MISSIONE 4
ISTRUZIONE
RICERCA

FOSSR
Fostering Open Science in Social Science Research

Deliverable 5.8

Rocco Paolillo[1], Mario Paolucci[1]

1. CNR - Institute for Research On Population and Social Policies

08/02/2024

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Project Acronym	*FOSSR*
Agreement number	IR0000008
Project Full Title	Fostering Open Science in Social Science Research
Funding Scheme	Public notice of the Ministry of University and Research for the presentation of project proposals for “Strengthening and creating Research Infrastructures” to be financed under the PNRR

DELIVERABLE INFORMATION

Deliverable Number	D5.8
Deliverable Name	Complex modelling and artificial populations for agent-based simulations
Dissemination level	Confidential
Funding Scheme	Public notice of the Ministry of University and Research for the presentation of project proposals for “Strengthening and creating Research Infrastructures” to be financed under the PNRR
Grant Agreement number	IR0000008
Contractual date of delivery	31/01/2024
Deliverable Leader	Rocco Paolillo
WP/Task Responsible	Andrea Giovanni Nuzzolese
Keywords	Synthetic reconstruction, Complexity, Agent-based modeling, Algorithms, Iterative Proportional Fitting
Abstract	The FOSSR open-cloud will provide plenty of data to support social research and policymaking. However, details and abundance of data might not be enough to understand the emergence of the dynamics of collective phenomena such as, for instance, segregation or polarization. Agent-based modeling is a computational method that serves this scope, enabling the construction of artificial societies comprising virtual agents that simulate human cognitions and behaviors to interact with their context. By studying the interaction between agents as a drive to collective behavior, the method enables experimentation on the actual emergence of collective phenomena. Synthetic populations are a set of algorithms to assure that artificial populations are representative of their target population in terms of attributes of agents and their distribution, which includes also the reconstruction of data from independent sources of information. The goal of WP5.5 is to enable algorithms for synthetic populations and synthetic reconstruction in the FOSSR open cloud,

	providing a service for researchers who will use data from the FOSSR infrastructures to build reliable agent-based models. In this deliverable, we report our activities to the date. We provide an overview of complexity science, agent-based modeling, and their contribution to the FOSSR infrastructure. We show our formalization of the Iterative Proportional Fitting algorithm for synthetic reconstruction and its validation, and steps towards the integration in the workflow of the future FOSSR infrastructure in collaboration with other WPs.

DISCLAIMER

The content of this Deliverable reflects only the author's view. The European Commission and MUR are not responsible for any use that may be made of the information it contains.

Document history

Date	Version	Contributor(s)	Description	Supervision
08/01/2024	0.1	Rocco Paolillo	Layout	All authors
20/01/2024	0.2	Rocco Paolillo	Initial draft	All authors
24/01/2024	0.3	Mario Paolucci	First revision	All authors
25/01/2024	0.4	Rocco Paolillo	Second version	All authors
30/01/2024	0.5	Mario Paolucci, Rocco Paolillo	Final revision	All authors
31/01/2024	1	Rocco Paolillo	Last editing	All authors

Summary

List of Figures

List of Equations

List of acronyms

ACRONYM	*Description*
ABM	Agent-based modeling
CA	Certificate Authority
CESSDA	Consortium of European Social Science Data Archives
CNR	Consiglio Nazionale delle Ricerche
FOSSR	Fostering Open Science in Social Science Research
GIS	Geographic Information System
HIPF	Hierarchical Iterative Proportional Fitting
HTTPS	Hyper-Text Transfer Protocol Secure
IDE	Integrated development environment
IPF	Iterative Proportional Fitting
IT	Information Technology
MUR	Ministero dell'università e della ricerca (Ministry of University & Research)
NGOs	Non-governmental Organization(s)
PNRR	Piano Nazionale Ripresa e Resilienza
R&I	Research and Innovation
REST-API	Representational State Transfer Application Programming Interface
RIs	Research Infrastructure
RISIS	Research infrastructure for research and innovation policy studies
SAAS	Software as a service
SHARE	Survey of Health, Ageing and Retirement in Europe
SSL	Secure Socket Layer
TAE	Total Absolute Error
URI	Uniform Resource Identifier
WP	Work Package

1. Introduction

The background mission of the FOSSR project relies on creating and strengthening awareness and knowledge of data and methodologies used in empirical social sciences among a wide audience, through maintaining and reinforcing relevant Research Infrastructures (RI), while fostering the development of a research and social environment conducive to an open, shared and simplified data access via innovative interfaces. The project will concretely contribute to the effective implementation of the 'open science' for social science researchers by providing innovative tools and services for research, shared virtual environment for data access, and a wide and varied package of training courses, sessions and programmes (FOSSR DoW, 2022).

FOSSR adopts the common theme of the development of Open Science in the Italian context with the goal of creating a framework of tools and services for the social science scholar community involving the Rls in social sciences coordinated by CNR, namely CESSDA, SHARE and RISIS. The framework should take the form of an integrated knowledge sharing platform, a single point of access to all the tools and services made available by the Italian nodes of social science infrastructures.

FOSSR fosters the building of an Italian Open Science Cloud, along the lines of the European Open Science Cloud project, in which to integrate innovative services developed by the project for data collection, data curation and fairness and data analysis on economic and societal change.

FOSSR wants to promote toward multiple audiences, widespread knowledge and awareness of the data and methodologies employed in empirical social science, fostering the growth of a broad societal environment favourable to further thriving of social science research in Italy, providing easy, open, streamlined access to social science data through innovative interfaces

1.1. FOSSR Objectives and Ambition

FOSSR has the general aim of promoting, towards multiple audiences, a widespread knowledge and awareness of the data and methodologies employed in empirical social science, by providing (i) systematic and organised knowledge (also through summary harmonised data) about available social science data resources in Italian data archives, especially the CESSDA Archive, already object of the grand infrastructural proposal; (ii) resources supporting methodological advancement as to data collection and data analysis, especially important for RISIS to understand the design, the implementation, and the outcome of research and innovation policies, which can improve the robustness of empirical evidence produced for policy makers and to deal with new research questions, and (iii) tools and services to make publicly available advanced probability panels for longitudinal analyses to support important survey such as SHARE, complementing them with a network of online laboratories. The integration of this pool of resources shall concretely contribute to the realisation of open science for scholars in social sciences, going with an important program of scientific training for the production and analysis of social science based on FAIR[1] empirical data.

