Journal of Statistical Software

November 2015, Volume 68, Issue 2. *doi: 10.18637/jss.v068.i02*

sms: Microdata for Geographical Analysis in R

Dimitris Kavroudakis
University of the Aegean

Abstract

Spatial microsimulation is a methodology aiming to simulate entities such as households, individuals or businesses in the finest possible scale. This process requires the use of individual based microdatasets. The package presented in this work facilitates the production of small area population microdata by combining various datasets such as census data and individual based datasets. This package includes a parallel implementation of random selection with optimization to select a group of individual records that match a macro description. This methodological approach has been used in a number of topics ranging from measuring inequalities in educational attainment (Kavroudakis, Ballas, and Birkin 2012) to estimating poverty at small area levels (Tanton, McNamara, Harding, and Morrison 2007). The development of the method over recent years is driving computational complexity to the edge as it uses modern computational approaches for the combination of data. The R package **sms** presented in this work uses parallel processing approaches for the efficient production of small area population microdata, which can be subsequently used for geographical analysis. Finally, a complete case study of fitting geographical data with the R package is presented and discussed.

Keywords: microdata, geographical analysis, spatial microsimulation, R.

1. Spatial microsimulation

Individual based models are among the most popular tool sets for understanding and analyzing trends or patterns of a population (for a description of population models, see Caswell 2001). Microsimulation models can also be seen as a form of population projection model (Imhoff and Post 1998). Microsimulation methodologies may be used to create small area population microdata by combining datasets and then using the results for geographical analysis (for a description of the method see Ballas, Rossiter, Bethan, Clarke, and Dorling 2005). The microsimulation method has been used in the past by economists with successful results to generate data for individuals and then check the effects of public policies at the smaller aggregation level (for use of microsimulation in economics, see Bourguignon and Spadaro 2006). Nevertheless, microsimulation models developed by economists lack the aspect of geographical space which is essential in cases where the spatial distribution of microunits is crucial. Taking into consideration the spatial characteristics of a population of entities (individuals, households or businesses) may reveal patterns and trends that may lead to a more comprehensive understanding of inequalities or other structures and help towards providing more equal public policies. Spatial microsimulation can be defined as a methodology for creating large population microdata which may be later used for the analysis of the impact of policy at the microlevel (Ballas and Clarke 2003) such as census output areas (the smallest geographical unit for which census data are available) or a post-code neighborhood. In the last two decades there have been some cases of using this method along with geographical data to produce spatial microdatasets for spatial microsimulation models (Birkin, Clarke, and Openshaw 1995a; Williamson, Birkin, and Rees 1998; Ballas and Clarke 2001). The microdata used in various geographical analysis may contain detailed information about individuals or households or business units, which have geographical reference and may be used for the geographical analysis of various scientific fields (for more on geographical microdata, see Birkin *et al.* 1995a; Dale, Fieldhouse, and Holdsworth 2000; Bartelsman and Doms 2000). Spatial microsimulation has been used in various spatially-aware research fields such as: understanding spatial inequalities of housing (Hooimeijer and Oskamp 1996; Deng, Wu, and Wang 2010; Cervero 1996), geographical analysis of education (Kavroudakis, Ballas, and Birkin 2013; Kavroudakis *et al.* 2012; Kavroudakis 2009), crime (Rephann and Öhman 1999; Bowers and Hirschfield 1999; Bosse and Gerritsen 2008), health (Ballas, Clarke, Dorling, Rigby, and Wheeler 2006; Caldwell 1996) and various economic topics such as pension provision (Ballas *et al.* 2006; Caldwell 1996) and taxes (Sutherland 2007; Creedy 2002; Sutherland 2000; Rudas, Szivós, and Tóth 1998; Redmond, Sutherland, and Wilson 1998; Bekkering 1995; Lewis, Michel, and Institute 1990; Lietmeyer and Dickhoven 1986). Spatial microsimulation is a powerful population modeling approach which may be used in policy making for the understanding of the spatial characteristics of populations. The production of spatially microsimulated microdata is a product of a complex methodology which requires the combination of *individual based dataset*s as well as macro-scale datasets (such as the *census of population*). This combination of data requires good understanding of fitting algorithms and combinatorial optimization approaches first developed and applied in a population geography context by Williamson *et al.* (1998). Recent reviews of relevant work and the state of the art include the work of Birkin and Clarke (2011), Tanton and Edwards (2013), Lovelace and Ballas (2013) and Hermes and Poulsen (2012).

2. The need for small area population microdata

The combination of geographical macrodata (census) and individual based data produces a spatially aware microdataset for further analysis. The **sms** package (Kavroudakis 2015) facilitates the combination of the datasets in a number of stages such as: production of small area microdata, the visualization of the process results and finally the presentation of the intermediate stages of the fitting process. Macrodata, such as the census of population, offer limited numbers of variables about the population of a geographical area. On the other hand, microdata or individual based data (IBD) such as the British Household Panel Survey (BHPS; BHPS 2007; IISER 2006) and The European Union Statistics on Income and Living

Conditions (EU-SILC; Whelan and Maître 2007; Figari, Levy, and Sutherland 2006) consist of many variables but they are typically only available at a coarse geographical level such as regions at best (Frees 2004). Individual based data are suitable for research because of the great number of variables but inappropriate to conduct geographical analysis (Hsiao 2003) mainly because of the lack of geographical reference. The combination of rich individual based data with census data produce a spatial microdataset which is ideal for geographical analysis and population simulation. The construction of such data is necessary as it provides researchers with geocoded datasets with individual based information. Spatial microsimulation is a methodology which uses such microdata to perform geographical analysis and population simulations (Kavroudakis *et al.* 2012). The combination of datasets is computationally intensive, requiring good knowledge of the simulated objects and the simulation platform. There is a necessity for a generic mechanism that should prepare microdata without specialized knowledge of the field for the end-user.

3. The sms package

The R environment for statistical computing and graphics (R Core Team 2015) is nowadays among the most common scientific programming tools used by a number of scientists around the world. The R statistical language is a modern statistical programming language and a great number of scientific methods are provided in R but there has so far not been any spatial microsimulation package. The software environment R includes many basic statistical functions and offers a mechanism of loading a great number of extra packages for more specialized tasks. There are some R packages for geographical analysis such as **spatstat** (Baddeley and Turner 2005), as well as **splm** (Millo and Piras 2012) and the **UScensus2000** "suite of packages" (Almquist 2010). Nevertheless, there is a relative paucity of R packages dealing with spatial microsimulation methods. The package presented in this work facilitates the development of individual based microdata and provides a structural skeleton for examining the results and intermediate states of a spatial microsimulation data-fitting process. In some geographical analysis cases, the number of areas and the population size require an efficient approach in simulation and combination of datasets. In order to increase the speed of the data-fitting process, the **sms** package uses a parallel processing approach which divides the main simulation process into smaller parts for parallel processing. Figure 1 shows the use of computational time of the CPU (central processing unit) between tasks execution in concurrent and parallel processing strategies. The concurrent approach uses the processing time either for task 1 or for task 2 but on the other hand, parallelism uses multiple dedicated cores (1 core for each task) to achieve a better performance (Birkin, Clarke, and George 1995b; Azencott 1992). The **sms** package automatically identifies the number (n) of available CPU cores in the user's system and uses $n/2$ parallel threads for the preparation of microdata. For example if a system has 6 CPU cores, the package will use $6/2 = 3$ parallel threads that will process the workload and handle one geographical area after the other (Figure 2). The **sms** package has two distinct advantages which make it a useful tool for researchers focusing on spatial microsimulation analysis and small area population estimates. These are: scalability and generalization. The **sms** package has the ability to handle data of various sizes and to conduct large scale simulations almost as efficiently as small scale simulations. This is possible mainly because of the parallel processing approach which makes use of available CPU cores of the system. The efficiency of the results depends heavily on the quality of input data. For

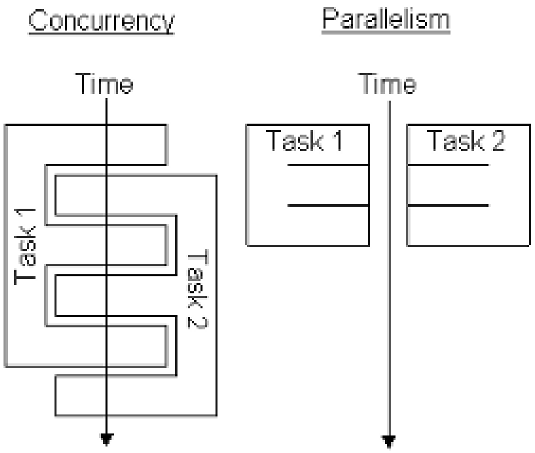

Figure 1: Concurrent and parallel strategies in computationally intensive calculations.

