Comparing different spatial microsimulation frameworks

Tomintz M, Kosar B

Carinthia University of Applied Sciences
Department of Geoinformation and Environmental Technologies

November 07, 2014

Summary

Spatial microsimulation modelling was introduced in the 1980s but there is still a lack of open and easy to use frameworks. This obstacle was recognized by researchers and now there are frameworks available to access/download. The aim of this paper is to compare three different spatial microsimulation frameworks, as all of them have different algorithms implemented, that are available for free to the research community. The expected results are to find the best approach to model the Austrian smoking population for municipalities based on Austrian datasets and to list the pros and cons of the frameworks based on defined criteria.

KEYWORDS: comparing spatial microsimulation framework; smoking; simSALUD; Flexible Modelling Framework; spatial microsimulation with R

1. Introduction

Spatial microsimulation modelling is a novel method for creating small area estimates of data that are not available for small geographical scales. There are different approaches for microsimulation modelling but due to complexity, missing easy or free to use software frameworks, approaches can often not be used and therefore are implemented from scratch. There is an emerging need for comparing different spatial microsimulation methods (Clarke and Harding, 2013). Nowadays there are frameworks published for the community including handbooks and tutorials, so that people can use them.

The aim of this paper is to use and compare different spatial microsimulation frameworks, including FMF, Spatial Microsimulation with R and simSALUD, to find the best method for estimating the smoking population at municipality level for Austria. Criteria are defined for comparison, including preparing and import of input data files, speed of calculation, etc. The most interesting fact why doing this comparison is to see which method performs best for the existing input data.

2. Spatial microsimulation frameworks

For this paper, three spatial microsimulation frameworks are selected for comparison because they have the same aim but differ in the following facts: implemented algorithm, different programming and scripting languages respectively, the way of importing the input data and exporting the output data (simulated results), the access to the framework (installation, Web-based access), to name a few. The following subchapters describe each software framework in more detail based on the aforementioned criteria.

1.1. FMF (Flexible Modelling Framework)

The flexible modelling framework (FMF) is a software application which was created by Kirk Harland and his members of the Multi Agent Systems and Simulation (MASS) research group (Harland et al. 2012). The FMF software framework has been developed in the Java programming language which presents a graphical user interface (GUI) for connecting spatial analysis and other tools to data. The FMF has initially been used for generating realistic populations of Leeds and is currently being used to examine trends and processes within retail markets. The framework itself requires one single flat (records with no structured relationships) survey file (e.g.: csv) as well as for each constraint one single aggregated data (e.g.: census) file. The used algorithm is an iterative optimisation algorithm simplified from simulated annealing and creates synthetic population for the areas that distributes individuals from the microdata in the population areas. No programming skills are required and the application can be run on every computer without installing the software. The application allows the users to validate the results using a various number of validation methods such as cell percentage error, percentage error, Standard absolute error, total absolute error, standardized root mean square error, etc.
The framework was also developed as a generic box of plugins which allows external users to create their own plugins to contribute it to the project for further development and testing. Subsequently other useful tools have been added, including a cluster hunting tool to identify clusters in geographical data.

1.2. Spatial Microsimulation with R

Another approach to create a spatial microsimulation model was done by Robin Lovelace using the free software package R (Lovelace 2014). His decision of using the software R was the low-level language of R compared with other statistical programs based on a strong graphical user interface in combination with the great flexibility (many pre-made functions) for analysing and modelling the data. Bit high-level, compared to other general purpose languages such as C or Python. The model is based on the deterministic method to allocate individuals to areas called iterative proportional fitting (IPF). The tool can be used to either adapt the code on the input data of the user or to write it from scratch to learn the language R and the algorithm, as illustrated in Lovelace (2014). The tool requires as input data one individual-level dataset (csv format) and depending on the used function one aggregated area data or for each constraint one csv file. In general the framework requires to have some knowledge in the language R, at least for adopting the parameters for the own data. In contrast to the aforementioned framework, the software R offers packages to visualize the spatial data in R (Lovelace et al. 2014). Additionally the tool provides some integerisation methods to only allocate whole people to the result (Lovelace 2013).

1.3. simSALUD

simSALUD is a spatial microsimulation framework which was created as part of the research project SALUD (SpatiAL microsimUlation for Decision support) at the Carinthia University of Applied Sciences (Tomintz et al. 2013). The application was designed as a web spatial microsimulation application and consists of three wizard-based modules: simulation, validation, visualization. Currently, the algorithm implemented within simSALUD is a deterministic reweighting approach after Ballas 2005, O'Donoghue 2013 with an extension to integerise (Ballas 2005) the model outputs. Based on a static microsimulation approach also this framework requires as input one non-spatial survey population file (csv) and for each constraint one single csv file containing demographic and socio-economic population data for small areas (e.g.: municipalities). As the framework FMF, simSALUD requires no programming skills and needs to have only an internet connection for using the application. All three steps are guided through a wizard and allow the user an easy-to-use handling. After running the model, also this framework provides the user to validate and verify the models robustness. Available validation methods are for example: total absolute error, percentage error, percentage error, simple regression, etc. One feature of the framework is that the outputs of the simulation can be either visualized in form of a map on the simSALUD web-based framework itself or exported after the model run for further analyses in common geographic information software products.