One further key objective is setting a new generation of young researchers in social sciences, by hiring several researchers and technologists with fixed time contracts, which will become highly skilled human resources in data science and data management in social science research, and by funding 20 PhD positions to train early career researchers in the field.

FOSSR is also aimed at fostering the growth of a broad societal environment favourable to further thriving of social science research in Italy, providing easy, open, streamlined access to social science data through innovative interfaces (data exploration portals, time series, interactive visualisations), and through online data analysis software aimed at students, along with divulgation resources about social science methodology. These aspects are particularly aimed at civil society organisations (NGOs, etc.), students, ordinary citizens, to foster a widespread societal awareness of social science data, results, methods to promote an easier and more user-friendly dissemination.

The main investment shall be on the creation of a suitable IT infrastructure and a network of data centers to provide researchers access online and onsite, and to support the workflow of the proposed collaborative projects. One of the subsequent goals of FOSSR will be to develop innovative tools and services for data collection and data analyses, and to support participating users in the acquisition and use of advanced equipment and software for social science research needed for collaborative studies and projects. The achievement of the above-mentioned goals can be reached by means of an online platform – the Open Cloud, acting also as a dissemination layer, intermediating between data producers, archives, and the broader shop floor of users (both scholars and stakeholders), assuring access intermediating between data producers, archives, and the broader shop floor of users (both scholars and stakeholders), assuring access also to multiple software interfaces, geared at different audiences, methodological content, and training materials.

1.2. Purpose and scope of this document

Social science is one of the branches of science devoted to studying societies and the relationships among individuals within those societies. The branches of social science include anthropology, economics, political science, psychology, and sociology. Population growth and the complexity of modern society have made the social sciences of utmost importance. Thanks to them, it is possible to thoroughly understand social dynamics to act by preventing or solving related problems. This approach requires much information to build models to properly simulate and understand individual and societal behaviours. Therefore, the interaction of the scientific community and the ability of scholars to access this information quickly and efficiently is of great importance.

[1] Findable, Accessible, Interoperable, Reusable

The aim of FOSSR is the creation of an Italian Open Science Cloud for the Social Sciences, which shall provide innovative tools and services to investigate issues related to contemporary societies' economic and societal change. To achieve such a result, referring to RIs in the social sciences that make FAIR-type data available is of primary importance. This type of data will be obtained by the RIs involved in social sciences coordinated by CNR, namely:

- CESSDA (Consortium of European Social Science Data Archives) provides the scientific community with facilities, tools, datasets, and certified services to conduct research activities of excellence in the social sciences domain.
- SHARE (Survey of Health, Aging, and Retirement in Europe) is an interdisciplinary and longitudinal survey on the economic, social, health, and well-being conditions of the 50+ population in twenty-seven European countries (plus Israel).
- RISIS (Research Infrastructure for Research and Innovation (R&I) Policy Studies), which provides data and services to support the development of a new generation of analyses and indicators for the study of science, technology, and innovation processes based on three main perspectives: actors involved to understand the role they play, topics addressed to understand the directionality of the R&I efforts, and geography of science and innovation.

FOSSR shall incorporate tools (hardware and software) and methods functional to research practices traceable to the paradigms of e-science, behavioural economics, and computational social sciences. In operational terms, the functionalities for which important innovations are expected are: data collection, data integration, data curation, data sharing, the creation of a survey facilitator, the construction of a social listening structure, and the activation of an artificial population facility

In operational terms, this framework should be an integrated knowledge-sharing platform, a single point of access to all the tools and services made available by the Italian nodes of social science infrastructures.

FOSSR has the general aim of promoting, towards multiple audiences, widespread knowledge and awareness of the data and methodologies employed in empirical social science by providing (i) systematic and organised knowledge about available social science data resources in Italian data archives; (ii) resources supporting methodological advancement as to data collection and data analysis, (iii) tools and services to make publicly available advanced probability panels for longitudinal analyses to support important survey The integration of this pool of resources shall concretely contribute to the realisation of open science for scholars in social sciences, going with an essential program of scientific training for the production and analysis of social science based on FAIR empirical data.

The WP5 focuses on designing and developing a national platform based on a distributed cloud computing infrastructure aimed at creating a single integrated system of the national nodes of CESSDA, RISIS, and SHARE.

This document aims to describe the activity 5.5 to the date concerning the modeling of complex systems via agent-based modeling. The core topic is the implementation of synthetic population and algorithms for synthetic reconstruction. The work of the package conveys the future

implementation of software as a service in the FOSSR infrastructure, together with other services offered by the project. Activities to this goal are presented.
The deliverable proceeds as follows: (i) Chapter 2 introduces the topic of complexity science and agent-based modelling and their contribution to the FOSSR project (ii) Chapter 3 delves into the specific topic of synthetic populations in the FOSSR project (iii) Chapter 4 shows the algorithm implemented so far. Finally (iv) Chapter 5 outlines activities and requirements for the implementation of algorithm(s) in software as a service in the FOSSR infrastructure, and (v) Chapter 6 drives conclusions and outlines the next steps of WP5.5.

2. COMPLEX MODELING AND ARTIFICIAL POPULATIONS FOR THE SOCIAL SCIENCES

2.1. The paradigm of social complexity

Most social issues of societies such as segregation or polarization are complex phenomena. The term social complexity indicates that these phenomena are the unintended consequence of the aggregated behavior of individual citizens and social institutions, which requires thinking of the society as a system of interacting components [1], [2], [3]. Understanding the complexity of such phenomena points to studying the dynamics of emergence of distributed action [4]. Causality of collective phenomena in this framework requires the consideration of at least two levels of the dynamic phenomenon: a micro-level of individual actors interacting in a constrained context of action, and a macro-level of the emerged observable phenomenon [5], [6]. Such ideas of causality that nowadays are fully embraced by academic areas such as analytical sociology [5], trace back to the 1990’s with James Coleman boat diagram [6]. This metaphor of social macro-phenomena shows how changes in the state of the system at the macro-level occur via generative mechanisms of situated action at the lower levels of the system, i.e. individual actors who take action to respond to their context (see Figure 1). We can describe the micro-macro causal emergence with the example of Schelling’s model, a notorious model to explain spatial segregation [7], [8]. The model wants to show how even mild preferences for similar ones (e.g. co-ethnics) wanted in the proximal neighborhood can generate a high level of spatial segregation due to cascade effects occurring between people relocating to satisfy their preferences. The macro constraint in the model is the density of the population that affects the distribution of people in space, thus affecting their neighborhood composition. At the micro-level, people hold a threshold of the desired percentage of similar ones in their proximal neighborhood. Action-formation mechanisms occur when citizens react to undesired neighborhoods relocating to a random location. The transformative mechanisms that underline the emergence of spatial segregation at the macro-level occur because of the cascade effects of the relocation of each individual citizen on the other, accumulating until reaching a tipping point where spatial segregation of neighborhoods cannot be reversed.