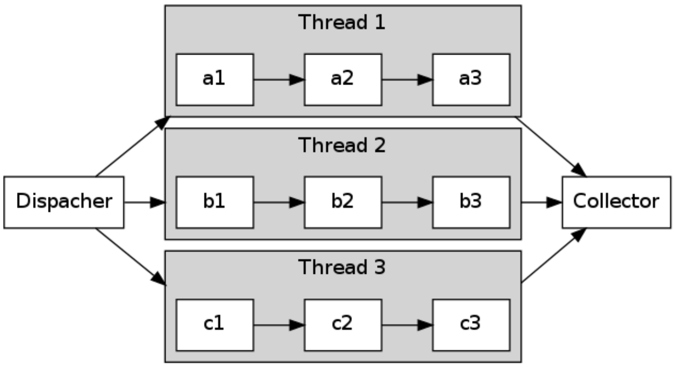

Figure 2: The representation of the parallel processing approach of the **sms** package for simulating multiple geographical areas at the same time.

more details on scalability issues in software engineering, see Smith and Williams (2002a,b); Laitinen, Fayad, and Ward (2000). Additionally, the generalized character of the **sms** package refers to the data-agnostic abilities of the package.

The generalized structure of the code enables the usability of parts of the package with other basic R methods such as `plot` because the **sms** package objects extend the well known data structures of R such as: `vectors`, `lists` and `data.frames` (for more details on data abstraction and generalization, see Gannon, McMullin, and Hamlet 1981; Cardelli and Wegner 1985; Liskov and Guttag 1986). Also, the '`microsimulation`' class of the package includes a number of `data.frames` which are accessible to other generic functions such as: `summary` or `str`. The **sms** package is open source and publicly available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=sms` and can be installed in two ways: either from a local file or directly from a CRAN server. If there is a working internet connection the following command downloads, installs and loads the **sms** package:

```
R> install.packages("sms", dependencies = TRUE)
R> library("sms")
```

The user is now ready to combine data and prepare microdata with the use of the **sms** package methods. The data fitting process for the preparation of microdata requires combinatorial optimization approaches for the selection of the more accurate group of individual data records for each geographical area. This process requires the use of optimization algorithms for the estimation of the accuracy and the selection of the optimal combination of data records. More specifically, in the context of this work, the optimization algorithms are used to select the optimal group of individual records that represent more accurately a geographical area. The main algorithms used for this data-fitting process are simulated annealing (SA) and hill climbing (HC). Simulated annealing is more advanced and returns relatively more accurate representations of populations for small areas. The hill climbing algorithm is a greedy heuristic algorithm for finding the best group of individual records. Its main disadvantage is that it can be "trapped" in local optimum solutions. On the other hand simulated annealing, as a more advanced heuristic algorithm, advances faster towards better solutions (solutions with less error) than hill climbing as it initially accepts less accurate solutions (combination of individuals that increased the current total absolute error). This initial tolerance of SA may lead to better data combinations later in the optimization process. This flexibility enables the avoidance of local optimum solutions and this is one of its key advantages compared to hill climbing making it ideal for this type of combinatorial optimization approach. Williamson describes the effectiveness and the advantages of simulated annealing in more detail (Voas and Williamson 2000; Williamson *et al.* 1998). This algorithm is an analogy of the annealing of solid material in physics and it was initially conceptualized in a paper published by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and describes the simulation of the cooling of materials in heat bath, which is also known as "annealing". Simulated annealing was first used in the small area data fitting context by Williamson *et al.* (1998) and there have been several applications that applied or built upon that work (Kavroudakis *et al.* 2012). The **sms** package includes both simulated annealing and hill climbing algorithms in order to provide the user with a choice of optimization techniques for result comparison purposes.

4. Case study

Due to the complexity involved in understanding the microdata preparation process, an illustrated case study will be used for the presentation of **sms** package functions and discussion of the results. This section will present the preparation of microdata for a geographical area in Greece (Lesvos island) and will illustrate the use of the **sms** package for small area data preparation. The example data used are simplified artificial data which resemble the structure of actual census and the individual based dataset available for Greece from Eurostat (Eurostat 2012). The example data are also included in the **sms** package for teaching and demonstration purposes. The island of Lesvos consists of 13 constituencies. Table 1 shows the census data of the geographical area that will be used as input data for the microsimulation process. The total population of the island is 900 individuals (Table 1), which includes 459 individuals with higher education (he) degrees and 428 female individuals. The census dataset in Table 1 reflects the counts of individuals in a specific point in time for each geographical area of the study area. Usually census tables include a limited number of variables (ESRC 2008; Rees, Martin, and Williamson 2002) and this is why we need to combine Table 1 with an individual based dataset in order to produce a rich individual based dataset which reflects the actual aggregate census counts. The following sample dataset is included in the **sms** package and

areaid	population	he	females
8301	56	46	42
8302	73	42	15
8303	58	12	10
8304	78	43	21
8305	73	17	60
8306	77	15	11
8307	66	37	20
8308	78	41	42
8309	77	56	10
8310	78	55	26
8311	58	19	60
8312	68	20	60
8313	60	56	51

Table 1: The demonstration data for Lesvos island, Greece, which are a simplification of the census of population.

pid	he	female	agemature	car_owner	house_owner	working	annualIncome
6001	0	1	1	1	0	1	16567
6002	1	1	1	1	1	1	2458
6003	0	1	0	0	1	0	9437
6004	1	0	0	1	0	0	22130
6005	1	0	0	0	1	1	3936
6006	1	0	0	1	0	1	16695

Table 2: The first 6 rows from the example individual based dataset for Greece which is a simplification of the available individual based data for Greece.

resembles the actual census and survey data for Greece. After loading the **sms** package we use the following two commands to load the census and survey data into the R work space:

```
R> data("census", package = "sms")
R> data("survey", package = "sms")
```

Table 1 shows the census information for the study area and Table 2 shows the first six rows from the available individual based dataset for Greece that consists of 200 individual records with a number of variables for each record (rows).

Now the `census` variable is a `data.frame` containing census data, as shown in Table 1. The individual based dataset can also be loaded in the same way and is also a `data.frame` in the form of Table 2, where each column represents a variable for individuals. The variables have been binary recoded in order to increase the efficiency of the calculations. This is, instead of one column named `sex` with two possible variable states: male/female, the column is a representation of a variable state `female` and the value is binary either 0 (no), or 1 (yes). This conversion may increase the number of columns, but on the other hand it ensures a consistent data format of binary values helpful during intense calculations of the microdata selection process. Also, it is necessary for the `census` dataset to have a column with the name population indicating the total number of individual records for each area. After loading the data, it is necessary to associate data variables between the two data.frames. For a generic approach, a complex data structure is needed. The **sms** package uses a "data lexicon" which is an R data.frame which provides information about the relationship of the variables of each data source. The following code shows the construction of the data lexicon for this case study:

```
R> in.lexicon <- createLexicon()
R> in.lexicon <- addDataAssociation(in.lexicon, c("he", "he"))
R> in.lexicon <- addDataAssociation(in.lexicon, c("females", "female"))
R> in.lexicon
```

```
           con_1   con_2
census_row    he females
survey_row    he  female
```

The lexicon is a data.frame which in this example holds 2 data connections. Each column of the lexicon represents a data association between census and survey data. The first connection is described in the first column of the data.frame under the name con_1 and indicates that census column with name he is associated with column he in the survey dataset. The data.lexicon holds the "association information" between datasets for the fitting process.

