Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Intro

Finanziato dall'Unione europea
NextGenerationEU

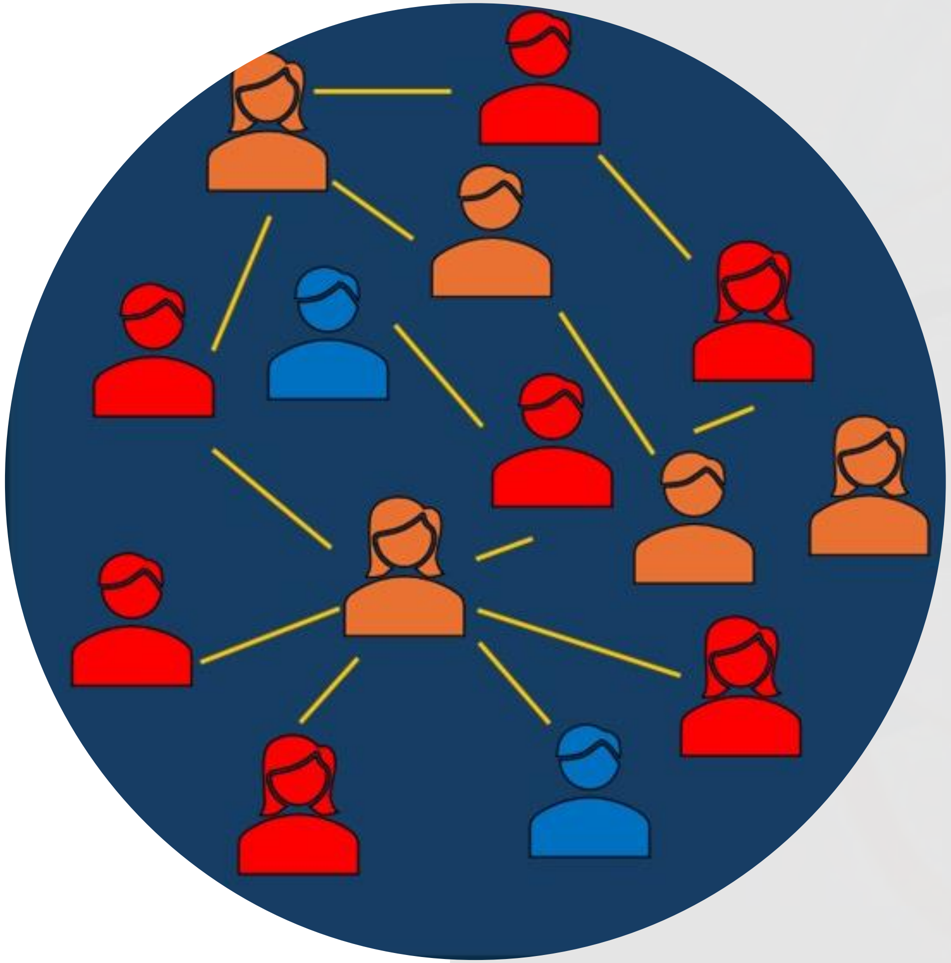
Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

IRPPS

part one: Intro Complexity & ABM
part two: A case study and a tool for ABM
part three: Synthetic Populations & FOSSR service



Rocco Paolillo
CNR-IRPPS

rocco.paolillo@cnr.it
@roccopaolillo.bsky.social

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NRRP Grant agreement n. MUR IR0000008

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA

Consiglio Nazionale
delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

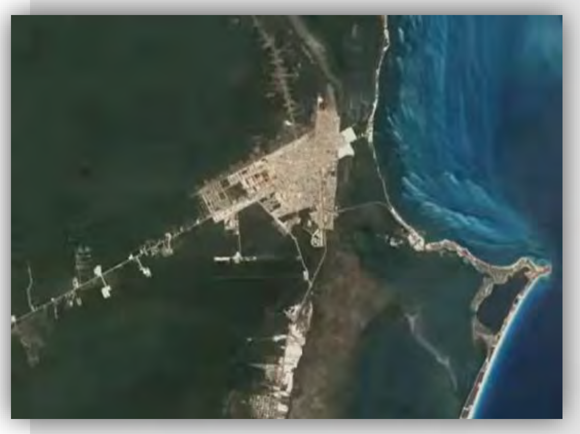
CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Part One

Social complexity and social mechanisms



- Most of social phenomena are inherently collective phenomena
- They assume the definition of a system made of interacting components (e.g. market place, urban landscape, welfare state) where the phenomenon unfolds
- Common scope of different disciplines is to unfold the mechanisms that unfold the dynamics of the phenomenon

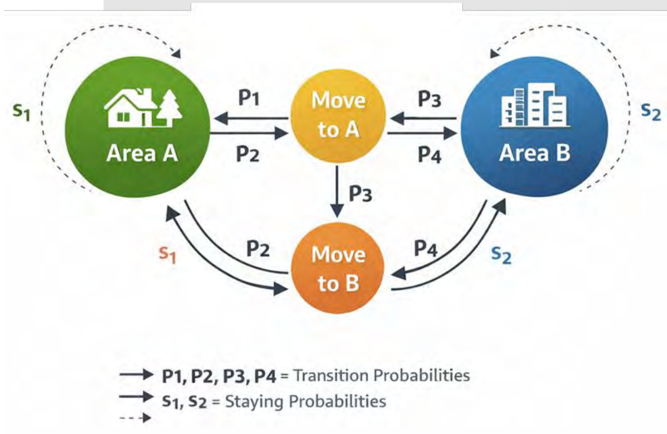
If the concepts of system and the goal to identify mechanisms to disclose the phenomenon is common to many disciplines, different perspectives apply

Different perspectives on the concept of *emergence* and *mechanisms*

Macro

Macro to Macro

- Markov Chain
- Evolution of processes depending on the previous state of the system as unit of analysis

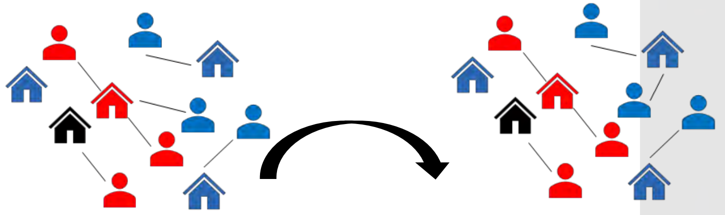


P1, P2, P3, P4 = Transition Probabilities
S1, S2 = Staying Probabilities

Meso

Meso to Macro

- Random graph models
- Evolution of networks as units of analysis



	🟥	🟦	🟦	⬛	🟦
🔴	1	0	0	0	0
🔵	0	1	0	0	0
🔵	0	0	0	0	1
🔴	0	0	0	1	0
🔵	1	0	0	0	0
🔴	1	0	0	0	0

	🟥	🟦	🟦	⬛	🟦
🔴	1	0	0	0	0
🔵	0	1	0	0	0
🔵	0	0	0	0	1
🔴	0	0	0	1	0
🔵	0	0	0	0	1
🔴	1	0	0	0	0