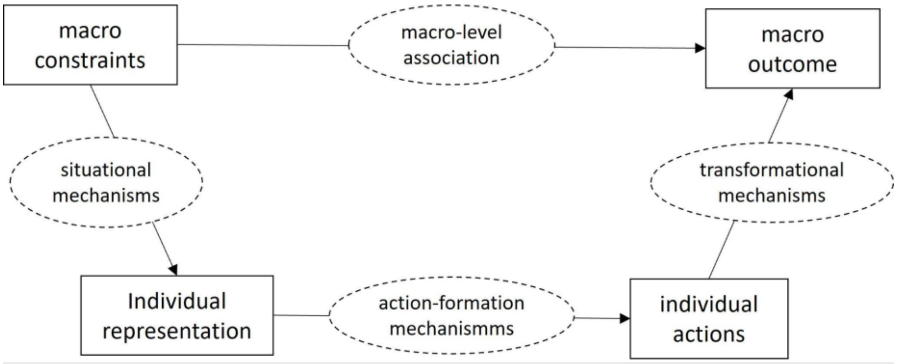


Figure 1. Coleman's boat of social mechanisms, adapted from Hedström & Ylikoski (2015)

2.2. Agent-based modeling for social complexity

The challenge posed by social complexity is that, as Epstein and Axtell stress out, many social phenomena are in fact already “emerged” in reality (p. 20) [9]. Therefore, the methodological challenge posed by the paradigm is how to reconstruct backward the micro-macro dynamics of social emergence so to experiment on them. Agent-based modeling (ABM) is a methodology that fits this scope, building artificial societies to study the emergence of collective distributed behavior. Constitutive elements of such societies are agents, i.e. virtual objects representing social actors such as citizens or institutions [4], [10]. Agents are provided with dynamic and stable attributes, along with desires, beliefs, and intentions to interact and adapt to their world [11], [12], [13]. The core of agent-based modeling is the *interaction* between agents as the key driver for the emergence of collective and distributed phenomena [4]. Taking the example of the cascade effects in the Schelling model described above, interactions at the micro-level of agents generate inter-dependent outcomes that, accumulated through time, give rise to macro phenomena that cannot be considered scalable attributes of individual agents [14]. Coding is fundamental in the translation of these concepts into a running machine to study such dynamics [15]. Writing a programming code, modelers can implement parameters for regulating the global environment of virtual agents, e.g. density of population, and micro-rules of agents' behavior [10], [15]. Outcomes of the interaction properly measured can be collected during the execution of the simulation and for all the conditions generated by the intersection of parameter levels, i.e. the parameter space [15]. The ultimate goal of the method is to experiment on and formalize the dynamics of the micro-macro link of social phenomena as a causal path to social complexity. *What-if scenarios* are an instrument to the case, running the simulation with different initial parameter settings, so to formalize what conditions underline what outcome by affecting the dynamics of the emergence of the phenomenon [16]. Initial conditions of what-if scenarios can be based either on theoretical speculations or empirical observations.

2.3. Agent-based modeling integration in the FOSSR open-cloud

Agent-based modeling can reply to several inquiries in the social sciences, spanning from theory-driven abstract models [17] to data-driven and descriptively detailed models [18]. The first set includes the description of social phenomena and the theoretical exposition of their dynamics, identifying key patterns in the parameter space [19]. The second set focuses either on the causal explanation of a phenomenon in a targeted empirical context, or the prediction of future events through the initialization with empirical data [19].[2] In both cases agent-based modeling can be a useful tool for both theoretical knowledge of social phenomena and for policy research [21], being a relevant methodology for the long-term goals of the FOSSR project. In post-ante analysis, pilot policies can be implemented in what-if scenarios to test scientifically the possible effects of policies, otherwise impossible in the real world without consequences to the population. Additionally, ABM can provide knowledge on the dynamics of the phenomenon the policy wants to intervene in, being useful for the phase of policy design. To sum up, agent-based modeling can be a complementary tool to the FOSSR open-cloud, integrating data from the infrastructure in experimental settings so to understand and predict the emergence of collective behavior. The implementation of agent-based modelling can be particularly useful to test the consequences of policies to be implemented in the Italian context. The goal of WP5.5 is to enable the usage of agent-based modeling in the FOSSR infrastructure as a service to researchers interested in the amount and diversity of data of the open-cloud. Considering the data-driven nature of the FOSSR project, the first concern is to enable agent-based models representative of the target population, which will be the focus of the rest of the deliverable.

3. SYNTHETIC POPULATIONS

Synthetic populations are a set of procedures and practices to ensure that artificial populations are representative of the target population [22], [23]. This is a requirement to ensure the reliability of the results of policy scenarios from agent-based simulations. The long-term goal of WP5.5 is to formalize algorithms to perform initialization of synthetic populations for agent-based modeling from the data available in the FOSSR server.

3.1. Literature Background

Literature on synthetic populations stems from urban geography, focusing on micro-simulations able to reproduce information conveyed in census tract units [24], [25]. Synthetic populations

[2] See the Second Online Seminar FOSSR on Agent-based Modeling and related discussion [20]

need to be a simplified representation of the real target population [23]. However, as Chapuis, Taillandier, and Drogul stress out in their review [23], the additional value of synthetic populations lies in being a “microscopic” (p.1) representation of such reality, meaning that each entity of the artificial system (e.g. a single household, person, etc.) needs to hold the attributes of its target and that the aggregated distributions of such simulated characteristics fit the measures in target data. This richness of information and comparability improves the reliability of results from simulation scenarios [26], [27]. Joint distribution of agents’ characteristics is a critical topic in this regard, due to the availability of detailed information it provides.
Historically, the literature identifies two main streams in the implementation of synthetic populations (see [23], [28], [29]):

- *Combinatorial Optimization*: this is the case where all information needed is known from data available, for instance, the joint distribution of population characteristics. The task of the modeler, in this case, is to scale down the size of the real population to the simulation scenario without modifying the percentage representation of each subclass of agents' distributions.

- *Synthetic Reconstruction*: in this case, much information is unknown to the modeler, and it needs to be “reconstructed” using estimators computed via additional procedures. This is the case, for instance, where joint distributions of agents’ characteristics are not available, but aggregated records are available for each independent variable (marginal distributions).