5. Fitting and evaluation

The fitting process produces a number of results which may be used for the analysis and interpretation of the population attributes. The results of this process can also be used for public policy analysis and hypothetical what-if scenarios. The results for a geographical area may deviate from the actual counts of the census variables. This deviation is related to the number of variables used as well as the fitting mechanism of the microsimulation process. The data-fitting is the process of fitting individual based data to the census description of an area which is an optimization process where groups of individual records are randomly selected from a database and are evaluated against the profile of a geographical area. This iterative process extracts and replaces individual based records until small area constraints are satisfied.

This optimization process evaluates a selection of individual records and calculates an error value during each selection. During this process the same individual record may be used more than once, which is a desirable behavior because the aim of the process is to accurately re-create the total attributes of an area. If the selection of a specific record improves the accuracy of the area's representation, then it may be selected multiple times. Consequently, the **sms** package aims to find the optimum selection of individual records for each geographical area that minimizes an error function. The **sms** package uses the total absolute error (TAE) formula to quantify the absolute deviation of the microdata from census data description of an area. Equation 1 shows the TAE formula which is the sum of the absolute deviations of the simulated values (s_i), from the census counts of the areas (c_i).

$$TAE_i = \sum |s_i - c_i| \tag{1}$$

```
R> mysms <- new("microsimulation", iterations = 90, census = census,
+    panel = survey, lexicon = in.lexicon)
R> myseed <- 1800
R> try01 <- run_parallel_HC(insms = mysms, inseed = myseed)
R> try02 <- run_parallel_SA(insms = mysms, inseed = myseed)
```

After loading the datasets and constructing the data lexicon, the user is ready to produce small area population microdata for each of the 13 constituencies of Lesvos island. As can be seen in the following code, the `run_parallel_SA` function returns the best possible microdataset for all areas in `census` after 90 iterations by using individual records from the `survey` dataset. This function uses the simulated annealing algorithm for the selection of the best possible combination of individual records minimizing the TAE. Simulated annealing is a "tolerant" heuristic algorithm that initially accepts combinations which do not improve the error value, while gradually this tolerance decreases because this may lead to better combinations as it may help escape local optimum solutions.

There is also an additional method available in the **sms** package which uses the hill climbing algorithm for the construction of the microdataset. This method is `run_parallel_HC` which uses the hill climbing heuristic algorithm which is a "greedy" heuristic algorithm that accepts only combinations that improve the error value. Figure 3 depicts the comparison of the two algorithms for the development of a small area microdata population. The bottom part of the figure depicts the fitting process with the use of the `run_parallel_SA` method (simulated annealing algorithm) and the top part shows the fitting process of the same data with the use of the `run_parallel_HC` method (hill climbing algorithm).

In both graphs of Figure 3 the horizontal axes indicate the number of iterations of the fitting process and the vertical axes show the total absolute error of each selection. Filled dots represent selections which are saved as "best current selection" and round unfilled circles indicate selections which are not saved as they have unacceptable TAE which is greater than previous TAE values. During the first iterations the simulated annealing algorithm is tolerant to combinations that increase the TAE which eventually leads to a solution with smaller overall TAE than hill climbing (top graph) which is a "greedy" algorithm accepting only combinations decreasing the TAE.

Both `run_parallel_HC` and `run_parallel_SA` methods return a '`microsimulation`' object containing a number of other objects and results. The `try02` object contains results for the 13 areas of Lesvos island. The '`microsimulation`' object holds 5 slots of information regarding the data fitting process of each area. The following code shows the number of elements in the results of the fourth area and explores each element:

```
R> oneAreaResult <- try02@results[[4]]
R> length(oneAreaResult)

[1] 5

R> names(oneAreaResult)

[1] "areaid"       "selection"    "tae"          "tries"
[5] "error_states"
```

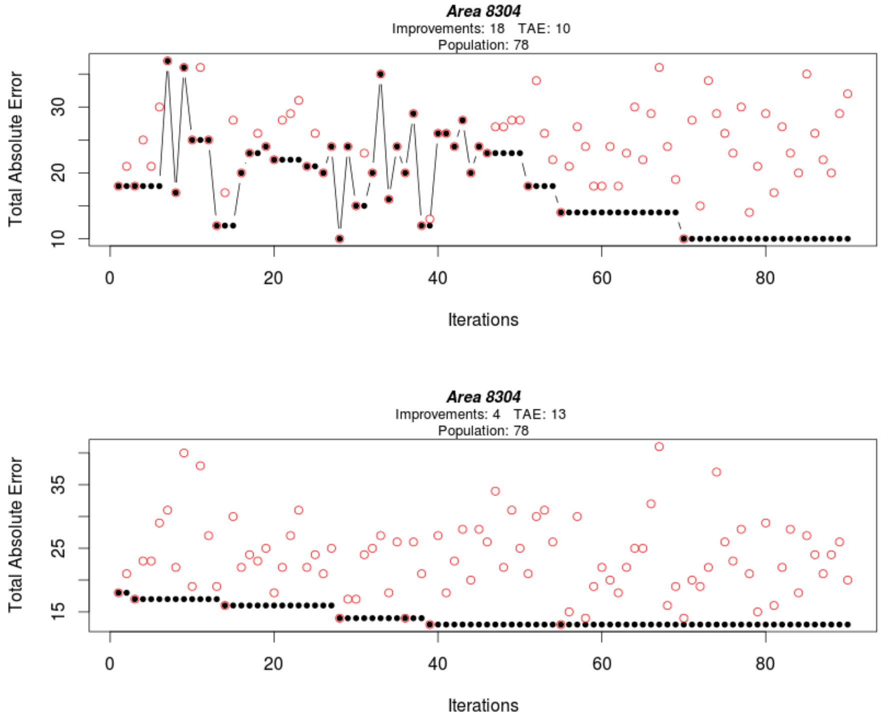

Figure 3: Comparison of hill climbing (top) and simulated annealing (bottom) algorithms for the construction of population microdata of a single geographical area.

The 5 list elements contain the following information about the construction of population microdata:

- `areaid`: This is the unique identification of the area which identifies the area in the census database.
- `selection`: This object holds the best combination of individual records satisfying the census constraints. It is also called "synthetic micro-population" and may be used for further geographical analysis.
- `tae`: This object holds the value of the total absolute error of the results. This is a measure of the absolute deviation of the results from the census data constraints of the simulated area. The aim of the fitting process is to minimize this value. The smaller this value, the greater the overall accuracy of the results.
- `tries`: This object holds all intermediate *TAE* values of the fitting process. Those values are useful for understanding the fitting process as they represent the *TAE* of the intermediate selections illustrating the algorithm's progress. The number of *TAE* values is the same as the number of iterations of the fitting process. These values can be plotted as they are enclosed in an R `vector` class.
- `error_states`: This slot holds the *TAE* values of the accepted selections of the selection process. This vector represents the progress of the fitting algorithm while advancing

towards the global optimum solution. The values of this vector represent the best current *TAE*.

