation program, and then assemble the pieces back together to form small areas. Consideration was given to the choice of software platform – the final decision was made to use ESRI's ArcINFO product, and code the GIS operations into a set of linked macros using the Arc Macro Language (AML).

Examination of the alternative approaches, together with mapping geocoded household data made available from the Central Statistics Office, suggested that using the residential delivery point counts as the criterion attribute and the greedy allocation strategy produced the most satisfactory set of boundaries.

3.2 Addressing

Most properties in urban areas in Ireland have what we might term a 'well formed address'; that is, an address which uniquely identifies the property. This might be some combination of a number or house name, street name, locality name, and district name. However, this is not always the case, particularly in rural areas. In some small villages the road running through the village does not have a name, and the houses along it have neither numbers nor names; the address for each delivery point is just the name of the settlement. It is due to the local knowledge of the postman that letters are delivered to the right households, although the arrival of a new postman to an area can cause problems. The problem of non-uniqueness of address is greatest in the most rural areas; an examination of the addresses in GeoDirectory suggests that some 66% of addresses in the county of Roscommon are not unique. Whilst GeoDirectory does indicate whether an address is unique or not, the BUILDINGS table does indicate whether a property is on a named throughfare. Most unnamed thoroughfares are in rural areas.

3.3 Incorporating Natural Boundaries

Some enhancements were sought. While the SA boundaries nest into their parent ED boundaries, they do not take into account 'natural' boundaries such as watercourses or railway lines. The question of modifying the algorithm to take into account these boundaries was then explored.

The Thiessen polygon algorithm in ArcINFO takes a set of points and returns a set of polygons which have Thiessen properties which can be clipped using the ED boundary. The locations of the polygon boundaries cannot be influenced – there is no way of modifiying the algorithm.

A raster equivalent is to use the costallocation function in ESRI's GRID module together with the locations of the delivery points to produce an allocation grid in which any cell is closest to the cell representing its residential delivery point and no other. Creating the grid of residential delivery points is easy – the pointgrid function is used. The other input required is a cost grid – each cell contains the cost of traversing one unit of distance. If every cell in the cost grid contains the same value, the allocation grid which is output will correspond to a rasterised version of the Thiessen polygon vector coverage of the type created in the pilot. The key to modifying the allocation is to introduce varying traverse costs to represent the 'importance' of the various natural boundaries. Initally all cells in the cost grid were set to 1, and then those which were crossed by watercourse, a railway line, or a main road were set to 10000. The resulting output grid is then vectorised to obtain polygons.

Clearly the raster size is important here – larger cells require less processing than smaller cells, but smaller cells are closer to the original boundaries when the output allocation grid is vectorised. Raster sizes of 1m and 2m were used. This raised problems which will be discussed further below.

The watercourse, railway and road centrelines were rasterised from Ordnance Survey Ireland’s large scale data using the linegrid function. This caused problems – for instance, the Royal Canal which flows through Maynooth is represented by separate lines for its north and south banks. Each track on the railway that runs along the canal is represented by a separate line. In essence, this level of data provided too much detail for the polygon formation, and a decision was made to use the representations from OSi's 1:50000 vector data where watercourses have, counterintuitively, a richer feature coding, but are represented by single centrelines.

3.4 Segments without Delivery Points

During testing it became clear that there were some segments in the road centreline data that lacked delivery points. In general these were outside the urban areas, and usually along the longer segments. This resulted in some of the rural skeletons growing in rather unexpected ways. This was a problem whether the delivery point count or segment length was used as the choice criterion.

A solution to this problem was to treat the 'urban' small-area and 'rural' small area formation as separate tasks and then merge the 'urban' and 'rural' proto small areas. The question arises of deciding what parts on an ED are urban and which are rural.

There are smaller officially-defined geographical areas in Ireland than EDs which are known as Townlands. There are some 50000 of these and their boundaries reflect a pre-medieval division of the landscape. Like EDs townlands vary widely in size, shape, and population. In most cases they nest into EDs, but their boundaries are not always aligned. Intersecting the townland and delivery point coverages allows us to count the numbers of delivery points on both named and un-name throughfares in each townland. Examination of the results for the pilot areas suggested that if any delivery point in a townland was on a named thoroughfare, then the townload could be regarded as urban, otherwise it was treated as rural.