A third framework, more recent and less used is that of *statistical learning*, where the joint distribution of agents’ characteristics is computed not from the record of individual agents, but from regression models using a probabilistic approach to approximate the aggregated distribution of data from which computing the individual distribution of agents (see [30], [31]).

3.2. Synthetic Reconstruction for FOSSR services

As much literature highlights, synthetic reconstruction is the main venue for the synthetic population, and we identify it as the most promising approach for FOSSR server service for several reasons. First, synthetic reconstruction allows for the generation of new estimated data, mainly the distribution of joint characteristics, by integrating independent sources of information. This is a scenario very plausible in the FOSSR open cloud due to the abundance of different sources of information and different scales. Moreover, once synthetic reconstruction allows for estimating the distribution of joint characteristics of agents in an artificial population, this can be scaled to every size, allowing for great flexibility in the modeling practice. We believe these features of synthetic reconstruction best fit the goals of FOSSR's open science mission and build on the potentiality of data available in the cloud server. The long-term goals of WP5.5 are therefore to identify algorithms for the extraction of synthetic populations, in particular in the synthetic reconstruction framework, and enable their usage in the FOSSR infrastructure. In the next sections, we show our work done so far in this direction, breaking down the first steps of algorithm coding and implementation of the software as a service.

4. ALGORITHMS IMPLEMENTATION

We describe the planning and current state of WP5.5.[3] We envisage the construction of algorithms for synthetic reconstruction and complementary methods in a bottom-up style, increasing the level of complexity of scenarios the algorithms will be applied to. We show the current stage and developments for the implementation of the products of our work in a software as a service within the FOSSR infrastructure.

4.1. Iterative Proportional Fitting

4.1.1. Introduction

The Iterative Proportional Fitting (IPF) is the archetypal technique in synthetic reconstruction [23], [24] from which further extensions have been developed. We therefore first start from this algorithm as the first block of WP5.5 development. The algorithm serves to reconstruct integrated data from two independent variables. It is the equivalent of raking in statistics, where it is used to provide weights for categories underrepresented in the sample. At the agent-based modeling level, the algorithm is used for initialization with joint characteristics not available in the data. We choose to start from this algorithm for two reasons. First, despite its simplicity, it allows for a new estimation of joint distributions of agents' attributes which can be valuable for policy analysis with the simulation method. Second, most of the methods for synthetic reconstruction in literature were developed as an extension to the algorithm to overcome its limits in terms of its complexity and usability.

4.1.2. Formalization

The Iterative Proportional Fitting takes two sets of data into a contingency table, where the independent empirical data are positioned as marginal distributions in rows (variable 1) and columns (variable 2). Each resulting cell is a cross-category of the joint distribution of the two variables that need to be estimated by updating its value by a weight. This weight, as for all cells, is computed in an iterative procedure by rows (*r*) and columns (*c)* (see Figure 2). In statistical raking, each cell has an initial number equal to the representation of that category in the sample, with weights computed out of the known marginal distributions in the population. At the ABM level, where the algorithm is used for estimating the percentage of each cross-

[3] Software developed to date for the implementation of synthetic reconstruction and validation of the algorithm is available at: https://github.com/RoccoPaolillo/IPF_FOSSR5.5.git

category, the value of each cell can be set to 1. This is a scenario plausible when constructing abstract models.[4] Starting from the rows, for each cell $x_{r,c}$, its weight w_x^r is computed by dividing the *target* empirical marginal distribution of its row (sum of cells of that row: $\sum_T r$ in Equation 1) by the *fitted* computed distribution of that row ($\sum_F r$ in Equation 1). *Target* and *fitted* distributions are the equivalent of the sum of all cells of that row, which is updated at each iteration. Each cell is updated by multiplying its current value by the computed weight. At this point, the rows should fit the observed marginal distribution, but also the values of columns will be updated to the new *fitted* marginal distribution, due to the updated cells. The procedure is therefore re-iterated by the columns, updating each cell $x_{r,c}$ by weight w_x^c computed by dividing the *target* empirical marginal distribution of its column (sum of cells of the column: $\sum_T c$ in Equation 2) by the *fitted* computed distribution of that column ($\sum_F c$ in Equation 2). The update of columns will then change the fitted marginal distribution of rows again, moving back to the computation of weights by rows. The procedure stops when a benchmark chosen by the modeler is reached. Several benchmarks can be used (for a list, see [23]); we selected the Total Absolute Error (TAE), i.e. the sum of the absolute difference between observed marginal distribution and fitted marginal distribution, since it is a simple measure useful to our goal.

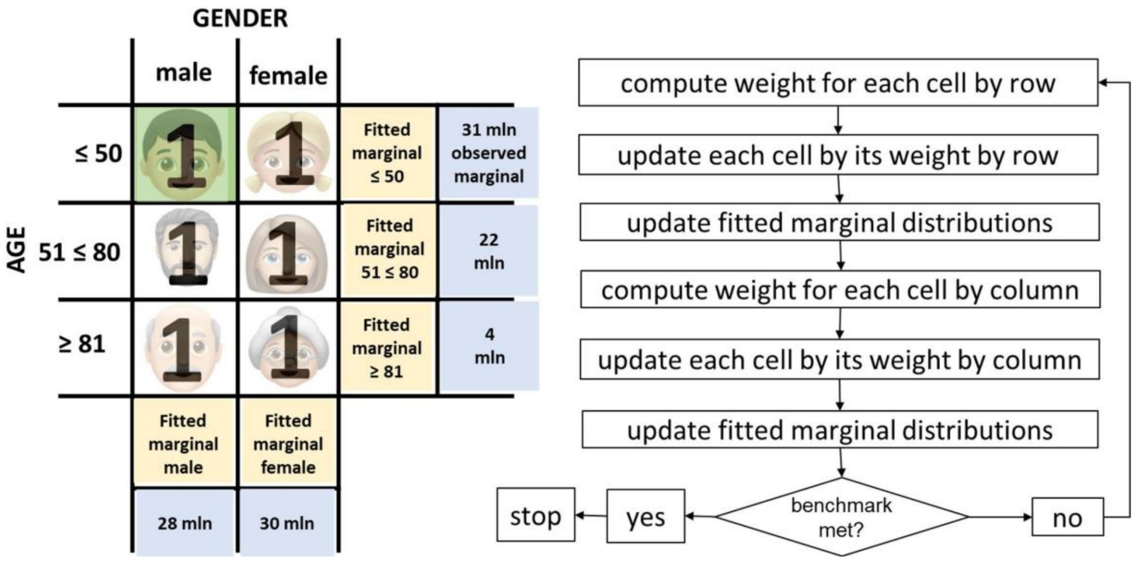

GENDER
male
female
AGE
≤ 50
51 ≤ 80
≥ 81
1
Fitted marginal ≤ 50
Fitted marginal 51 ≤ 80
Fitted marginal ≥ 81
31 mln observed marginal
22 mln
4 mln
Fitted marginal male
Fitted marginal female
28 mln
30 mln
compute weight for each cell by row
update each cell by its weight by row
update fitted marginal distributions
compute weight for each cell by column
update each cell by its weight by column
update fitted marginal distributions
benchmark met?
yes
stop
no