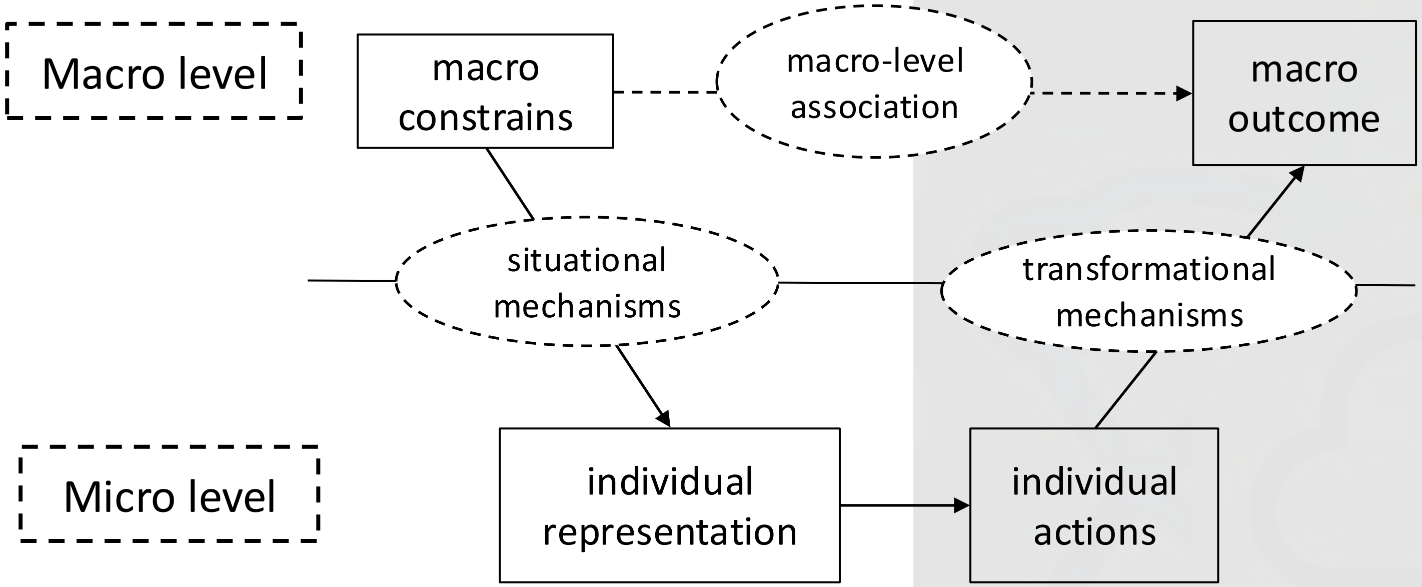
Micro

- Interactions of individuals and institution as the interacting components of the system

Micro to Macro

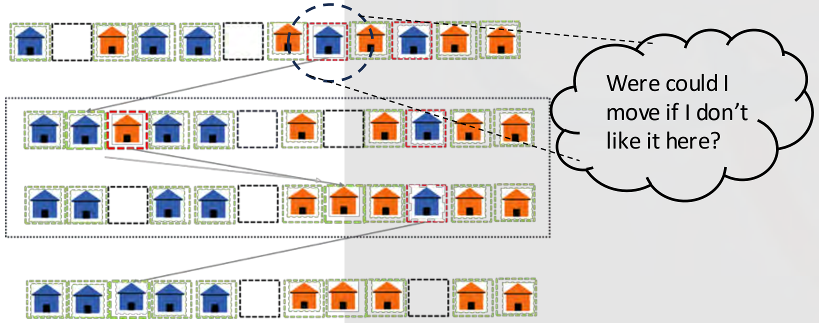
- But sometimes, we might be interested in those mechanisms that move from the micro level, e.g. citizens/institutions with their attitudes, motivations and course of action, but get **outside of the individual agency and inglobe the interaction** of individuals as explicative mechanism of emergence

- The phenomenon is an aggregated, mutual adaptation of individuals rather than the sum of individual action



v

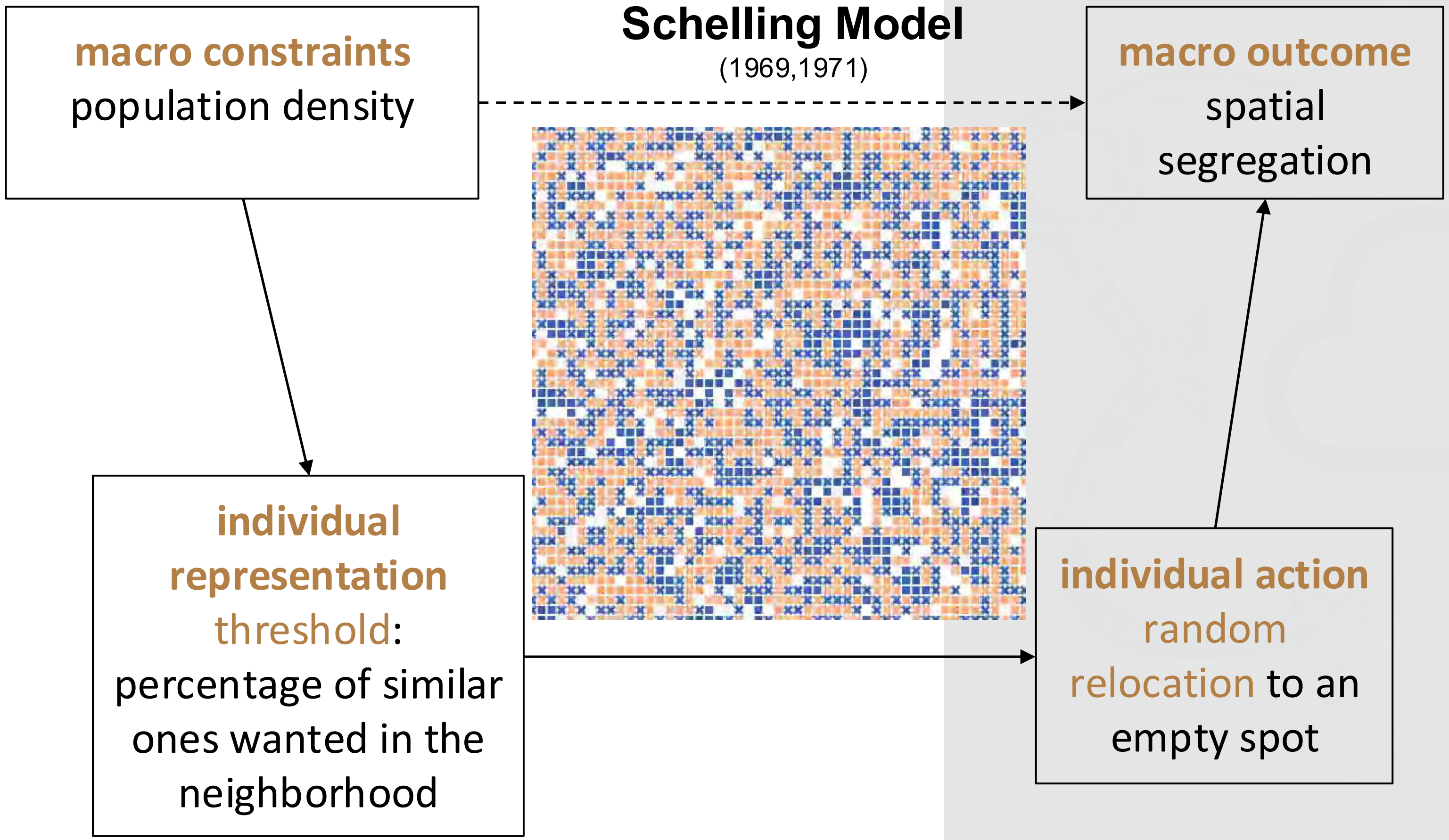
Analytical Sociology
Agent-based Modeling



Coleman Boat (1994), additions by Hedström and Ylikoski (2010), adapted

Example of spatial segregation

- **actors:** households evaluating their neighborhood
- **individual behavior:** preference for a percentage (threshold) of similar ones in their neighborhood (homophily)
- **mechanism of emergence:** cascade effects where the behavior of one household influences the composition of neighborhood and preference of others
- **unexpected outcome:** high levels of spatial segregation, also for mild threshold preferences