3.5 Rural Allocation

The initial processing of the data for an ED was to decide which were the rural townlands. The townland residential delivery point counts and a townland adjacency matrix were extracted from the ArcINFO coverages in ASCII form and passed to a FORTRAN program. It was decided that a brute force approach would be taken. A townland is selected at random and its un-allocated neighbours examined. The one with the largest property count was joined. The unallocated neighbours of these fused areas are examined and the one with the largest property count is added until the delivery point count reaches the threshold of 65. This continues until all townlands have been processed. The mean and variance of the delivery counts is computed. The process is iterated from different random starting points until the solution with the lowest mean and variance is obtained (this is usually in fewer than 10000 iterations).

Analogous allocation problems occur with this method as with the skeleton building process. Orphan and singleton allocations appear. Orphans are dealt with by reallocating their component townlands among any

neighbours. Singletons are flagged to be treated using the skeleton building method.

3.6 Rural/Urban Merging

With the development of the separate procedure for handling the rural parts of each ED the final set of operations require merging of the separate sets of proto small areas.

3.7 Random Numbers

In the pilot phase the system random number generator was used – the software runs on a Sun workstation under Solaris 5.9. It is clear that applying this process to the whole of Ireland will require enormous streams of reliable random numbers. The programs were redesigned to use to Mersenne Twister (Matsumoto and Nishimura 1998). This has been shown be reliable in operational situations using spatial data (van Niel and Laffan 2003). The seeds are set using the system clock so every run will use a different stream of random numbers.

3.8 Testing and Validation

Producing a prototype algorithm is just a starting point. An algorithm which appears to work on two EDs out of over 3400 might not be expected to work in a wider variety of contexts. Some EDs are entirely 'urban', some EDs are entirely 'rural'. Some EDs have interesting geometry. Nine additional Eds were selected to test the algorithm for stability. The test EDs are listed below in Table 1.

Table 1. Test ED characteristics

ED Type	ED Name(s)
Rapidly expanding commuter town	Maynooth (Kildare)
	Leixlip (Kildare)
Old inner city area	Merchants Quay A (Dublin)
New urban development	Ashtown A (Dublin)
Large county town	Longford Urban No. 1
Small rural town	Abbeyleix (Laois)
Rural	Ardamine (Wexford)
Rural with holiday homes	Moy (Clare)
Island community	Inishmore (Clare)
Unusual geometry: Doughnut	Kilkenny Rural
Ground truthing	Botanic A (Dublin)

Whilst the initial algorithm had been developed using Maynooth and Leixlip, these areas were included in Stage II to assess the impact of the refinements. Both Eds are in areas of very rapid urban expansion, and Maynooth is an ED with many student houses. Merchant's Quay is in the Liberties area of Dublin – an old inner city area which has been gentrifying but has a good mix of old corporation housing, new apartments, and some low value older owner occupied houses. Ashtown A by contrast is undergoing rapid development on green/brownfield sites although there are some older estate developments; it lies on the edge of Dublin. Abbeyleix in a classic medium-size rural ED with a town in the centre. Inishmore is a challenge: it contains the three Aran Islands with a combined population of about 1300. Kilkenny Rural is one of several polygons which surrounds completely one or more EDs – these doughnut polygons were to prove an interesting challenge. Botanic A was the subject of initial manual exploration: it lies in north-central Dublin.

The results of the application of the algorithm to the test areas is in Table 2:

Table 2. Summary of Small Areas for Test EDs

ED	Small Areas	Mean Size	Max Size	Min Size	Total H/holds
Leixlip	35	137	250	70	4825
Longford	13	126	215	74	1629
Maynooth	33	118	226	66	3887
Moy	4	107	133	80	428
Merchant's Quay	5	200	308	103	1002
Inishmore	6	109	198	65	656
Abbeyleix	8	134	217	80	1071
Kilkenny Rural	40	132	391	68	5311
Ardamine	13	137	283	80	1793
Ashtown A	17	133	266	66	2261
Botanic A	11	124	174	68	1359

The table shows some summary statistics for the EDs chosen as the test areas. As well as the ED name, the table reports the number of Small Areas created within the ED, the average number of residential delivery points in each Small Area in the ED, the minimum number of delivery points, the maximum number of delivery points and the total number of residential delivery points in the ED.