Figure 2. Iterative Proportional Fitting procedure

[4] It is possible in the code shared, however, to initialize the population of agents with a random distribution

$$x_{r,c} = x_{r,c} * w_x^r \qquad w_x^r = \frac{\sum_T r}{\sum_F r}$$

Equation 1

$$x_{r,c} = x_{r,c} * w_x^c \qquad w_x^c = \frac{\sum_T c}{\sum_F c}$$

Equation 2

4.1.3. Verification and Validation

We implemented the algorithm both in NetLogo and Python, see the next session for more details on the process and steps towards the implementation of the software as a service. For validation of the algorithm, we used ISTAT data on gender and age distribution in Italy for the year 2022.[5] The dataset was chosen because it provides both marginal and joint distribution data on the two variables, so that it can be used for the validation of the algorithm output. Gender had two levels: male and female. Age was an ordinal variable from “0 years old” to “100 and more” years old. The code takes the national level of reference and clusters the variable age in “0<= 50 years old”, “51 to 80 years old”, and “>=81 years old”. Gender categories are unaffected. The algorithm is run over such categories and percentages for each crossed category computed. The percentages from the implementation of the algorithm in both programming languages were compared with the empirical percentages of ISTAT dataset (see Figure 3). The comparison between NetLogo and Python (docking) shows no difference, meaning that the algorithm produces the same results. The comparison with ISTAT data confirms the algorithm results, smaller differences are due to decimal precision and expected in the execution of the Iterative Proportional Fitting (see [23], [24]).

[5] https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,POP,1.0/POP_POPULATION/DCIS_POPRES1/IT1,22_289_DF_DCIS_POPRES1_2,1.0 data reported on January 1, 2023. Last check on January 24, 2024.

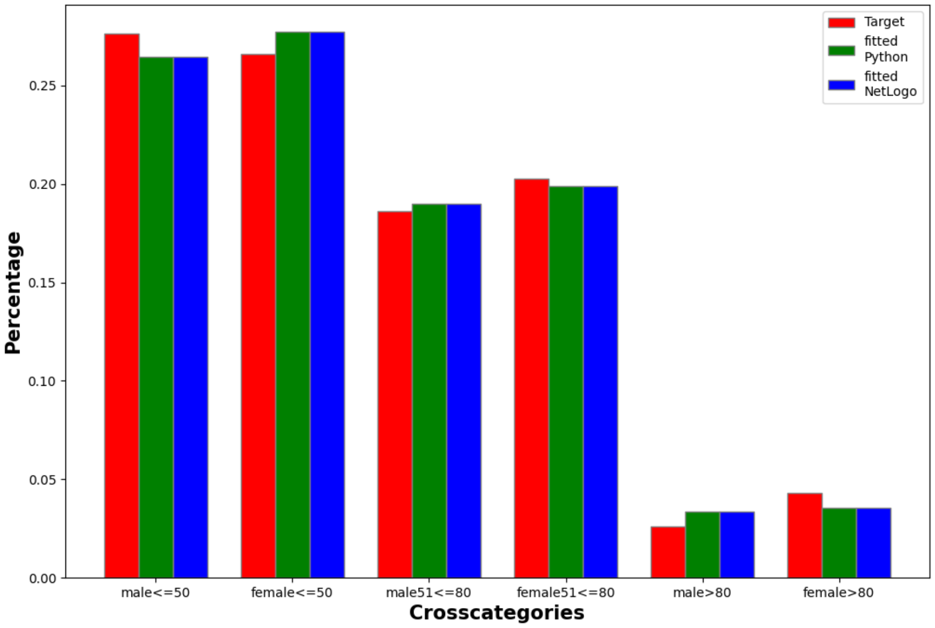


Figure 3. Validation test: comparison of IPF fitted cross-category percentages with empirical ISTAT percentages

5. Design of the software as a service

The activities of WP5.5 head to the implementation of a software as a service (SAAS) where users can run algorithms formalized over data from the FOSSR open-cloud to initialize synthetic populations. To this goal, WP5.5 will work closely with WP6 and WP7 who are in charge of the actual implementation of the physical server. Consultations with WP6 and WP7 are ongoing. The strategy suggested is that of *grey boxes*, i.e. to operate independently in the development of our algorithms and incorporate them in the open-cloud server once completed together with other FOSSR services. We first show a detailed implementation of the algorithm and its reproducibility. We then provide an overview of the requirements for the software and workflow as expected from the point of view of the user.

5.1. Implementation of the code

In line with FAIR principles FOSSR is built on, we implemented the Iterative Proportional Fitting in two open software, NetLogo and Python, for several reasons. NetLogo is a well-known and extremely user-friendly agent-based simulation platform, coming with an IDE that enables coding and a specific language. This reflects the principle of accessibility. Python is a high-level general-purpose programming language, meaning it focuses on a high level of abstraction and on the programming logic, independent of the hardware component of the machine where it is run, and using a language similar to natural human language to increase its usability. While each model written in NetLogo will produce an independent software to be executed, code written in Python can be collected into a script and be easily incorporated into other source code. Thus, Python offers higher interoperability than NetLogo. Out of this characteristic, Python increases the chance of collaboration between WPs for the implementation of algorithms in the FOSSR cloud server. We implemented the IPF in both languages and made the code available for reusability (see footnote 3).

In the *ipf.py* file (in Python), the algorithm is executed by *ipf_update* function, while in the *ipf.nlogo* file (in NetLogo) by the command block *update_weights*.