Schelling Model
(1969,1971)

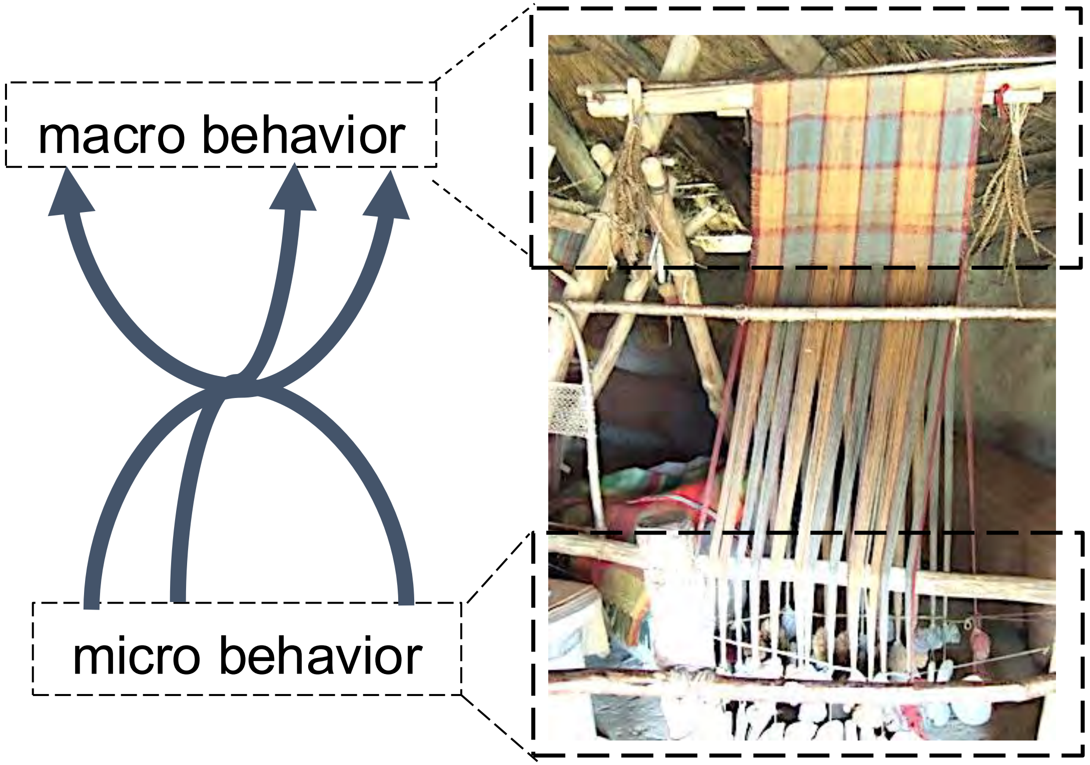
macro constraints
population density

macro outcome
spatial segregation

individual representation
threshold: percentage of similar ones wanted in the neighborhood

individual action
random relocation to an empty spot

Agent-based Modeling



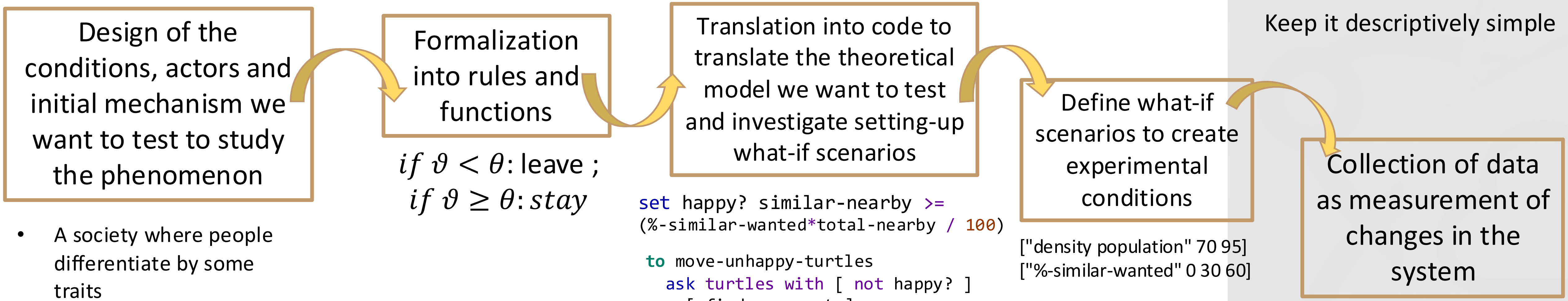
macro behavior

micro behavior

- **Simulation method** tailored to model the interacting components that constitute the system, e.g. agents representing citizens in an **artificial society**
- We can manipulate both attributes and plan of actions of agents and observe the consequences of interaction of agents executing their plans.
- By manipulating plans and conditions where the agents interact and adapt, we can experiment on and formalize the dynamics of emergence of the collective phenomenon

Social computing with agent-based modeling

KISS
Keep it short simple, stupid
KIDS
Keep it descriptively simple



Design of the conditions, actors and initial mechanism we want to test to study the phenomenon

- A society where people differentiate by some traits
- They stay in a neighborhood if certain threshold of similarity are satisfied
- Can segregation emerge even if the threshold is not that high?

Formalization into rules and functions

$if\ \vartheta < \theta$: leave ;
$if\ \vartheta \geq \theta: stay$

Translation into code to translate the theoretical model we want to test and investigate setting-up what-if scenarios

```
set happy? similar-nearby >=
(%-similar-wanted*total-nearby / 100)

 to move-unhappy-turtles
   ask turtles with [ not happy? ]
     [ find-new-spot ]
 end
```

Define what-if scenarios to create experimental conditions

["density population" 70 95]
["%-similar-wanted" 0 30 60]

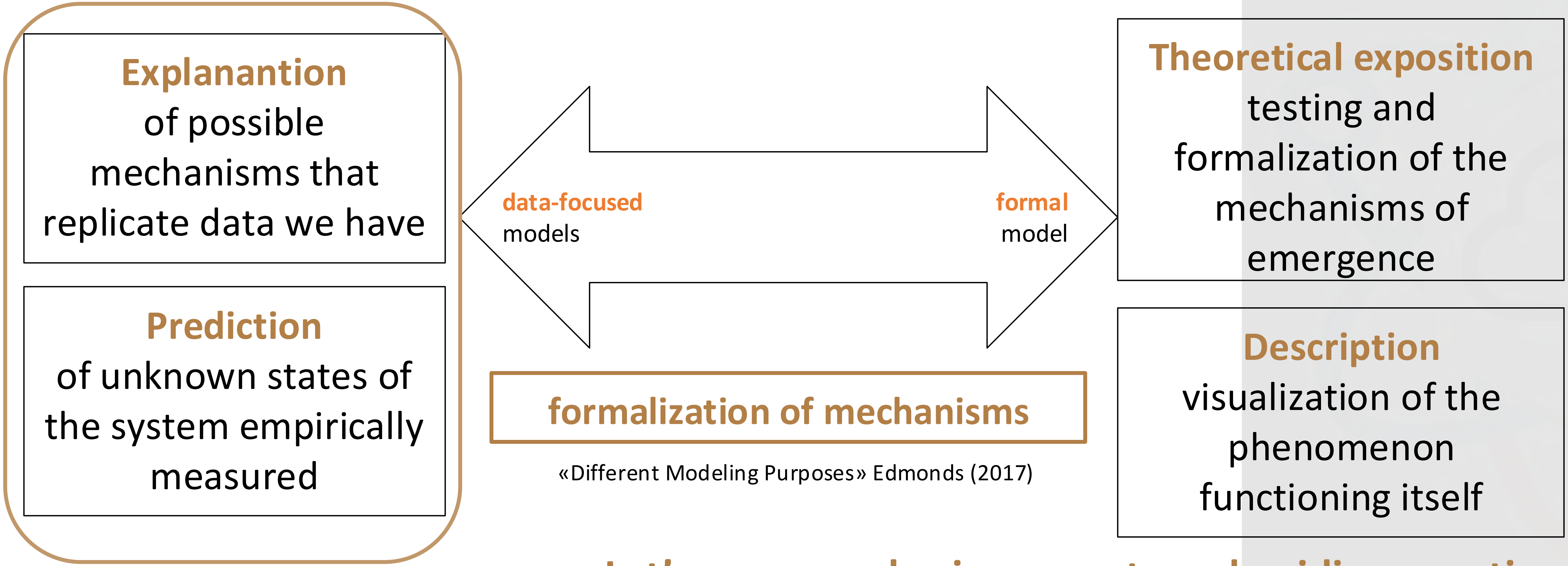
Collection of data as measurement of changes in the system