It is clear that there's quite a strong and re-assuring relationship between the number of small areas which the algorithm produces and the number of households in an ED. The average number of households is roughly on a par with that average for the Output areas in Northern Ireland. The minimum threshold of 65 is not breached. As expected there are anomalies.

The statistics for Merchant's Quay are skewed by the number of apartment blocks along a small number of street segments. However, the initial impressions of this testing stage were that the algorithm was robust and was able to produce sensible results.

3.9 From Prototype to Production

The transition from a two stage pilot to the full roll out took 18 months. Applying the algorithm to 11 EDs was a relatively simple process, and the initial estimates were that, given the average processing time on one of the authors' laptop, that running the whole county on a Sun workstation would be relatively straightforward. This proved not to be the case.

The prototype of the algorithm consisted of a series of loosely linked macros and two fortran programs. The data from GeoDirectory has been pre-processed and the extracts from the buildings table made for each ED. The macros were run from a simple GUI where the user clicked on an ED name and a macro was called which ran macros to extract the road centrelines, natural boundaries, townland boundaries, dumped the data, ran the external programs, and then merged the results back together creating a coverage/shapefile called small-areas, and some two dozen intermediate coverages which were used in processing.

It was decided that the EDs in an entire county would be processed in a single run. The results would be stored in a subdirectory named after the county, and underneath would be separate subdirectories for each ED. Within the ED subdirectory would be the extracted data, intermediate coverages and grids, and the final coverage named small-areas, so that when problems arose, we could track backwards through the formation process for any ED in the county or country.

The data for the whole country was conveyed to the NCG on a 500Gb disk drive – 212Gb of the drive were the large scale orthophotographs which would be used during the verification stage of the process. Handling such large amount of data requires some thought. Copying the data took several hours, and the orthos were reduced in physical size by conversion from TIFF to JPEG format – this took over 8 hours.

The macros were rewritten to make them amenable to batch style processing with the name of the county as a parameter. After some early setbacks it became clear that we would need to retain all the printed output from the ArcINFO commands. These are stored in 'watchfiles' which can grow unexpectedly large.

4 Production Algorithm

The implementation is through a set of linked macros and four fortran programs. The macros total some 1300 lines of ArcINFO AML, and the fortran programs are just short of 3000 lines of code. There is extensive pre- inter- and post-processing of the outputs from the various fortran programs in the GIS. Some of this arises because of the requirements of the fortran, other manipulations are required because of the architecture of the small-areas creation method, and a final group of manipulations might be described as housekeeping – these are due to the data representation in ArcINFO.

The allocation for the urban and rural parts of the EDs has been described above. It is useful to consider how the various component operations are brought together, and what parameters are needed to control the system.

For any ED there are 90 separate GIS operations required, as well as the running of four external programs, to create the final set of small-areas. The editing phase is then followed by a final merge of the EDs in each county to create a county set of small-areas.

There is a small set of utility routines, in particular, a clipping macro. The clipping macro takes an ED boundary and returns points, lines, or polygons which lie within the ED boundary, and also removes any internal polygons which are not part of the ED – this assists in the processing of doughnut polygons. As with many 'simple' operations provided as part of GIS software, the desired result is often the outcome of a series of linked operations. The clipping routine is an example of a generic routine, so it was coded separately rather than the operations being coded 'in-line' in the main processing macro.

4.1 Data Extraction

There are several components required for the creation of the small areas within an ED. The townland and ED boundaries are misaligned – they do not form a neat spatial hierarchy. At the county edges, there are slight misalignments in the source data which create thousands of sliver polygons – these must be removed before further processing continues. After the application of the clipping routine, slivers of less than $1m^2$ are removed but the processing is organised so that the external edge of the clipped townland coverage remains coterminous with the ED boundary. This requires 8 separate GIS operations, including an edit session to remove the

resulting pseudo-nodes (in the ArcINFO terminology nodes with a valency of 2 are referred to as pseudo-nodes).

The road centreline segments are extracted from the national road centrelines data – this is provided as part of the 1:50000 scale data, but the centrelines are based on 1:1000 and 1:2500 source. The building centroids are extracted from the GeoDirectory coverage.

An important parameter is the spatial tolerance. In the ArcINFO model this is known as the fuzzy tolerance and determines when nodes will be snapped and whether coordinates are moved. As positional accuracy is important, a spatial tolerance of 0.001m is used. This ensures that the final boundaries can be adjusted to be coterminous with the supplied ED boundaries.