Each element of the results can be exported individually for analysis and interpretation as follows:

```
R> oneAreaResult$areaid

[1] 8304

R> head(oneAreaResult$selection)

     pid he female agemature car_owner house_owner working
15  6152  1      1         0         0           1       0
56  6042  0      0         1         1           0       1
193 6046  1      0         0         1           1       0
199 6031  0      0         1         0           1       1
118 6197  0      0         1         1           0       1
168 6029  0      1         1         1           0       0
    annualIncome
15         11947
56         19157
193        24515
199        10229
118        16944
168        22979

R> oneAreaResult$tae

[1] 10

R> oneAreaResult$tries

 [1] 18 21 18 25 21 30 37 17 36 25 36 25 12 17 28 20 23 26 24 22 28 29
[23] 31 21 26 20 24 10 24 15 23 20 35 16 24 20 29 12 13 26 26 24 28 20
[45] 24 23 27 27 28 28 18 34 26 22 14 21 27 24 18 18 24 18 23 30 22 29
[67] 36 24 19 10 28 15 34 29 26 23 30 14 21 29 17 27 23 20 35 26 22 20
[89] 29 32

R> oneAreaResult$error_states

 [1] 18 18 18 18 18 18 37 17 36 25 25 25 12 12 12 20 23 23 24 22 22 22
[23] 22 21 21 20 24 10 24 15 15 20 35 16 24 20 29 12 12 26 26 24 28 20
[45] 24 23 23 23 23 23 18 18 18 18 14 14 14 14 14 14 14 14 14 14 14 14
[67] 14 14 14 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
[89] 10 10
```

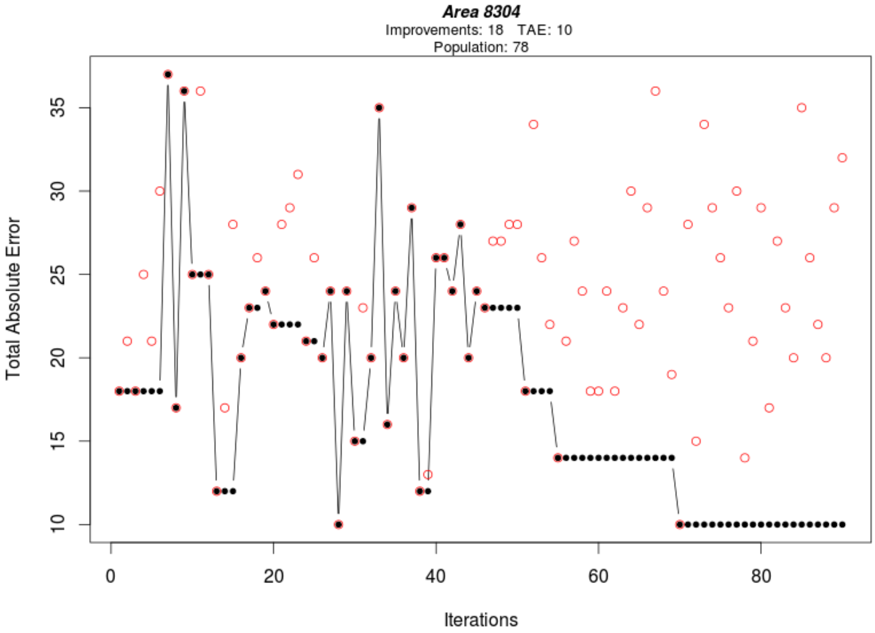

Figure 4: The progress of the total absolute error (*TAE*) during the fitting process with the use of the simulated annealing algorithm.

Figure 4 depicts the progress of the *TAE* values during small area population fitting. Using the function `plotTries`, the user can visualize the *TAE* states during the fitting process of a single area. Horizontal axes represent the iterations of the fitting process and vertical axes represent the *TAE* values.

```
R> plotTries(insms = try02, number = 4)
```

The simulated annealing algorithm initially creates a random population of individuals, which has `TAE = 10`. This combination is selected (circle line around a filled dot) as there is no other better solution yet in the process. In later stages, the algorithm tries to find the best selection by evaluating other combinations of individual records from the individual based dataset. The empty round circles represent unused selections of individual records with greater *TAE*. The filled dots represent selected combinations which are the best solution until that iteration. During the progress of the fitting process (from left to right) the algorithm keeps in memory only the best selection until that iteration and tries to find other possible combinations with smaller total absolute error. In this example the microdata for geographical area 8304 have been prepared after 90 iterations with final `TAE = 10`. This case study illustrated the use of the **sms** package for the development of small area microdata. The microdataset (`oneAreaResult$selection`) is used for further analysis and interpretation. These results can be aggregated and analyzed geographically as they contain individual records which have been placed in a virtual geographical space to produce a virtual representation of Lesvos island, Greece. This combination of individual based data and census datasets is a valuable source of information for scientific analysis as it is a unique source of information with geographical attributes. This type of microdata offers the opportunity to analyze a geographical area in the smallest possible scale by examining the characteristics of the individual records of each population. During the last years, the use of microdata has become a popular approach among organizations and public entities aiming to examine local effects of potential public policies and evaluate the potential effects of local reforms.

6. Geographical analysis of microdata

The microdata produced using the **sms** package is a representation of the Lesvos island population. The accuracy of this representation depends on the number of algorithm iterations as well as the number of common variables between the census of population and the individual based dataset. The greater the number of iterations during the fitting process, the more the probabilities of selecting suitable individual based records for the accurate representation of the area's population. Additionally, the more common variables between the two datasets, the more accurate is the final population representation as individual records have increased probabilities to fulfill census data constraints. The following sequence of maps (Figures 5 to 10) depict the results of the data fitting process. For depiction simplicity, percentage counts have been used where possible. One of the most important variables to map is the simulated mean income, which is not typically available from public sources and which can be extremely useful for the analysis of the geographical implications of government policies and for estimating poverty and wealth at the local level (Ballas, Clarke, Dorling, and Rossiter 2007; Campbell and Ballas 2013; Miranti, McNamara, Tanton, and Harding 2011). This information was not available before data-fitting at this geographical level. These maps may be used for comparison among areas and depict various population characteristics. Such characteristics include house and car ownership, sex, age group and working status. This group of variables was not available for this geographical scale. This synthetic population may be used for relevant geographical analysis in a finer possible scale. One example could be the analysis of working individuals in the localities of the island as this can be potentially useful for the examination of the island's workforce and the provision of new jobs or targeted unemployment campaigns.

A different approach of presenting the result of the fitting process is the preparation of graphs showing relationships between population attributes by locality. Figure 11 depicts the relationship between percentage of employed population and percentage of population with house ownership. This scatterplot shows a positive relationship between the two variables which indicates that according to the results of the fitting process employment turnover is somehow adequate for buying a house in Lesvos island. In Figure 12 we can see a positive relationship between percentage of population in employment and percentage of population having at least one car. The blue line is the line of "best fit" from a linear regression between the two variables and the value of `coefficient` is depicted in the subtitle of the scatterplots showing the type and extend of relationship between the two variables. According to the results, employment opportunities are normally distributed across all 13 localities of the island with some localities showing a relatively smaller percentage of employment. These localities can be targeted for new employment projects and possible government investment policies. Policy makers can use these results to analyze the characteristics of the island's population and understand the local characteristics of each area. The results of the data fitting process represent individuals and can be analyzed in various aggregation levels (city, locality, constituency, region) making it ideal for policy making simulations and examination of local effects of government policies.

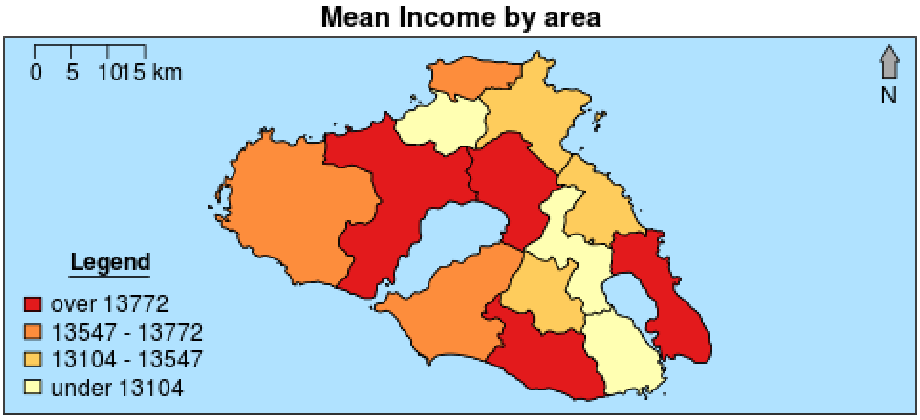

Figure 5: Results of data fitting, showing mean income by constituency in Lesvos island.