mean exposure to similars in the neighborhood of agents when no one relocates anymore

'If you can grow it, you have explained it' (Epstein, 2006)

'If you don't know how you grew it, you didn't explain it.' (Macy & Flache, 2009, p.263)

What is agent-based modeling useful for?



Explanantion of possible mechanisms that replicate data we have

Prediction of unknown states of the system empirically measured

formalization of mechanisms

«Different Modeling Purposes» Edmonds (2017)

Theoretical exposition testing and formalization of the mechanisms of emergence

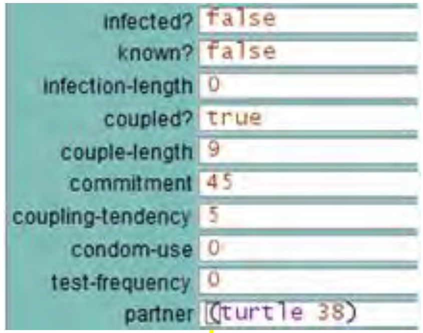
Description visualization of the phenomenon functioning itself

Let’s see some basic concepts and guiding questions

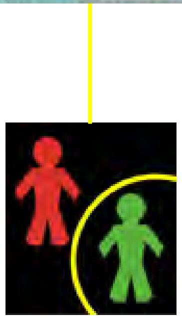
Agents – Who are the actors involved in the phenomenon?

Agents: a virtual object capable of elaborating information and able to execute an action (individuals, institutions, households...)

- **Intentionality**: acting based on goals or plans
- **Proactivity**: initiating actions rather than waiting
- **Reactivity**: responding to external stimuli or changes
- **Prosociality**: acting in coordination with others (social agents)



infected?	false
known?	false
infection-length	0
coupled?	true
couple-length	9
commitment	45
coupling-tendency	5
condom-use	0
test-frequency	0
partner	(turtle 38)



State variable: what characteristics cannot change through time? Ethnicity, gender

Dynamic variable: what characteristics can change through time? Opinions, preferences

Global variable: shared by all agents

Local variable: shared by specific agents or class of agents

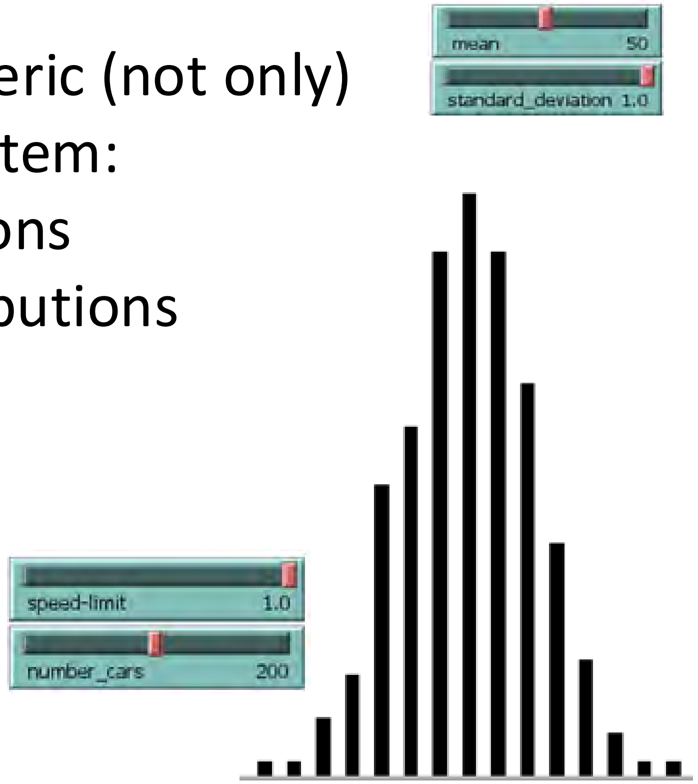
Heterogeneous vs **Homogeneous** (attributes distribution)

Attributes, Beliefs, Desire, Intentions (A+BDI)

Parameters – What are the conditions affecting the phenomenon?

Parameters: tunable numeric (not only) variables to modify the system:

- calibrate initial conditions
- agents' variables distributions



mean 50

standard_deviation 1.0

speed-limit 1.0

number_cars 200

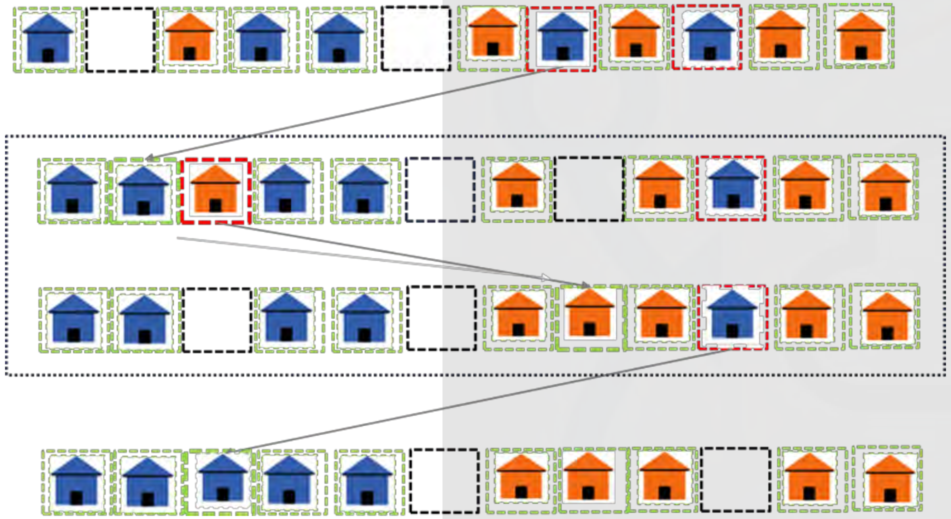
Global parameter: variable affecting all agents, and every agent can interact with
Belief shared by all agents

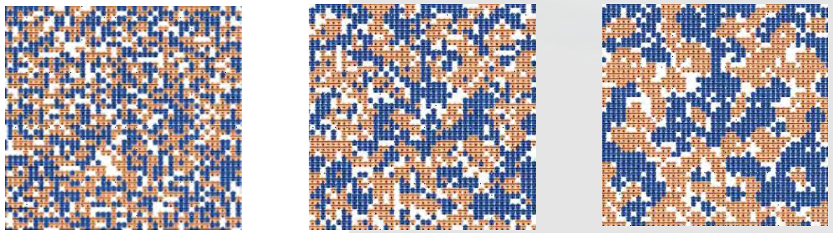
Local variable: accessible only to some specific agents
Norm specific to a class of agents

Evolution (I) – How the phenomen emerges through agents' interaction?