A parallel operation is the extraction of a list of orthophotographs which cover this ED. The orthos are used in plotting the results for individual EDs. There are 25500 orthos covering the extent of the Republic, so for any ED only a handful are required as backdrops. As part of the initial data preparation for the project we created an index coverage which contains the name and extent of every ortho. This is intersected with the ED boundary to obtain a list of the orthos which cover the ED.

4.2 Rural Townland Processing

Initial experiments using the EDs in Kildare as a test bed revealed that the original criterion from Stage II for the identification of 'rural' townlands was too crude. After some investigation it was decided that two thresholds were important: the proportion of delivery points which were not on named throughfares and the total number of delivery points in each townland. Townlands with more than 27.5% of delivery points on unnamed thoroughfares and with fewer than 232 residential delivery points are deemed to be rural, as is any townland with no delivery points. These thresholds were determined after analysis of Townland characteristics in County Kildare. Kildare contains a wide range of settlement types from very small villages to large commuter towns.

The centres of doughnut EDs have to removed – they are not considered as part of the outside polygon in the ArcINFO model so have to be separately flagged at each stage in the processing. The polygon attribute table (which contains the ID of each rural townland) and the arc attribute table (which contains the ID of the polygons on each side of any boundary) are dumped into ASCII files for input into the townland allocation program. The output is a list of proto-small areas codes and townland IDs. These are merged back with the original rural townland coverage, and internal

boundaries between adjacent townlands in the same small area are dissolved to create the set of rural proto small-areas.

4.3 Urban Townland Processing

Urban townlands and rural townlands flagged for processing as urban are extracted from the townlands coverage, and used to clip the road centreline coverage and the buildings coverage. The next stage is data pre-processing for input into the urban centreline allocation program.

The throughfare codes in GeoDirectory do not match any segment coding in the road centreline data. This precludes a simple tally of residential and commercial delivery points over road segments. A proximity analysis is carred out to determine the ID of the closest road segment to each residential delivery point, and the number of delivery points is then tallied over the segment IDs. These counts are merged back with the road centreline segment coverage.

A problem arises in a few cases where the proximity analysis misallocates building locations to an incorrect road segment – this usually occurs at road junctions. A re-allocator program was written in C++ to correct the misallocation – correction is possible in about 50% of cases.

Records from the arc attribute table for the roads coverage are dumped to an ASCII file – this contains, for each segment, the IDs of the start and end nodes of the segment, the ID of the segment itself, its length, and the number of residential delivery points. The skeleton creation program takes these as input and produces an ASCII file with the segment IDs and their associated skeleton IDs which is merged back with the road centrelines coverage.

The proximity analysis for the buildings also gives a segment ID to each delivery point, so the skeleton IDs are also merged with the buildings coverage – we now know in which skeleton each building lies – this includes commercial buildings as well as residential buildings.

4.4 Raster Processing

The next stage is to form the urban proto-small areas. As these are based on constrained Thiessen polygons, this stage is carried out in a raster environment. There is a rich collection of raster processing functions, based on Tomlin's Map Algebra (Tomlin 1990), available in the GRID module which is part of ArcINFO.

A fundamental aspect of raster processing is deciding on the size of the raster to be used. Large cells can be processed quickly, but suffer from low

spatial resolution when they are re-vectorised. Small cells take longer to process, halving the side length quadruples the number of cells, but creates vectors which it was thought can be more easily merged back into the townland based results. We shall discuss this problem further below. After some experimentation, 1m was chosen as the raster size.

The urban townlands are extracted from the boundary and rasterised at 1m relative to the lower left corner of the bounding box of the ED. This will be used as both a window and a masking grid. The urban parts of the rail, motorway, primary road, watercourse, and path coverages are clipped (using the clipping routine described earlier) and rasterised – cells through which any of these constraints pass are given a passage cost of 10000 and all others are given a passage cost of unity. The resulting grid is the cost grid. The buildings coverage is finally rasterised, with the skeleton code as the attribute.

The building skeleton code grid and the cost grid are used as arguments to the costallocation function. The output from costallocation is a grid in which each cell contains the skeleton code of the building to which it is closest. Vectors are then recovered from this grid using the gridpoly function. This gives us the vector boundaries of the feature constrained Thiessen polygons. The polygon IDs are adjusted by the addition of 30000 to differentiate them from the rural proto small-area polygons when the two sets of polygons are merged.