3. Input data: case study "smokers"

For this paper the case study "smokers" is chosen to do the comparison with the different software frameworks. The aim is to estimate the smoking population starting at the age of 15 at municipality level. Therefore two datasets are of interest: the Austrian Health Survey 2006/07 and the registered-based census 2011. One requirement for the modelling process is the availability of common variables in both datasets that predicts being a smoker best as possible. Therefore, statistical pre-analysis are performed on the Austrian Health Survey 2006/07 that holds more than 15.000 persons, i.e. chi-square and regression analysis. Finally four variables are selected as constraints for the spatial microsimulation modelling, i.e. age (eleven categories), sex (two categories), marital status (four categories) and last completed education (three categories).

The input data is accessible from the Statistics Austria noting that the micro-dataset (Austrian Health Survey 2006/07) can be requested free of charge but the registered-based census 2011 (demographic and socio-economic variables at municipality level) is with costs.

The advantage, however, is the common requirement of all three frameworks for the input data files, which are comma-separated values (csv) files. So there is no additional huge amount of data preparation when testing all three frameworks. The only difference is that the spatial microsimulation with R framework (see 1.2) requires all data in one file whereas the other two frameworks (see 1.1 and 1.3) require the input data in separate files separately for each file.

4. Expected results

The case study for this paper will identify the simulated smoking population for Austria at municipality level. As there are different spatial microsimulation approaches available, different ones are getting tested to explore which one brings the best result based on the Austrian input data. The results are tested using statistical analysis. One main validation source is the Total Absolute Error (TAE) where the simulated constraints are compared with the registered-based data constraints, as this method is used within all three software frameworks. Further, the results from all three frameworks are getting mapped to identify if there are huge spatial variations.

Further, all three frameworks are getting compared based on the criteria mentioned in the introduction. During the analysis it is possible that further criteria are getting defined. This will show which spatial microsimulation approach works best for the Austrian datasets and also their pros and cons in terms of usage and simulation time.

5. Conclusion and future work

This paper explores the best spatial microsimulation approach to model the smoking population in Austria among three approaches that are embedded in different frameworks. With it, the frameworks are tested based on certain criteria, including simulation speed, programming knowledge, etc. It is known that there is a lack of available open spatial mirosimulation frameworks and therefore it is good to see that first attempts are being made. However, there is need for testing different approaches and frameworks under comparison of different criteria.

As the comparison is based on Austrian datasets only, the next planned step is to use cross-national datasets, e.g. from the UK. This dataset has then a different number of individuals from the health survey and a different number of areas that are modelled. The results are valuable for further framework developments and to help people in their decision which framework fits best for their simulation aims.

6. Acknowledgements

This research project SALUD is funded by the Federal Ministry for Transport, Innovation and Technology (bmvit) and the Austrian Science Fund (FWF) (project number TRP280-G16). The registered based census data is accessed from "STATcube – Statistical Database of STATISTICS AUSTRIA".

7. Biography

Melanie Tomintz is senior researcher and lecturer at the Carinthia University of Applied Sciences (CUAS), Department of Geoinformation and Environmental Technologies. Her main research interests are in the area of health-GIS (Geoinformation for health science), spatial simulation, spatial analysis and usability evaluation.

Bernhard Kosar is research assistant at the Carinthia University of Applied Sciences (CUAS), Department of Geoinformation and Environmental Technologies. During his internship at the Louisiana State University he gained deep knowledge in the fields of Geoinformation in combination with the development of Web-based systems, spatial databases and spatial simulations.

References

Ballas D, Clarke G P, Dorling D, Eyre H, Thomas B and Rossiter D (2005). SimBritain: a spatial microsimulation approach to population dynamics, *Population Space and Place,* 11:13–34.

Clarke G and Harding A (2013). *Conclusions and the future of spatial microsimulation modelling*, in Spatial microsimulation: a reference guide for users, eds. Tanton, R and Kimberley, E, Springer.

Harland K (2013). Microsimulation Model User Guide (Flexible Modelling Framework)', School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom.

Harland K, Heppenstall A J, Smith D and Birkin M (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *JASSS*.15(1) 1.

Lovelace R (2013). Supplementary information: a user manual for the integerisation of IPF weights using R, 1–18. Available online on arXiv or from the University of Leeds.

Lovelace R (2014). Introducing spatial microsimulation with R: a practical. National Centre for Research Methods 08.

Lovelace R and Cheshire J (2014). Introduction to visualising spatial data in R. NCRM Working Paper. EloGeo.

Tomintz M N, Kosar B and Garcia-Barrios V M (2013). simSALUD – a Web-based spatial microsimulation application to support regional health planning in Austria, Conference paper, European Regional Science Association, Palermo, Italy.