It is not much **time** as a continuous variable, rather the evolution of the system along two interconnected concepts:

- **micro level: schedule** of activation of agents' behavior
- **macro level transition phase** of the system changing due to mutual adaptation of the agents





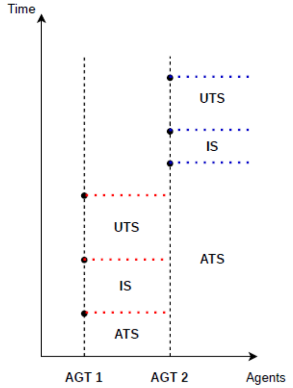
Cascade effect of the behavior of one agent on the neighborhood composition to other agents, affecting segregation at macro level

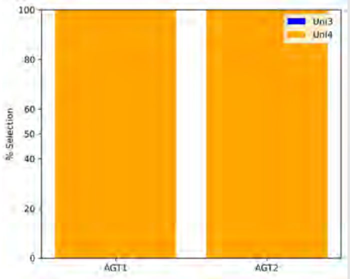
Evolution (II) – How does the order of agents’ action influence each other?

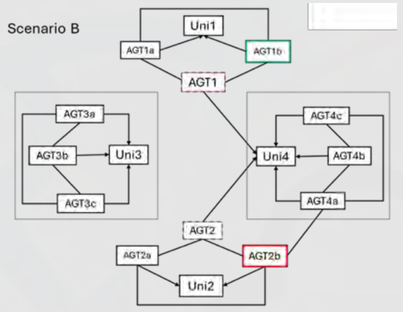
Parallelization: how the behavior is executed
Synchronous behavior: agents act together in parallel
Asynchronous behavior: agents act sequentially
(physical threads)

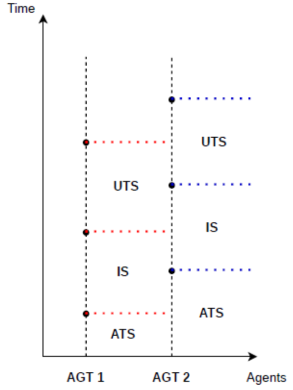
Synchrony: when the behavior is executed
Synchronization: agents decide based on the same knowledge of the world, including effects of actions of others
(they act «in parallel»)

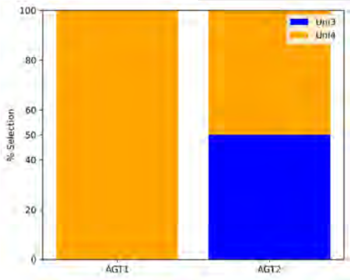
Example: scholar agents choose between two universities based on the chance to be introduced to elective authors based on potential shared connections. The sequential order of agents can affect the decision of those who select after

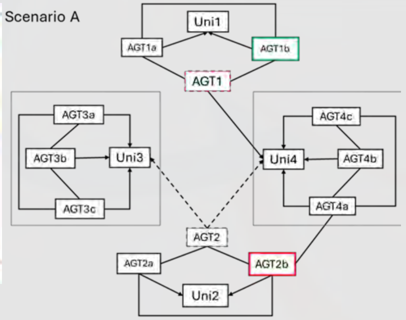










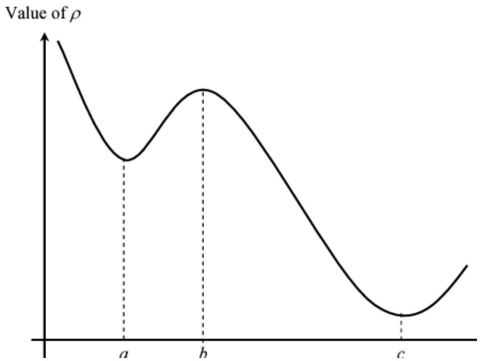


UTS > ATS: update shared knowledge

Longo, Paolillo, Ceriani (2026, forthcoming)

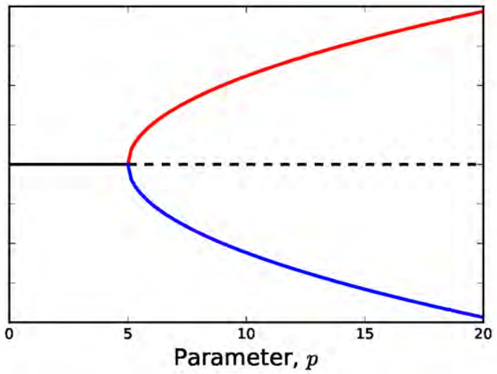
Outcome (I) – How can I read the evolution of the phenomenon?

Tipping points

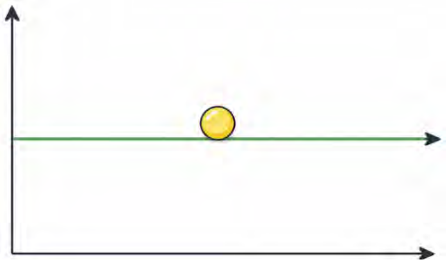


A sudden transition of the system is narrowed to one direction
e.g. in Schelling model the local level of segregation triggers relocation so that segregation becomes steady

Bifurcation

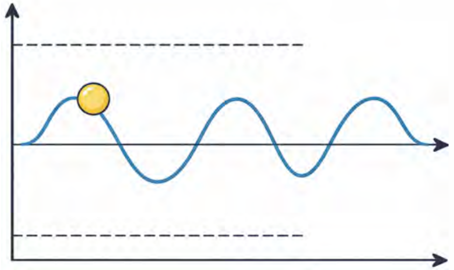


A moment where the phenomenon can diverge in two opposite directions with equal probability
e.g. in case of political polarization



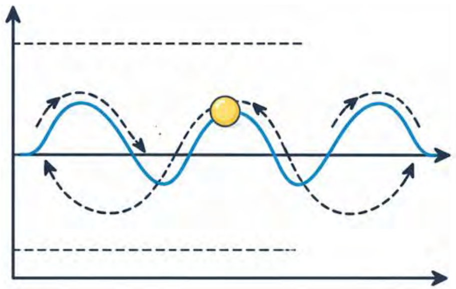
stable equilibrium

not mutable system
(Schelling, consensu)



cyclic equilibrium

system follows a trend of sequential cycles
(grass & sheeps, gentrification)