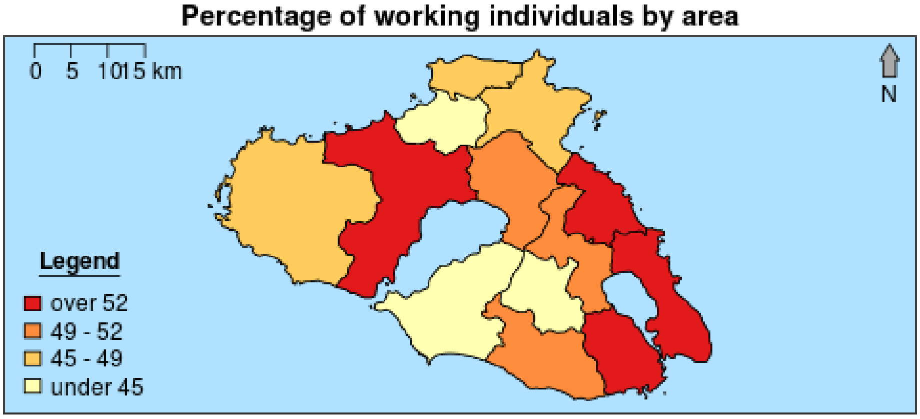

Figure 6: Results of data fitting, showing percentage of working individuals by constituency in Lesvos island.

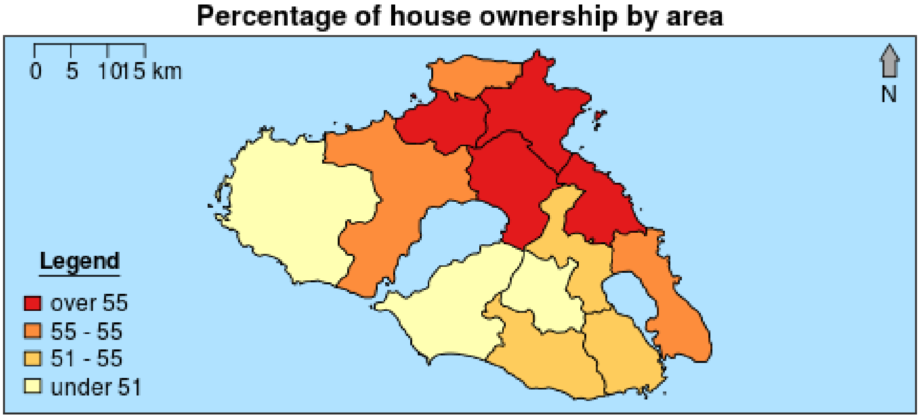

Figure 7: Results of data fitting, showing percentage of house ownership by constituency in Lesvos island.

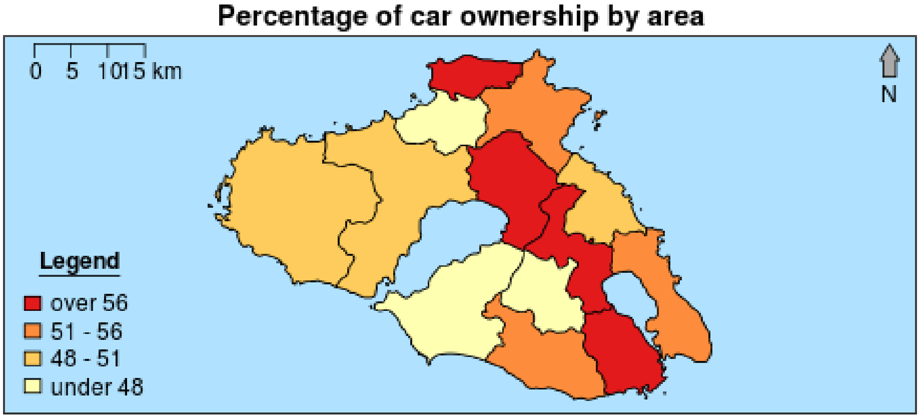

Figure 8: Results of data fitting, showing percentage of car ownership by constituency in Lesvos island.

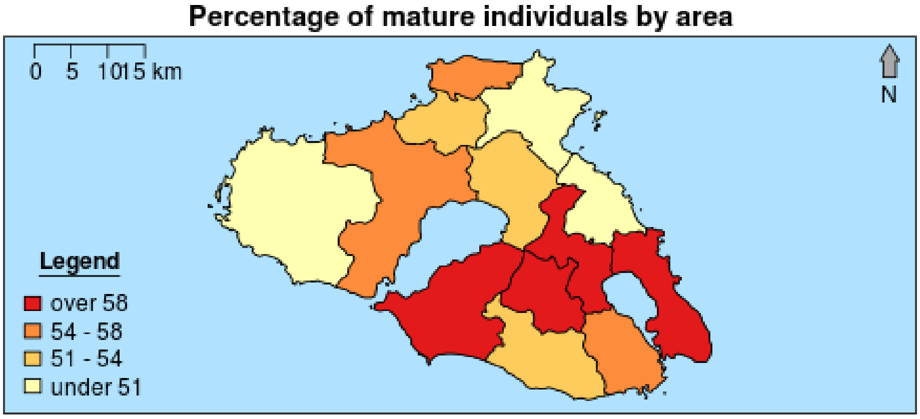

Figure 9: Results of data fitting, showing percentage of mature individuals by constituency in Lesvos island.

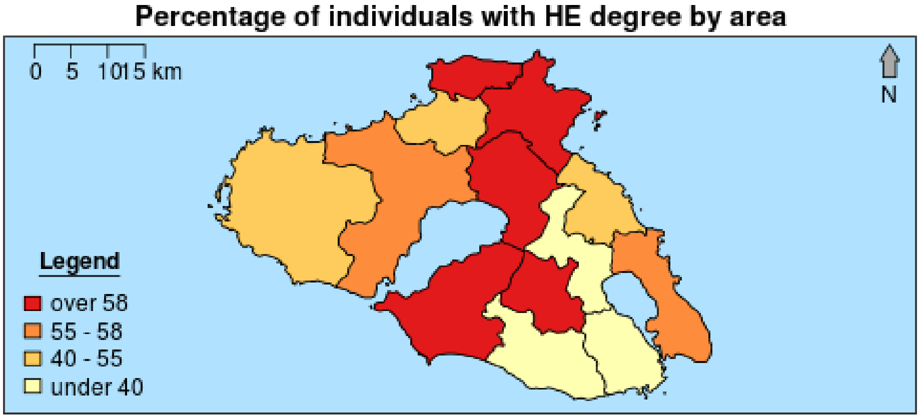

Figure 10: Results of data fitting, showing percentage of individuals holding a higher education degree by constituency in Lesvos island.

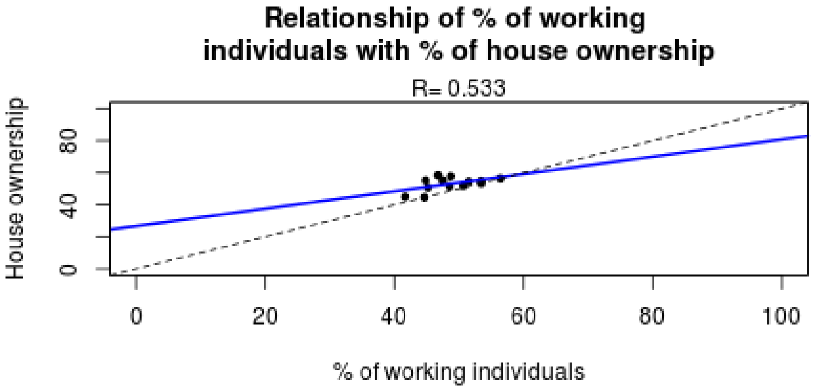

Figure 11: Relationship between percentage of working population and percentage of population with house ownership at the 13 localities of Lesvos island.