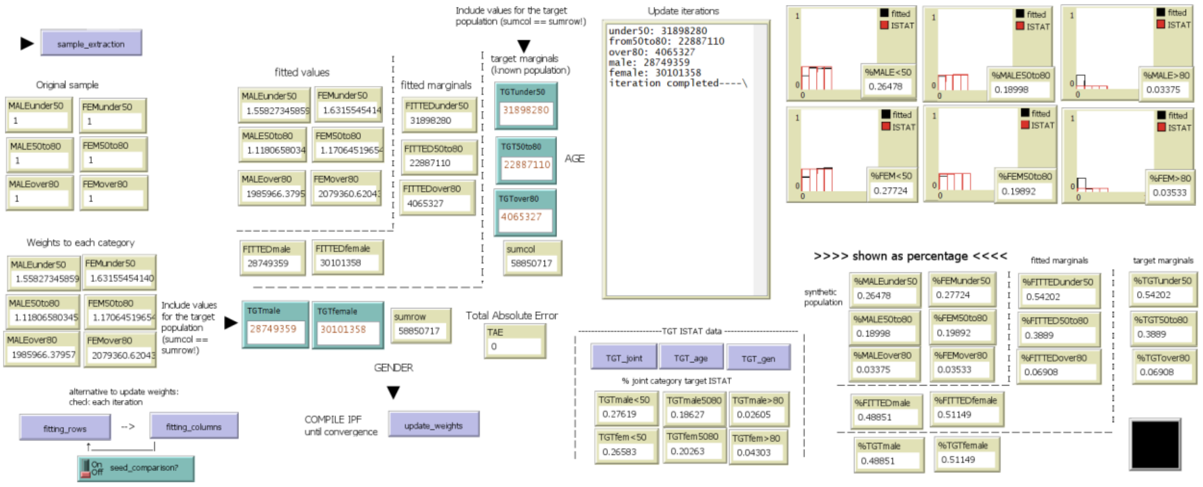
For both the programming languages, the pseudocode showing the main steps of the algorithm is as follows, for details see the repository where code is stored:

```
CALCULATE TAE
LOOP IF TAE > threshold:[6]
    UPDATE ROWS:
        FOR each cell
            compute weight by row
            update cell by weight
        ENDFOR
    UPDATE COLUMNS:
        FOR each cell
            compute weight by column
            update cell by weight
        ENDFOR
    UPDATE TAE
ENDLOOP
```

Figure 4. Pseudocode of the Iterative Proportional Fitting implementation

Both NetLogo and Python versions of the code provide a file with data output for the synthetic population initialization, in .txt format for NetLogo and in .csv format for Python.

[6] We set a threshold 0.0001 to be close to 0



Original sample
MALEunder50 1
FEMunder50 1
MALE50to80 1
FEM50to80 1
MALEover80 1
FEMover80 1

Weights to each category
MALEunder50 1.55827345859
FEMunder50 1.63155454140
MALE50to80 1.11806580345
FEM50to80 1.17064519654
MALEover80 1985966.37957
FEMover80 2079360.62043

fitted values
MALEunder50 1.5582734585
FEMunder50 1.6315545414
MALE50to80 1.1180658034
FEM50to80 1.1706451965
MALEover80 1985966.3795
FEMover80 2079360.6204
FITTEDmale 28749359
FITTEDfemale 30101358

fitted marginals
FITTEDunder50 31898280
FITTED50to80 22887110
FITTEDover80 4065327

target marginals
TGTunder50 31898280
TGT50to80 22887110
TGTover80 4065327
sumcol 58850717
TGTmale 28749359
TGTfemale 30101358
sumrow 58850717
TAE 0

```
under50: 31898280
from50to80: 22887110
over80: 4065327
male: 28749359
female: 30101358
iteration completed----\
```

%MALE<50 0.26478
%MALE50to80 0.18998
%MALE>80 0.03375
%FEM<50 0.27724
%FEM50to80 0.19892
%FEM>80 0.03533

%MALEunder50 0.26478
%FEMunder50 0.27724
%MALE50to80 0.18998
%FEM50to80 0.19892
%MALEover80 0.03375
%FEMover80 0.03533
%FITTEDmale 0.48851
%FITTEDfemale 0.51149
%TGTmale 0.48851
%TGTfemale 0.51149
%FITTEDunder50 0.54202
%FITTED50to80 0.3889
%FITTEDover80 0.06908
%TGTunder50 0.54202
%TGT50to80 0.3889
%TGTover80 0.06908

TGTmale<50 0.27619
TGTmale5080 0.18627
TGTmale>80 0.02605
TGTfem<50 0.26583
TGTfem5080 0.20263
TGTfem>80 0.04303

Figure 5. NetLogo interface for the Iterative Proportional Fitting

5.2. Implementation in the FOSSR Infrastructure

We have formalized the procedure we would expect when using the service provided by FOSSR for the extraction of synthetic populations. The expected workflow starts with the user accessing the FOSSR web-server, selecting the variables of interest, and asking for the server to run the execution of algorithms over the data selected. The workflow ends when the sender receives an output with data distribution computed for the synthetic population to be implemented in an agent-based model. To enhance the interoperability of the software as a service, the output of the algorithm execution should be available in the most common and readable formats, such as .csv, .json, or .txt file. The outcome with conditions for the initialization of agent-based models should be in such formats to adhere to the principle of interoperability of the FOSSR project.

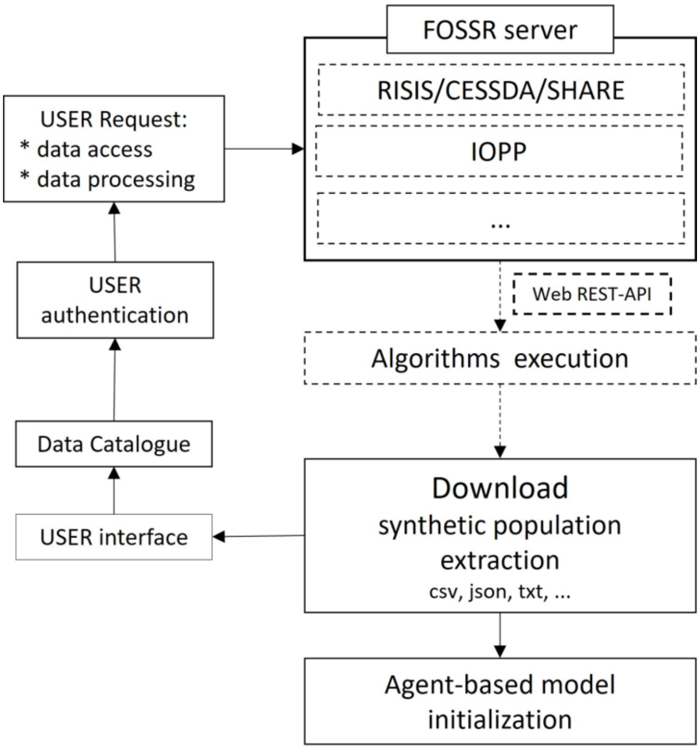
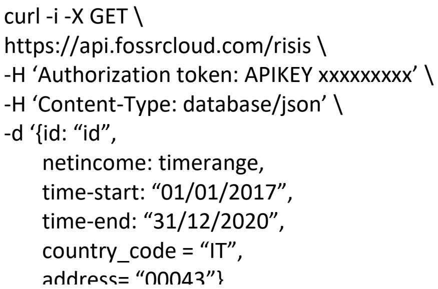