The result of vectorising the raster data is that lines which are not vertical or horizontal are created from a set of 'steps'. Aesthetically these are unpleasing when seen close-up, so they are smoothed to remvoe the worst effects. The spline operation is used in ArcEDIT to accomplish this; the requisite parameter is the grain tolerance, which is set at 2.5m.

4.5 Merging the Urban and Rural Proto Small Areas

The final stage in the algorithm turns out to be as complex as any of the previous steps. There are several special cases which require careful handling, in particular the preservation of the doughnut hole, uninhabited islands, and sliver polygons remaining after the vectorisation which have not been sufficently smoothed.

The proto small area coverages from the urban and rural processing are merged. Remaining slivers from the re-vectorisation are removed if they within a 2.5m buffer of the ED boundary, and the ID of the doughnut is adjusted to preserve its 'external' status. A final tally of residential delivery points is made for the proto small-area coverage which is then merged with the proto small-area coverage polygon attribute table.

Although the result of the spline smoothing in the vectorisation of the urban small area rasters is designed to produced asethetically pleasing results, the boundaries may not be conterminous with the original spatial constraint (rail, road, watercourse, path) locations, nor with townland boundaries. The ArcEDIT module allows a snapping operation to take place, such that any feature which is closer than, in this case, 1.415m of a snapfeature can be re-aligned with that snapfeature. Small area boundaries are snapped to roads, watercourses, railway lines, paths, urban townland boundaries, and finally to the ED boundary.

At this stage there may be a small residue of polygons with residential delivery point counts which are below the threshold of 65. These are merged with the adjacent polygon with which they share the longest border. Care has to be taken here to preserve the external and any doughnut boundaries. There may still be islands with a delivery point count lower than 65 – these are flagged. There are also a few complete EDs with a delivery point count which fails the threshold; these are flagged as well, as they will consist of only one small area.

The final stage is to assign small-area codes. Each ED has a 5 or 6 digit code of the form CCEEE or CCCEEE where the CC or CCC is a county code, and the EEE is a sequence number of the alphabetic order of the ED within the county. These codes are multiplied by 1000 and the sequence number of the small-area is added. Although the largest ED has 92 small areas, it may be that further building expansion within this ED results in more than 100 small areas – the coding system needs the redundancy to allow this.

4.6 Aesthetic Improvement

A final stage, which unfortunately is somewhat time consuming is to remove artefacts which arise out of the application of a completely automatic algorithm. These can be quite bizarre – a river and a railway running parallel to one another can produce something which looks like a heron's beak sticking from the side of a polygon. Some polygons have a bowtie shape, often in rural areas where a townland with zero delivery points has been added to one small area rather than another. These artefacts must be identified and removed by hand. A final external program computes a variety of shape statistics (Folk 1968, Moellering and Rayner 1979). Analysis of the distributions of the shape statistics reveals some helpful thresholds for identifying the various eccentrically shaped polygons which are candidates for manual adjustment. At the time of writing this process is being undertaken.

4.7 Time Requirements

A final consideration is the time required for the production. Processing is being undertaken on a Sun Blade 2500 workstation with an UltraSPARC chip – the CPU clock speed is 1.28GHz. The data are stored on a Dell PowerEdge 2950 server with 1TB of disk store. Kildare, which has 92 EDs takes around 8 hours to process, Dublin takes nearly 50 hours, Kerry, with fewer EDs but a more complex coastline takes nearly 60 hours.

5 Discussion

The preliminary results are shown in the table below and represent work in progress – there is much work to be completed before the final set of small Areas can be released.