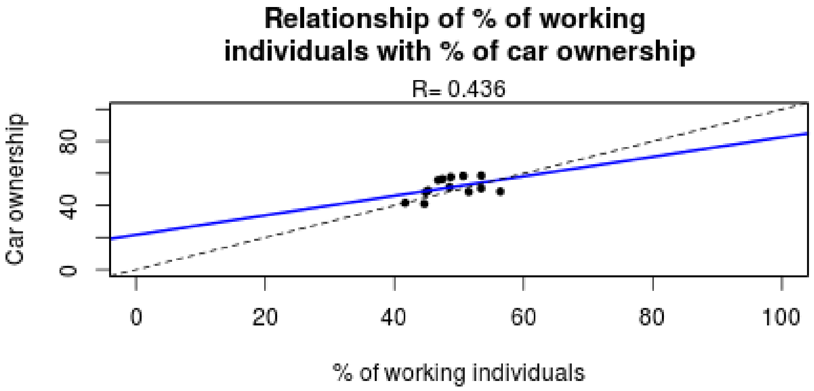

Figure 12: Relationship between percentage of working population and percentage of population owning at least one car at the 13 localities of Lesvos island.

7. Benchmarking the sms package

In order to illustrate the differences between the parallel and serial versions of the **sms** package methods, a number of comparisons have been made. In those comparisons the same data have been used by both methods and the elapsed time until the preparation of results was compared. In almost 90% of the cases a parallel approach was faster than a serial approach, making it ideal for large scale simulations. This section of the paper presents a set of evaluations of the **sms** package. The evaluation process includes a number of *runs* with various configuration sets. This benchmarking approach uses differently sized census data of: 10, 20, 30, 40 and 50 geographical areas. It also uses differently sized individual based survey data of: 100, 200, 300, 400, and 500 individual records. Additionally, it uses many iteration groups (100, 200, 300, 400, 500) before returning the results. This benchmarking mechanism prepares 125 different combinations of the above three factors (census data size, individual based dataset size, number of iterations) and evaluates the final *TAE* of each combination. The specifications of the testing machine are: Intel Core i7-3635QM CPU 8*2.40GHz, 64bit, 8GiB RAM. Figures 13 to 18 depict the benchmarking results in scatterplots with a regres-

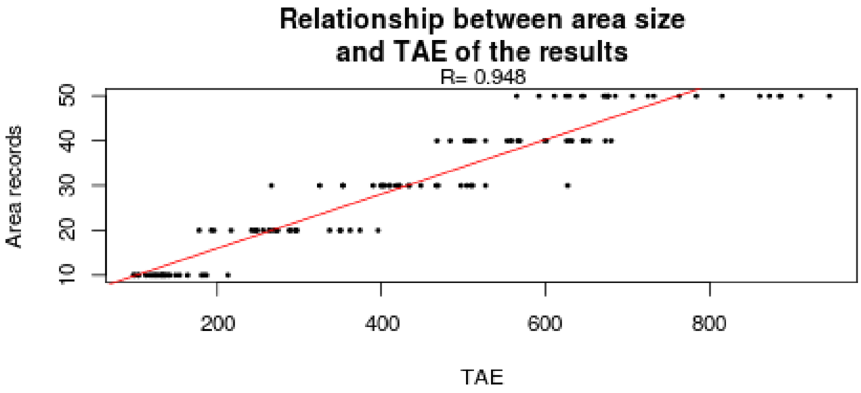

Figure 13: Relationship between census size and *TAE*.

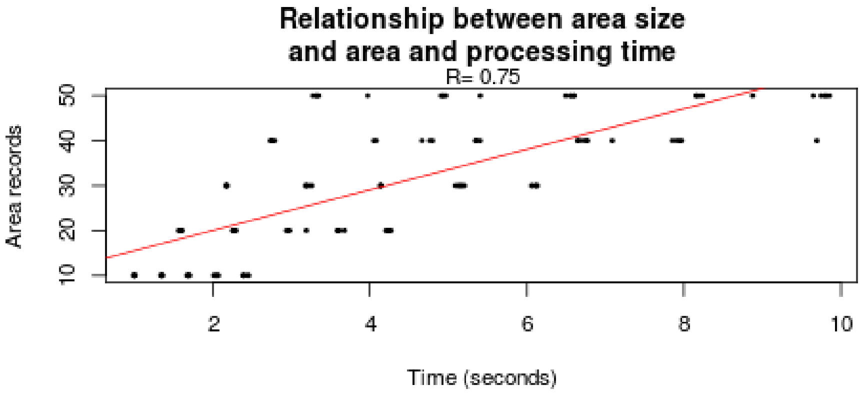

Figure 14: Relationship between census size and elapsed time.

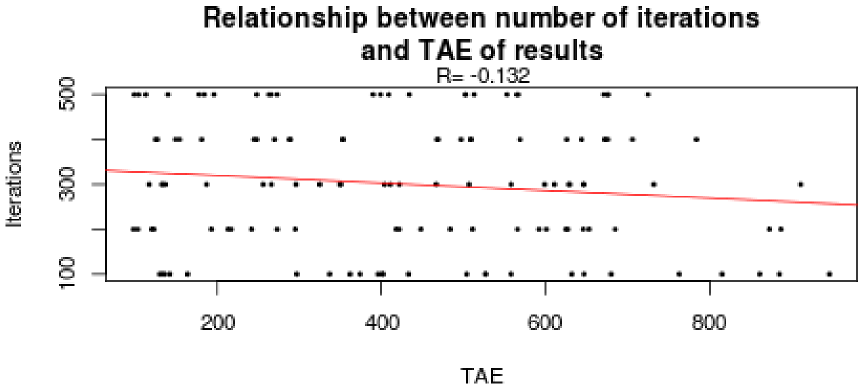

Figure 15: Relationship between number of iterations and *TAE*.

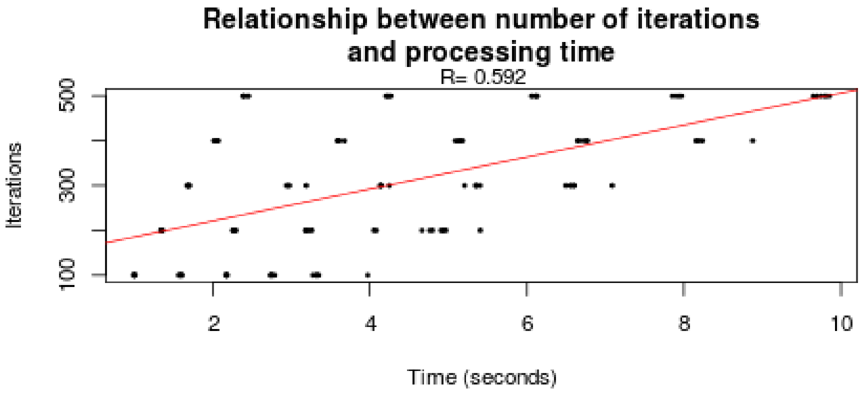

Figure 16: Relationship between number of iterations and elapsed time.

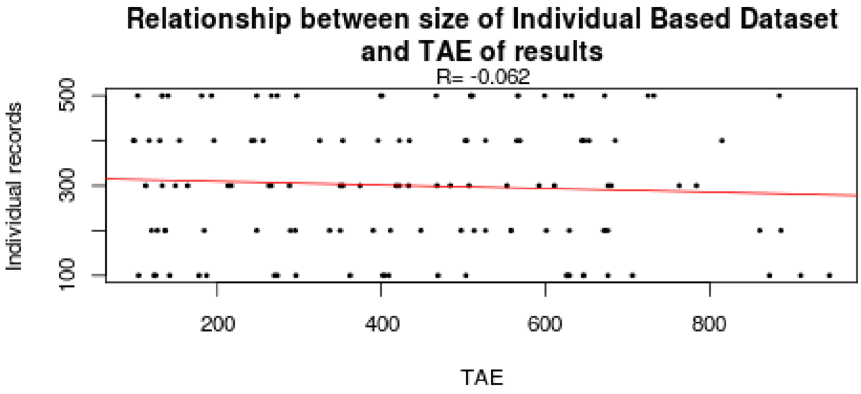

Figure 17: Relationship between individual based data size and *TAE*.