Figure 6. Workflow expected

5.2.1. Web REST-API

As for many services of the FOSSR server, the extraction of synthetic populations is expected to be run via a Web REST-API once the database with stored data is ready. The current implementation of the service and the associated decisions, similar to other web-based FOSSR services, falls under the purview of WP6 and WP7, with which we maintain ongoing communication. In this deliverable, we present a set of guidelines outlining the anticipated objectives for implementing the generated algorithms.

The acronym REST-API stands for *Representational State Transfer Application Programming Interface* and it proposes a set of guidelines for allowing communication between software over the Internet for the execution of a task. The endpoints in the workflow expected are the client user consumer and the FOSSR server provider. The communication between user and server will occur with encrypted requests written in HTTPS (Hyper-Text Transfer Protocol Secure). Once the server is implemented, it will be provided with a Uniform Resource Identifier (URI) and an encryption security is expected, such as a Secure Socket Layer (SSL), including a Certificate Authority (CA) for the identification of the user and to monitor access to the server information and services. As a measure of security, asymmetric encryption is advisable. In this strategy, the user receives a public key to encrypt their messages from the server, then the server uses a private key to decrypt those messages. The GET HTTPS method is expected to retrieve the resources over which the algorithms need to be executed (see fig. HTTP). The request should also include a manipulation of the data requested by applying the algorithms provided by WP5.5, so to end the procedure with a download to the user of data needed to initialize an agent-based model with the synthetic population. Examples of data extraction are the percentage distributions used in the verification and validation of the algorithm presented in Chapter 4. We expect the Web REST-API to comply with some best practices. These include a uniform interface: all requests and responses follow a common protocol via HTTPS messages; being stateless: each interaction between client and server is independent of others and complete in terms of information to satisfy the client request; cacheable: keeping memory of methods used by the client so to retrieve when re-used and facilitate automation of the task.



```
curl -i -X GET \
https://api.fossrcloud.com/risis \
-H ‘Authorization token: APIKEY xxxxxxxxx’ \
-H ‘Content-Type: database/json’ \
-d ‘{id: “id”,
      netincome: timerange,
      time-start: “01/01/2017”,
      time-end: “31/12/2020”,
      country_code = “IT”,
      address= “00043”}
```

Figure 7. Example of GET HTTPS request

6. Conclusions and steps ahead

Agent-based modeling can be a useful method for the long-term goals of the FOSSR infrastructure, allowing for the integration of data in simulation scenarios to experiment with the emergence of complex phenomena. While this can be useful for testing policies, we should guarantee that artificial populations are well representative of the context described by FOSSR data. To this goal, WP5.5 delves into the development of synthetic populations within the synthetic reconstruction framework, in collaboration with other WPs for the implementation of a software as a service in the FOSSR open-cloud. The first step of our agenda, i.e. the implementation of the Iterative Proportional Fitting was successful. The next steps are to overcome the limits known in the literature of the Iterative Proportional Fitting and that need to be accounted for in the long-term usage of the FOSSR open-cloud. First, the Iterative Proportional Fitting does not account for nested or multi-layered data, e.g. mapping individuals to households or households to neighborhoods [31], [32]. In the recent development of the literature, this is addressed via Hierarchical Iterative Proportional Fitting (HIPF) [33]. A generalizable implementation of the HIPF is the next step of WP5.5. An additional feature of this extension is the explicit spatial representation of agents mapping to GIS units [34]. In a further step, the algorithm should account for multiple dimensions that intersect within the same individual, a concept known in the literature as dimensionality course (see [35], [36]).

As for the implementation of the software as a service in the FOSSR infrastructure, we are in contact with WP6 and WP7 while the physical server and the infrastructure are developed. We will provide the WPs with the algorithms validated. When testing environments are available, we envisage test runnings for the user-friendly execution of algorithms, aligning with the requirements described above. Additionally, we will make the algorithms we formalize available via public repositories, in line with FAIR principles of FOSSR.

7. References

[1] P. Hedström e P. Bearman, *The Oxford Handbook of Analytical Sociology*. Oxford University Press, 2009.

[2] C. Cioffi-Revilla, *Introduction to Computational Social Science. Principles and Applications*. Springer London, London, 2014.

[3] M. W. Macy, «Social Simulation: Computational Models», in *International Encyclopedia of the Social & Behavioral Sciences*, 2015, pp. 701–705. doi: 10.1016/B978-0-08-097086-8.43092-8.

[4] M. W. Macy e R. Willer, «FROM FACTORS TO ACTORS: Computational Sociology and Agent-Based Modeling INTRODUCTION: AGENT-BASED MODELS AND SELF-ORGANIZING GROUP PROCESSES», *Annual Review of Sociology*, vol. 28, pp. 143–166, 2002, doi: 10.1146/annurev.soc.28.110601.141117.

[5] P. Hedström e P. Ylikoski, «Analytical Sociology», in *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, Elsevier Inc., 2015, pp. 668–673. doi: 10.1016/B978-0-08-097086-8.44071-7.

[6] J. S. Coleman, *Foundations of Social Theory*. Cambridge, MA: Harvard University Press, 1994.

[7] T. C. Schelling, «American Economic Association Models of Segregation», 1969.

[8] T. C. Schelling, D. Card, A. Mas, J. Rothstein, e T. C. Schelling, «Dynamic models of segregation», *The American Economic Review*, vol. 1, fasc. 2, pp. 143–186, 1971, doi: 10.1080/0022250X.1971.9989794.

[9] J. M. Epstein e R. L. Axtell, *Growing Artificial Societies: Social Science From the Bottom Up (Complex Adaptive Systems)*, 1St Edition. Brookings Institution Press MIT Press, 1996. [Online]. Disponibile su: http://www.worldcat.org/isbn/0262550253

[10] N. Gilbert e K. G. Troitzsch, *Simulation for the Social Scientist*. Maidenhead, UK: McGraw-Hill Education, 2005.