Table 3. Preliminary results

County	Eds	Pop'n	H'holds	SAs	Pop/SA	HH/SA
Carlow	54	46014	17195	194	237	89
Dublin	322	1122821	420429	4092	274	103
Kildare	89	163944	60957	635	258	96
Kilkenny	113	80339	29651	341	236	87
Laois	98	58774	22591	249	236	91
Longford	55	31068	12111	170	183	71
Louth	42	101821	38703	423	241	91
Meath	92	134005	53938	604	222	89
Offaly	87	63663	23769	264	241	90
Westmeath	106	71858	27064	322	223	84
Wexford	124	116596	45566	621	188	73
Wicklow	82	114676	42870	467	246	92
Clare	155	103277	38210	485	213	79
Cork	398	447829	167234	1900	236	88
Kerry	166	132527	48110	674	197	71
Limerick	173	175304	64225	749	234	86
Tipperary	175	140131	52367	645	217	81
Waterford	130	101546	38580	448	227	86
Galway	238	209077	78661	1011	207	78
Leitrim	78	25799	10646	164	157	65
Mayo	154	117446	43431	636	185	68
Roscommon	112	53774	20734	289	186	72
Sligo	82	58200	21480	302	193	71

Cavan	93	56546	21929	297	190	74
Donegal	149	137575	50415	719	191	70
Monaghan	70	52593	18655	255	206	73
RoI	3437	3917203	1469521	16956	231	87

While these are preliminary results, the average sizes of 231 residents and 87 households compare well with the equivalent areas in Northern Ireland which have an average of 326 residents and 125 households. The apparently anomalous result for Leitrim (65 per small area) is because 9 of the EDs already fail the household count threshold. Figure 1 shows the existing Electoral Division boundaries in County Kildare which extend about 70km north-south and 50km east-west. The small-areas created using the algorithm described in this paper are shown in figure 2. The detailed subdivision in the urban areas is quite clear.

Fig. 1. Electoral Divisions in County Kildare

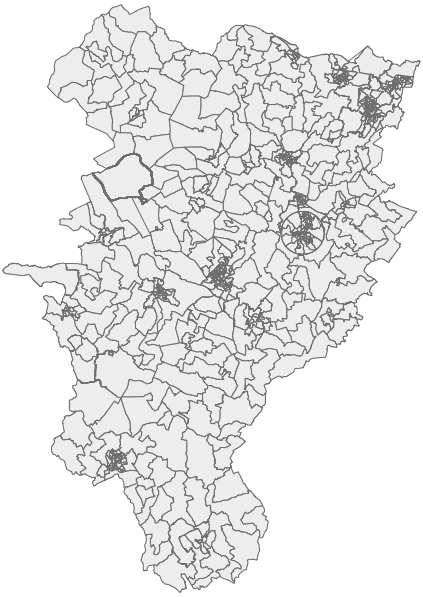

Fig. 2. Small Areas in County Kildare

A bespoke algorithm and methodology has been developed to generate small areas automatically with Electoral Divisions in Ireland. The method produces robust and sensible results. The small areas meet the design criteria and have a number of attractive features

- In urban areas the small areas are based on communities
- In rural areas the small areas are based on historic spatial units
- In urban areas streets are cohesive rather than dividing features
- The small areas boundaries take into account natural features
- The small areas are large enough for data to be reported without breaking any confidentiality thresholds. Where an ED fails to reach this threshold it will contain only one small area, and current official practice is to merge this and an adjacent ED for data reporting.
- The small areas are spatially similar to the Output Areas used in Northern Ireland which will greatly facilitate the creation of all-island datasets and encourage all-island analysis.

References

Fahey D, Finch F (2007) GeoDirectory Technical Guide. An Post/Ordnance Survey Ireland, Dublin

Folk RL (1968) Petrology of Sedimentary Rocks. Hemphill's, Austin TX

Martin D (1998) Census output areas: from concept to prototype. Population Trends 94:19-24

Matsumoto M, Nishimura T (1998) Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Transactions on Modeling and Computer Simulation 8(1):3-30

Moellering H, Rayner JN, (1979) Measurement of shape in geography and cartography, Numerical Cartography Laboratory Report No SOC77-11318. Ohio State University

van Niel K, Laffan S (2003) Gambling with randomness: the use of pseudo-random number generators in GIS. International Journal of Geographical Information Science 17(1):49-68

ONS (2007) Census Geography, URL:http://www.statistics.gov.uk/geography/census_geog.asp

OPCS/GROS (1992) ED/Postcode directory: Prospectus, 1991 Census User Guide 26. Office of Population Censuses and Surveys, Titchfield

Openshaw S (1977) A geographical solution to scale and aggregation problems in region building, partitioning and spatial modeling. Transactions of the Institute of British Geographers, NS 2:425-446

Tomlin CD (1990) Geographic Information Systems and Cartographic Modeling. Prentice-Hall, Englewood Cliffs, NJ.