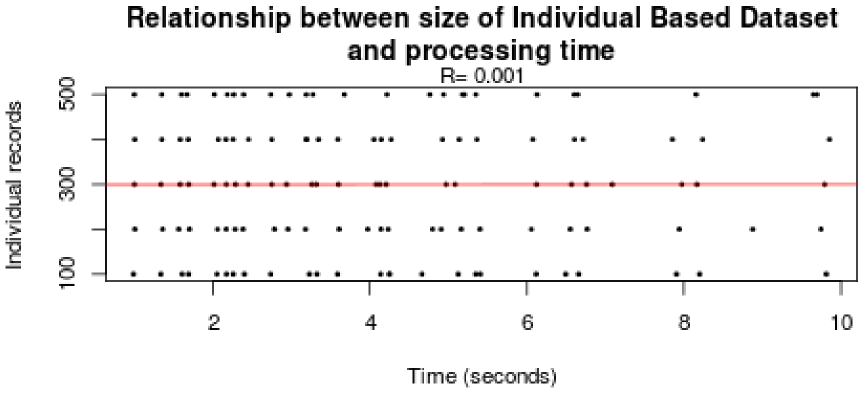

Figure 18: Relationship between individual based data size and elapsed time.

sion line (red line) and the correlation coefficient value in the subtitle, indicating the relative strength and type of correlation between the two variables. The results indicate a positive relationship between the size of the geographical area and the TAE as well as the relationship between size of the geographical area and elapsed time of the data fitting process (Figures 13 and 14). This is expected as there is a direct relationship between the size of a geographical area and the required computation time of simulating it. Also, algorithmically TAE is positively related to the census size of the area. There is also a negative relationship between the number of algorithm iterations and the value of TAE (Figures 15 which is expected because the more the iteration tries of the algorithm the more the opportunities for randomly finding a better state in the optimization process which will have a smaller TAE value. Finally as can be seen in 16, as the number of iterations increase, elapsed time is also increasing which is expected considering the computational effort required for evaluating an increasing number of possible solutions.

Finally, Figure 17 show that as individual based dataset size increases, the TAE is slightly decreasing which is expected as the size of the individual based dataset is directly associated with the number of individuals available for selection from the individual based dataset. The more the possible available combinations the more the possibilities for constructing a more acceptable solution. Figure 18 shows a somehow indifferent relationship between individual based dataset and the the total processing time. The more individual records are available, the higher are the chances that the small area constraints will be fulfilled, which means that TAE will drop. These benchmarks helped evaluate the **sms** package and quantify the relationship between the factors affecting computation time and TAE of the data fitting process. It is necessary to note that the complexity of the population characteristics in a geographical area as well as the representation of the individual based data are two relevant factors affecting scalability of the data fitting process.

8. Conclusion

The **sms** package presented in this paper prepares microdata for small geographical areas such as constituencies, census output areas or post code sectors for further geographical analysis. Microdata have many uses in contemporary research such as small area estimations and population projections as well as pubic policy analysis and simulation. This R package is unique as it uses the parallel processing abilities of the R language and can simulate relatively large datasets. The spatial microsimulation process is a methodology which uses small area microdata to simulate local effects of public policies in finer geographical scale such as household level or individual level. The use of this type of scientific tools will be increasingly important in the future for examining local effects of policies. Another potential use of this package is the preparation of microdata for the examination of the local effects of what-if scenarios and how population attributes (income, car ownership, housing, education, etc.) may be affected.

References

Almquist Z (2010). "US Census Spatial and Demographic Data in R: The **UScensus2000** Suite of Packages." *Journal of Statistical Software*, **37**(6), 1–31. doi:10.18637/jss.v037.i06.

Azencott R (1992). *Simulated Annealing: Parallelization Techniques*. John Wiley & Sons, New York.

Baddeley A, Turner R (2005). "**spatstat**: An R Package for Analyzing Spatial Point Patterns." *Journal of Statistical Software*, **12**(6), 1–42. doi:10.18637/jss.v012.i06.

Ballas D, Clarke G (2001). "Modelling the Local Impacts of National Social Policies: A Spatial Microsimulation Approach." *Environment and Planning C: Government and Policy*, **19**(4), 587–606. doi:10.1068/c0003.

Ballas D, Clarke G (2003). "Microsimulation and Regional Science: 30 Years of Spatial Microsimulation of Populations." In *50th Annual North American Meeting of the Regional Science Association International*. Philadelphia, USA.

Ballas D, Clarke G, Dorling D, Rigby J, Wheeler B (2006). "Using Geographical Information Systems and Spatial Microsimulation for the Analysis of Health Inequalities." *Health Informatics Journal*, **12**(1), 65–79. doi:10.1177/1460458206061217.

Ballas D, Clarke G, Dorling D, Rossiter D (2007). "Using **SimBritain** to Model the Geographical Impact of National Government Policies." *Geographical Analysis*, **39**(1), 44–77. doi:10.1111/j.1538-4632.2006.00695.x.

Ballas D, Rossiter D, Bethan T, Clarke G, Dorling D (2005). *Geography Matters: Simulating the Local Impacts of National Social Policies*. Contemporary Research Issues. Joseph Rowntree Foundation, York.

Bartelsman E, Doms M (2000). "Understanding Productivity: Lessons from Longitudinal Microdata." *Journal of Economic Literature*, **38**(3), 569–594. doi:10.1257/jel.38.3.569.

Bekkering J (1995). *A Microsimulation Model to Analyze Income Tax Individualization*. Tilburg University Press, Tilburg.

BHPS (2007). "ISER-Survey British Household Panel Survey." URL https://www.iser.essex.ac.uk/bhps/.

Birkin M, Clarke G, Openshaw S (1995a). "Using Microsimulation Methods to Synthesise Census Microdata." In *Census Users' Handbook*, pp. 363–387. GeoInformation, Cambridge.

Birkin M, Clarke M (2011). "Spatial Microsimulation Models: A Review and a Glimpse Into the Future." In *Population Dynamics and Projection Methods*, pp. 193–208. Springer-Verlag.

Birkin M, Clarke M, George F (1995b). "The Use of Parallel Computers to Solve Nonlinear Spatial Optimisation Problems: An Application to Network Planning." *Environment and Planning A*, **27**(7), 1049–1068. doi:10.1068/a271049.

Bosse T, Gerritsen C (2008). "Agent-Based Simulation of the Spatial Dynamics of Crime: On the Interplay Between Criminal Hot Spots and Reputation." In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems – Volume 2*, pp. 1129–1136. International Foundation for Autonomous Agents and Multiagent Systems.

Bourguignon F, Spadaro A (2006). "Microsimulation as a Tool for Evaluating Redistribution Policies." *The Journal of Economic Inequality*, **4**(1), 77–106. doi:10.1007/s10888-005-9012-6.

Bowers K, Hirschfield A (1999). "Exploring Links Between Crime and Disadvantage in North-West England: An Analysis Using Geographical Information Systems." *International Journal of Geographical Information Science*, **13**(2), 159–184. doi:10.1080/136588199241409.

Caldwell S (1996). "Health, Wealth, Pensions and Life Paths: The **CORSIM** Dynamic Microsimulation Model." In *Microsimulation and Public Policy: Selected Papers from the IARIW Special Conference on Microsimulation and Public Policy, Canberra, 5–9 December, 1993*, volume 232, pp. 505–522. Emerald Group.

Campbell M, Ballas D (2013). "A Spatial Microsimulation Approach to Economic Policy Analysis in Scotland." *Regional Science Policy & Practice*, **5**(3), 263–288. doi:10.1111/rsp3.12009.

Cardelli L, Wegner P (1985). "On Understanding Types, Data Abstraction, and Polymorphism." *ACM Computing Surveys*, **17**(4), 471–523. doi:10.1145/6041.6042.

Caswell H (2001). *Matrix Population Models*. John Wiley & Sons.