[11] V. Grimm *et al.*, «A standard protocol for describing individual-based and agent-based models», *Ecological Modelling*, vol. 198, fasc. 1–2, pp. 115–126, set. 2006, doi: 10.1016/j.ecolmodel.2006.04.023.

[12] V. Grimm, U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, e S. F. Railsback, «The ODD protocol: A review and first update», *Ecological Modelling*, vol. 221, fasc. 23, pp. 2760–2768, nov. 2010, doi: 10.1016/j.ecolmodel.2010.08.019.

[13] V. Grimm *et al.*, «The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism», *JASSS*, vol. 23, fasc. 2, p. 7, 2020, doi: 10.18564/jasss.4259.

[14] E. Bruch e J. Atwell, «Agent-Based Models in Empirical Social Research», *Sociological Methods and Research*, vol. 44, fasc. 2, pp. 186–221, mag. 2015, doi: 10.1177/0049124113506405.

[15] S. F. Railsback e V. Grimm, *Agent-based and individual-based modeling: a practical introduction. Princeton university press.* Princeton Press, 2011.

[16] F. Squazzoni, W. Jager, e B. Edmonds, «Social Simulation in the Social Sciences: A Brief Overview», *Social Science Computer Review*, vol. 20, fasc. 10, pp. 1–16, 2013, doi: 10.1177/0894439313512975.

[17] R. Axelrod, «Advancing the art of simulation in the social sciences», *Complexity*, vol. 3, fasc. 2, pp. 16–22, 1997.

[18] B. Edmonds e S. Moss, «From KISS to KIDS – an "anti-simplistic" modelling approach», in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, P. Davidsson, B. Logan, e K. Takadama, A c. di, New York, NY: Springer Berlin, 2004, pp. 130–144. [Online]. Disponibile su: http://cfpm.org

[19] B. Edmonds *et al.*, «Different Modelling Purposes».

[20] M. Paolucci e R. Paolillo, «FOSSR Second Online Seminar: Agent-based Modeling: contributions for the Social Sciences and FOSSR.» 26 gennaio 2024. [Online]. Disponibile su: https://doi.org/10.5281/zenodo.10571855

[21] N. Gilbert, P. Ahrweiler, P. Barbrook-Johnson, K. P. Narasimhan, e H. Wilkinson, «Computational Modelling of Public Policy: Reflections on Practice», *JASSS*, vol. 21, fasc. 1, p. 14, 2018, doi: 10.18564/jasss.3669.

[22] A. Adiga *et al.*, «Generating a synthetic population of the United States».

[23] K. Chapuis, P. Taillandier, e A. Drogoul, «Generation of Synthetic Populations in Social Simulations: A Review of Methods and Practices», *JASSS*, vol. 25, fasc. 2, p. 6, 2022, doi: 10.18564/jasss.4762.

[24] R. Lovelace, M. Birkin, D. Ballas, e E. Van Leeuwen, «Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique», *JASSS*, vol. 18, fasc. 2, p. 21, 2015, doi: 10.18564/jasss.2768.

[25] D. Ballas, Broomhead, Tom, e Jones, Phil Mike, «Spatial Microsimulation and Agent-Based Modelling», in *The Practice of Spatial Analysis: Essays in memory of Professor Pavlos Kanaroglou*, Springer Berlin Heidelberg, pp. 69–84.

[26] D. Ziemke, K. Nagel, e R. Moeckel, «Towards an Agent-based, Integrated Land-use Transport Modeling System», *Procedia Computer Science*, vol. 83, pp. 958–963, 2016, doi: 10.1016/j.procs.2016.04.192.

[27] J. Y. Guo e C. R. Bhat, «Population Synthesis for Microsimulating Travel Behavior», *Transportation Research Record*, vol. 2014, fasc. 1, pp. 92–101, gen. 2007, doi: 10.3141/2014-12.

[28] N. Jiang, A. T. Crooks, H. Kavak, A. Burger, e W. G. Kennedy, «A method to create a synthetic population with social networks for geographically-explicit agent-based models», *Comput.Urban Sci.*, vol. 2, fasc. 1, p. 7, feb. 2022, doi: 10.1007/s43762-022-00034-1.

[29] B. Farooq, M. Bierlaire, R. Hurtubia, e G. Flötteröd, «Simulation based population synthesis», *Transportation Research Part B: Methodological*, vol. 58, pp. 243–263, dic. 2013, doi: 10.1016/j.trb.2013.09.012.

[30] L. Sun, A. Erath, e M. Cai, «A hierarchical mixture modeling framework for population synthesis», *Transportation Research Part B: Methodological*, vol. 114, pp. 199–212, ago. 2018, doi: 10.1016/j.trb.2018.06.002.

[31] B. F. Yameogo, P.-O. Vandanjon, P. Gastineau, e P. Hankach, «Generating a Two-Layered Synthetic Population for French Municipalities: Results and Evaluation of Four Synthetic Reconstruction Methods», *JASSS*, vol. 24, fasc. 2, p. 5, 2021, doi: 10.18564/jasss.4482.

[32] F. Gargiulo, S. Ternes, S. Huet, e G. Deffuant, «An Iterative Approach for Generating

Statistically Realistic Populations of Households», *PLoS ONE*, vol. 5, fasc. 1, p. e8828, gen. 2010, doi: 10.1371/journal.pone.0008828.

[33] K. Müller e K. W. Axhausen, «Hierarchical IPF: Generating a synthetic population for Switzerland», 2011.

[34] K. Chapuis, P. Taillandier, M. Renaud, e A. Drogoul, «Gen*: a generic toolkit to generate spatially explicit synthetic populations», *International Journal of Geographical Information Science*, vol. 32, fasc. 6, pp. 1194–1210, giu. 2018, doi: 10.1080/13658816.2018.1440563.

[35] D. Casati, K. Müller, P. J. Fourie, A. Erath, e K. W. Axhausen, «Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking», *Transportation Research Record*, vol. 2493, fasc. 1, pp. 107–116, gen. 2015, doi: 10.3141/2493-12.

[36] R. J. Beckman, K. A. Baggerly, e M. D. McKay, «Creating synthetic baseline populations», *Transportation Research Part A: Policy and Practice*, vol. 30, fasc. 6, pp. 415–429, nov. 1996, doi: 10.1016/0965-8564(96)00004-3.