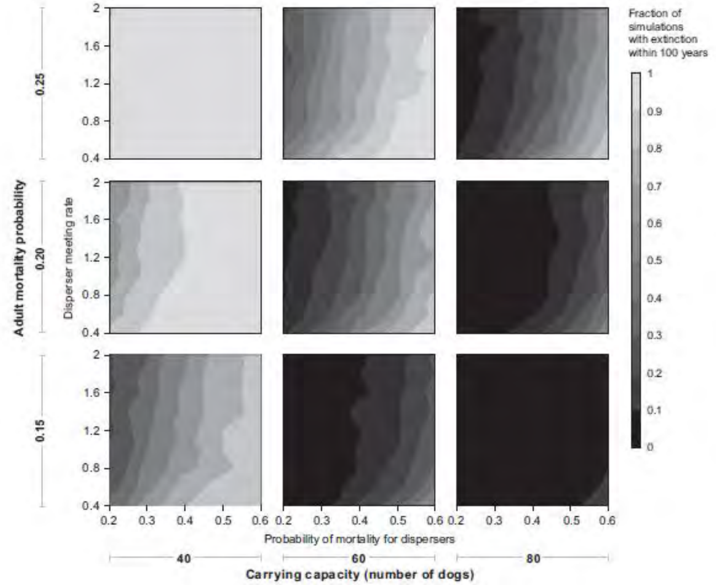


dynamic equilibrium

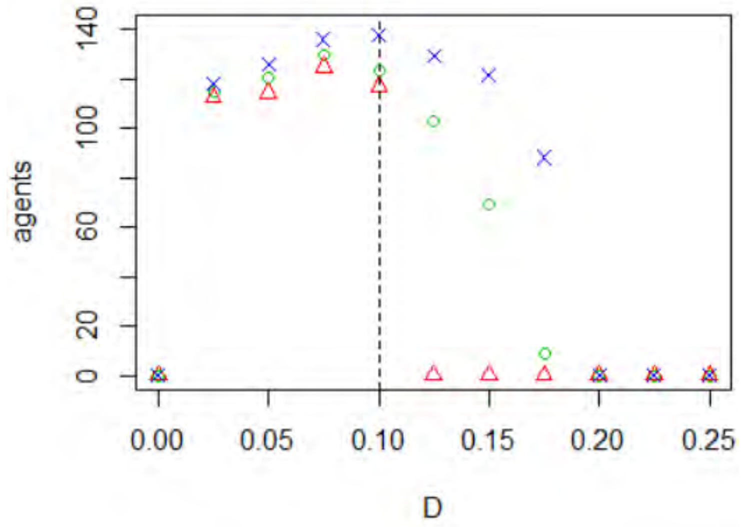
oscillations/inflows/outflows causing the system to apparently remain in balance
(e.g. supply & demand)

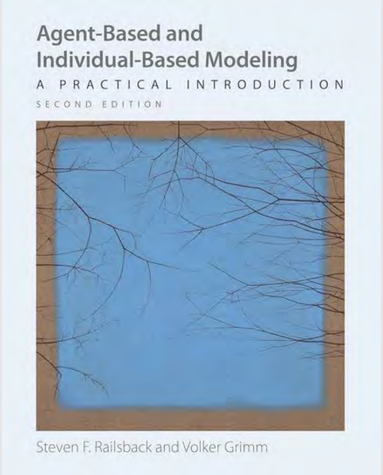
Outcome (II) – How can I identify how the conditions affect the phenomenon?

Global sensitivity analysis: pattern-oriented-modeling (POM) interaction between parameters (space of the model) to understand the overall mechanisms of the model



Local sensitivity analysis: one-factor-at-time (OFAT), focus on the effect of one specific parameter (nominal value) over the others





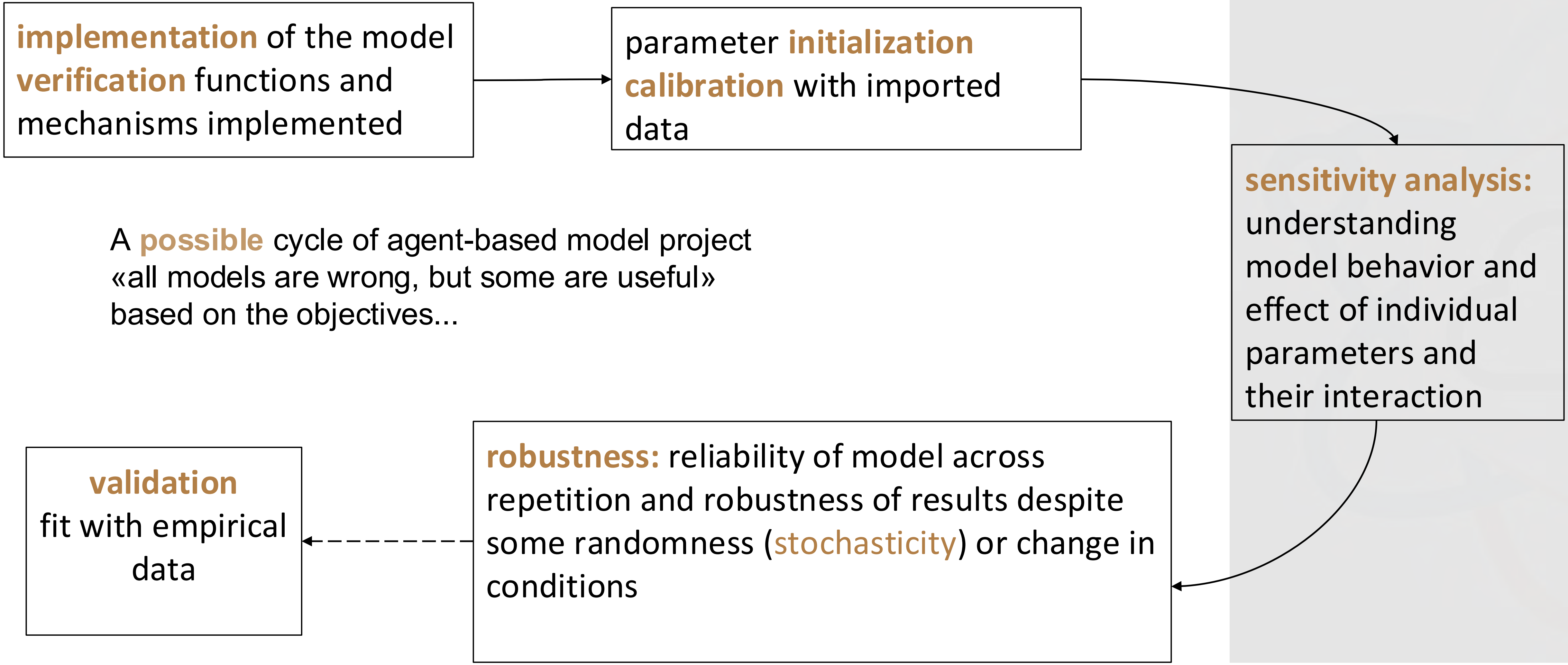
Agent-Based and Individual-Based Modeling

A PRACTICAL INTRODUCTION

SECOND EDITION

Steven F. Railsback and Volker Grimm

Compare what-if scenarios and measures



implementation of the model
verification functions and mechanisms implemented

parameter **initialization**
calibration with imported data

sensitivity analysis: understanding model behavior and effect of individual parameters and their interaction

robustness: reliability of model across repetition and robustness of results despite some randomness (stochasticity) or change in conditions

validation
fit with empirical data

A **possible** cycle of agent-based model project
«all models are wrong, but some are useful»
based on the objectives...

Parameters
What are the conditions affecting the phenomenon?

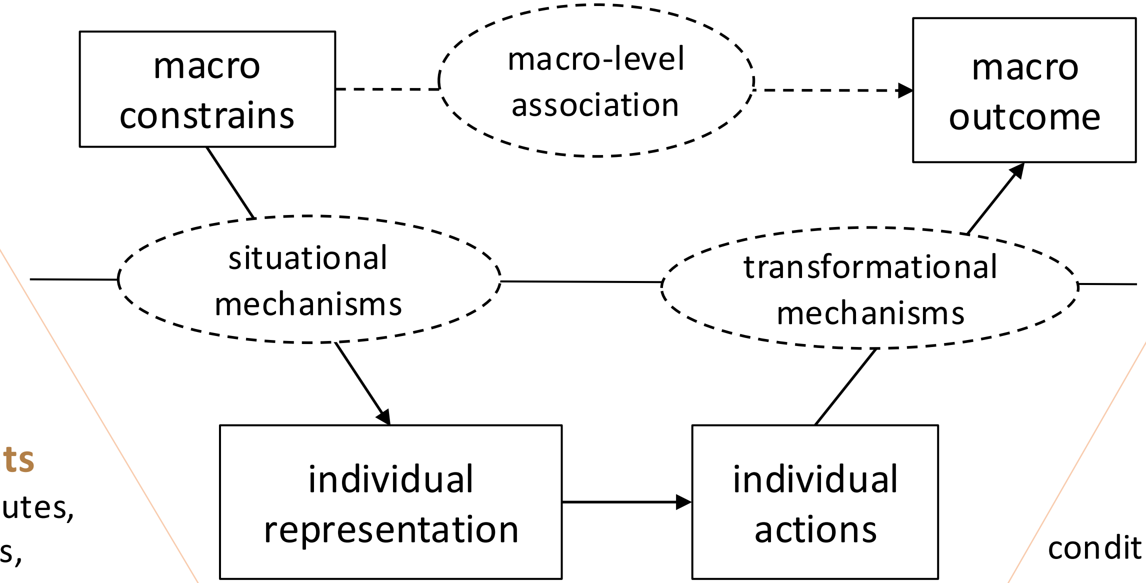
Design

Agents
Who are the actors involved in the phenomenon?