Cervero R (1996). "Mixed Land-Uses and Commuting: Evidence from the American Housing Survey." *Transportation Research Part A*, **30**(5), 361–377. doi:10.1016/0965-8564(95)00033-x.

Creedy J (2002). *Microsimulation Modelling of Taxation and the Labour Market: The Melbourne Institute Tax and Transfer Simulation*. Edward Elgar, Cheltenham.

Dale A, Fieldhouse E, Holdsworth C (2000). *Analyzing Census Microdata*. Arnold, London.

Deng C, Wu C, Wang L (2010). "Improving the Housing-Unit Method for Small-Area Population Estimation Using Remote-Sensing and GIS Information." *International Journal of Remote Sensing*, **31**(21), 5673–5688. doi:10.1080/01431161.2010.496806.

ESRC (2008). "Census of Population Programme." URL http://census.ac.uk/.

Eurostat (2012). "Eurostat Homepage." URL http://ec.europa.eu/eurostat/.

Figari F, Levy H, Sutherland H (2006). "Using the EU-SILC for Policy Simulation: Prospects, Some Limitations and Some Suggestions." In *Comparative EU Statistics on Income and Living Conditions: Issues and Challenges – Proceedings of the EU-SILC Cconference, Helsinki*, pp. 6–8.

Frees E (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press.

Gannon J, McMullin P, Hamlet R (1981). "Data Abstraction, Implementation, Specification, and Testing." *ACM Transactions on Programming Languages and Systems*, **3**(3), 211–223. doi:10.1145/357139.357140.

Hermes K, Poulsen M (2012). "A Review of Current Methods to Generate Synthetic Spatial Microdata Using Reweighting and Future Directions." *Computers, Environment and Urban Systems*, **36**(4), 281–290. doi:10.1016/j.compenvurbsys.2012.03.005.

Hooimeijer P, Oskamp A (1996). "A Simulation Model of Residential Mobility and Housing Choice." *Journal of Housing and the Built Environment*, **11**(3), 313–336. doi:10.1007/bf02496594.

Hsiao C (2003). *Analysis of Panel Data*, volume 34. Cambridge University Press.

IISER (2006). "Quality Profile: British Household Panel Survey Version 2.0." URL https://www.iser.essex.ac.uk/files/bhps/quality-profiles/BHPS-QP-01-03-06-v2.pdf.

Imhoff E, Post W (1998). "Microsimulation Methods for Population Projection." *Population: An English Selection*, **10**(1), 97–138.

Kavroudakis D (2009). *Spatial Microsimulation for Researching Social and Spatial Inequalities of Educational Attainment*. PhD. thesis, University of Sheffield, United Kingdom.

Kavroudakis D (2015). ***sms**: Spatial Microsimulation*. R package version 2.3.1, URL http://CRAN.R-project.org/package=sms.

Kavroudakis D, Ballas D, Birkin M (2012). "Using Spatial Microsimulation to Model Social and Spatial Inequalities in Educational Attainment." *Applied Spatial Analysis and Policy*, **6**(1), 1–23. doi:10.1007/s12061-012-9075-2.

Kavroudakis D, Ballas D, Birkin M (2013). "**SimEducation**: A Dynamic Spatial Microsimulation Model for Understanding Educational Inequalities." In R Tanton, K Edwards (eds.), *Spatial Microsimulation: A Reference Guide for Users*, number 6 in Understanding Population Trends and Processes, pp. 209–222. Springer-Verlag, Netherlands.

Laitinen M, Fayad M, Ward R (2000). "The Problem with Scalability." *Communications of the ACM*, **43**(9), 115–118. doi:10.1145/348941.349012.

Lewis G, Michel R, Institute U (1990). *Microsimulation Techniques for Tax and Transfer Analysis*. Urban Institute Press; Distributed by University Press of America, Washington, DC.

Lietmeyer V, Dickhoven S (1986). "Microanalytic Tax Simulation Models in Europe: Development and Experience in the German Federal Ministry of Finance." In *Microanalytic Simulation Models to Support Social and Financial Policy*, pp. 139–152. Elsevier.

Liskov B, Guttag J (1986). *Abstraction and Specification in Program Development*. MIT Press.

Lovelace R, Ballas D (2013). "'Truncate, Replicate, Sample': A Method for Creating Integer Weights for Spatial Microsimulation." *Computers, Environment and Urban Systems*, **41**, 1–11. doi:10.1016/j.compenvurbsys.2013.03.004.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953). "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics*, **21**(6), 1087–1092. doi:10.1063/1.1699114.

Millo G, Piras G (2012). "**splm**: Spatial Panel Data Models in R." *Journal of Statistical Software*, **47**(1), 1–38. doi:10.18637/jss.v047.i01.

Miranti R, McNamara J, Tanton R, Harding A (2011). "Poverty at the Local Level: National and Small Area Poverty Estimates by Family Type for Australia in 2006." *Applied Spatial Analysis and Policy*, **4**(3), 145–171. doi:10.1007/s12061-010-9049-1.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Redmond G, Sutherland H, Wilson M (1998). *The Arithmetic of Tax and Social Security Reform: A User's Guide to Microsimulation Methods and Analysis*. Cambridge University Press, Cambridge.

Rees P, Martin D, Williamson P (2002). *The Census Data System*. John Wiley & Sons, Chichester.

Rephann T, Öhman M (1999). "Building a Microsimulation Model for Crime in Sweden: Issues and Applications." In *Seminarium Om Ekobrottsforskning*, pp. 12–15, 20–25, 30–35.

Rudas T, Szivós P, Tóth I (1998). "**TÁRSZIM**: Hungarian Tax – Benefit Microsimulation Model." In *Workshop on Microsimulation in the New Millennium: Challenges and Innovations*. Cambridge.

Smith C, Williams L (2002a). "Performance and Scalability of Distributed Software Architectures: An SPE Approach." *Parallel and Distributed Computing Practices*, **3**(4).

Smith C, Williams L (2002b). *Performance Solutions: A Practical Guide to Creating Responsive, Scalable Software*, volume 1. Addison-Wesley Boston, MA;.

Sutherland H (2000). "**EUROMOD**: An Integrated European Benefit-Tax Model." In *Microsimulation in Government Policy and Forecasting*, pp. 575–580. Elsevier.

Sutherland H (2007). "**EUROMOD**: The Tax-Benefit Microsimulation Model for the European Union." In *Modelling Our Future: Population Ageing, Health and Aged Care*, volume 16, pp. 483–488. Elsevier.

Tanton R, Edwards K (2013). *Spatial Microsimulation: A Reference Guide for Users*. Springer-Verlag. doi:10.1007/978-94-007-4623-7.

Tanton R, McNamara J, Harding A, Morrison T (2007). "Rich Suburbs, Poor Suburbs? Small Area Poverty Estimates for Australia's Eastern Seaboard in 2006." In *1st General Conference of the International Microsimulation Association*.

Voas D, Williamson P (2000). "An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata." *International Journal of Population Geography*, **6**(5), 349–366. doi:10.1002/1099-1220(200009/10)6:5<349::aid-ijpg196>3.0.co;2-5.

Whelan C, Maître B (2007). "Measuring Material Deprivation with EU-SILC: Lessons from the Irish Survey." *European Societies*, **9**(2), 147–173. doi:10.1080/14616690701217767.

Williamson P, Birkin M, Rees P (1998). "The Estimation of Population Microdata by Using Data From Small Area Statistics and Samples of Anonymised Records." *Environment and Planning A*, **30**(5), 785–816. doi:10.1068/a300785.

Affiliation:

Dimitris Kavroudakis
University of the Aegean
Department of Geography
81100, Mytilene, Lesvos, Greece
Telephone: +30/22510/36427
E-mail: dimitrisk@geo.aegean.gr
URL: http://www.dimitrisk.gr/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/

November 2015, Volume 68, Issue 2 *Submitted:* 2012-07-11
doi:10.18637/jss.v068.i02 *Accepted:* 2014-10-31