Formalization

Agents
Attributes,
Beliefs,
Desires,
Intentions

Code



macro constrains

macro-level association

macro outcome

situational mechanisms

transformational mechanisms

individual representation

individual actions

Evolution
How does the order of agents' action influence each other

Outcome
How can I read the evolution of the phenomenon?

Evolution
How the phenomenon emerges through agents' interaction?

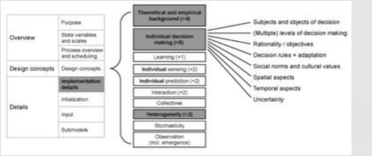
Outcome
How can I identify how the conditions affect the phenomenon?

What-if scenarios

Adapt your model to the grammars of a programming language

Let's implement to a case study and one programming tool (NetLogo)

ODD+D protocol to clarify ideas



Müller et al., 2013

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

References, suggested readings and tools

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. Annual review of sociology, 28(1), 143-166.

Edmonds, B. (2017). Different modelling purposes. Simulating social complexity: A handbook, 39-58.

Edmonds, B., & Moss, S. (2004, July). From KISS to KIDS–an 'anti-simplistic'modelling approach. In International workshop on multi-agent systems and agent-based simulation (pp. 130-144). Berlin, Heidelberg: Springer Berlin Heidelberg.

Epstein, J. M., & Axtell, R. (1996). Growing artificial societies: social science from the bottom up. Brookings Institution Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. Annual review of sociology, 36, 49-67.

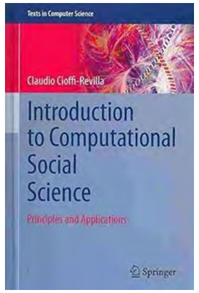
Macy, Michael W., and Andreas Flache. 2009. "Social Dynamics from the Bottom Up: Agent-based Models of Social Interaction." In Hedström, P. and Bearman, P. (Eds.) The Oxford Handbook of Analytical Sociology. Oxford, UK: Oxford University Press.

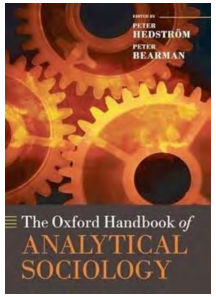
Schelling, T. C. (1969). Models of segregation. The American economic review, 59(2), 488-493.

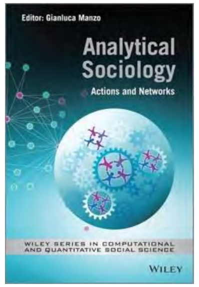
Schelling, T. C. (1971). Dynamic models of segregation. Journal of mathematical sociology, 1(2), 143-186.

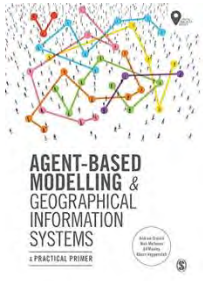
Longo, C.F., Paolillo, R., & Ceriani, M. (Forthcoming). T2B2T: The Ontology for adaptive Agent-driven seamless integration with the Semantic Web. In Proceedings of the Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), To appear on CEUR

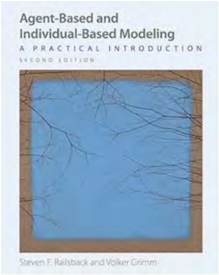
Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., ... & Schwarz, N. (2013). Describing human decisions in agent-based models–ODD+ D, an extension of the ODD protocol. Environmental Modelling & Software, 48, 37-48.


Texts in Computer Science
Claudio Cioffi-Revilla
Introduction to Computational Social Science
Principles and Applications
Springer


EDITED BY PETER HEDSTRÖM PETER BEARMAN
The Oxford Handbook of ANALYTICAL SOCIOLOGY


Editor: Gianluca Manzo
Analytical Sociology
Actions and Networks
WILEY SERIES IN COMPUTATIONAL AND QUANTITATIVE SOCIAL SCIENCE
WILEY


AGENT-BASED MODELLING & GEOGRAPHICAL INFORMATION SYSTEMS
A PRACTICAL PRIMER


Agent-Based and Individual-Based Modeling
A PRACTICAL INTRODUCTION
SECOND EDITION
Steven F. Railsback and Volker Grimm

COMPLEXITY EXPLORER
SANTA FE INSTITUTE

JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

essa

https://essa.eu.org/

MAILING LIST

SIMSOC@jiscmail.ac.uk

li.u

Institute for Analytical Sociology
Linköping University

Softwares free

NetLogo
https://www.netlogo.org/

MESA python
https://mesa.readthedocs.io/latest/

GAMA Platform
https://gama-platform.org/

Thank you! Questions?

rocco.paolillo@cnr.it

@roccopaolillo.bsky.social

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NRRP Grant agreement n. MUR IR0000008

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Part Two

Parameters
What are the conditions affecting the phenomenon?

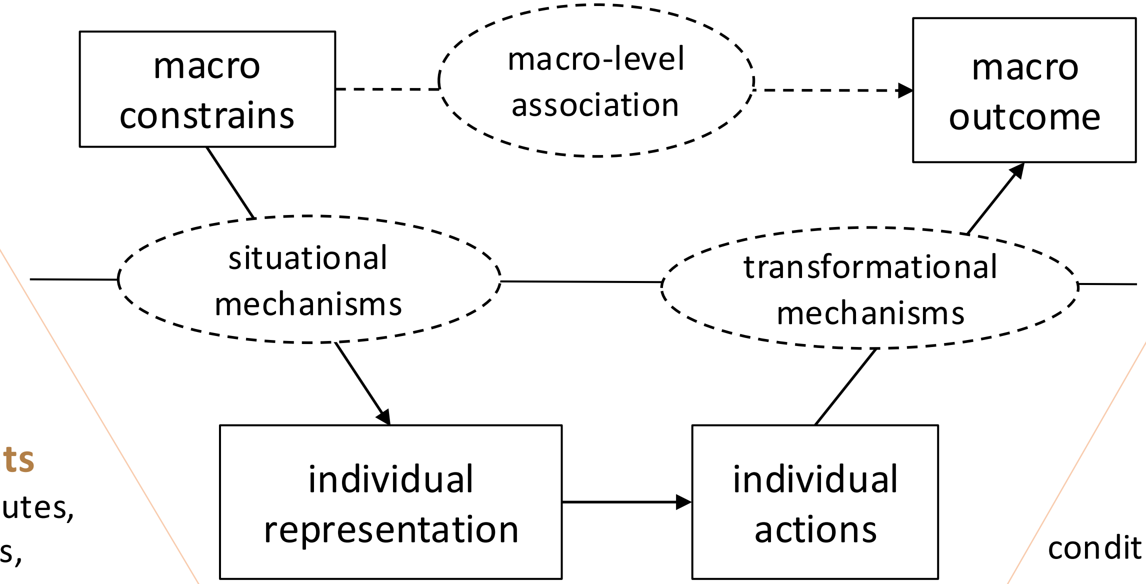
Design

Agents
Who are the actors involved in the phenomenon?

Formalization

Agents
Attributes,
Beliefs,
Desires,
Intentions

Code



Outcome
How can I read the evolution of the phenomenon?

Evolution
How the phenomenon emerges through agents' interaction?

Outcome
How can I identify how the conditions affect the phenomenon?

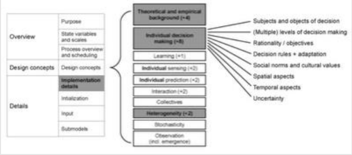
What-if scenarios

Evolution
How does the order of agents' action influence each other

Adapt your model to the grammars of a programming language

Let's implement to a case study and one programming tool (NetLogo)

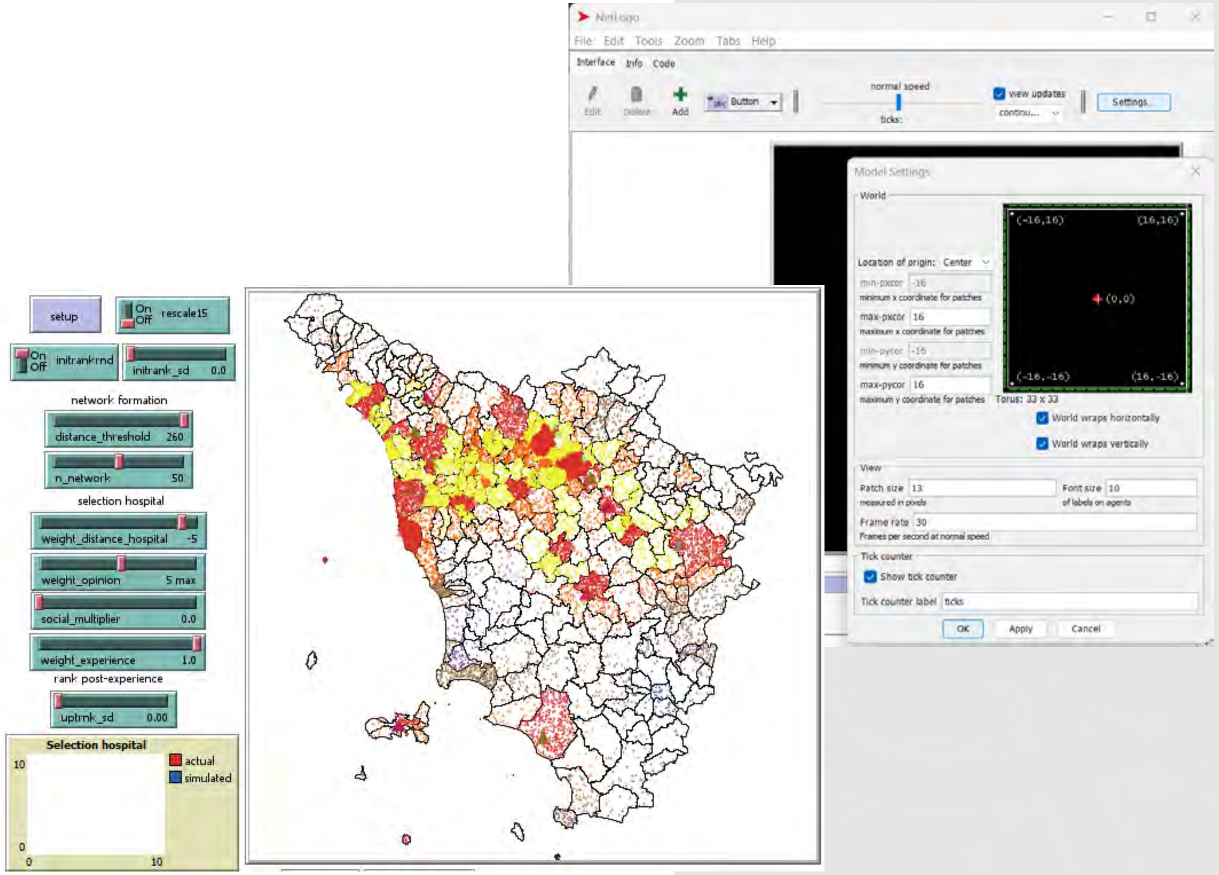
ODD+D protocol to clarify ideas


Overview
Purpose
State variables and scales
Process overview and scheduling
Design concepts
Design concepts
Details
Implementation details
Initialization
Input
Submodels
Theoretical and empirical background (+4)
Individual decision making (+8)
Learning (+1)
Individual sensing (+2)
Individual prediction (+2)
Interaction (+2)
Collectives
Heterogeneity (+2)
Stochasticity
Observation (incl. emergence)
Subjects and objects of decision
(Multiple) levels of decision making
Rationality / objectives
Decision rules + adaptation
Social norms and cultural values
Spatial aspects
Temporal aspects
Uncertainty

Müller et al., 2013

NetLogo

- Platform IDE (interfaccia) e programming language (java + starlogo) specific to agent-based modeling and experiments
- Open Source & User-friendly
- Allows many extensions (shapefile, csv import, random-wheel selection…)
- Widely used in the social science community and continuously maintained (7+ version)
- Programming language tailored to be as intuitive as possible and ready functions
- Supported by documentation (and Chat-GPT)
    - https://docs.netlogo.org/dictionary



On Off rescale15
On Off initrankrnd
initrank_sd 0.0

network formation
distance_threshold 260
n_network 50

selection hospital
weight_distance_hospital -5
weight_opinion 5 max
social_multiplier 0.0
weight_experience 1.0

rank post-experience
uptrnk_sd 0.00

🟥 actual
🟦 simulated

normal speed
ticks:
☑ view updates
continu...

Model Settings

World
Location of origin: Center
min-pxcor -16
minimum x coordinate for patches
max-pxcor 16
maximum x coordinate for patches
min-pycor -16
minimum y coordinate for patches
max-pycor 16
maximum y coordinate for patches
Torus: 33 x 33
☑ World wraps horizontally
☑ World wraps vertically

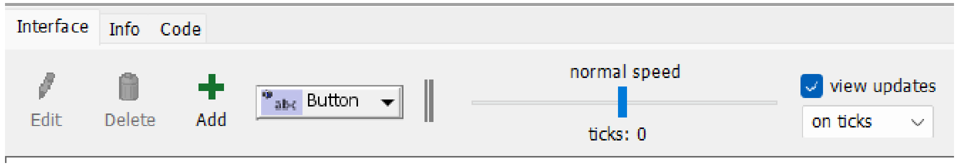
View
Patch size 13
measured in pixels
Font size 10
of labels on agents
Frame rate 30
Frames per second at normal speed

Tick counter
☑ Show tick counter
Tick counter label ticks

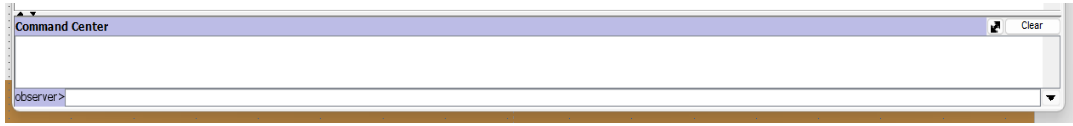
NetLogo
Simulate, Explore, Understand

https://www.netlogo.org/

NetLogo



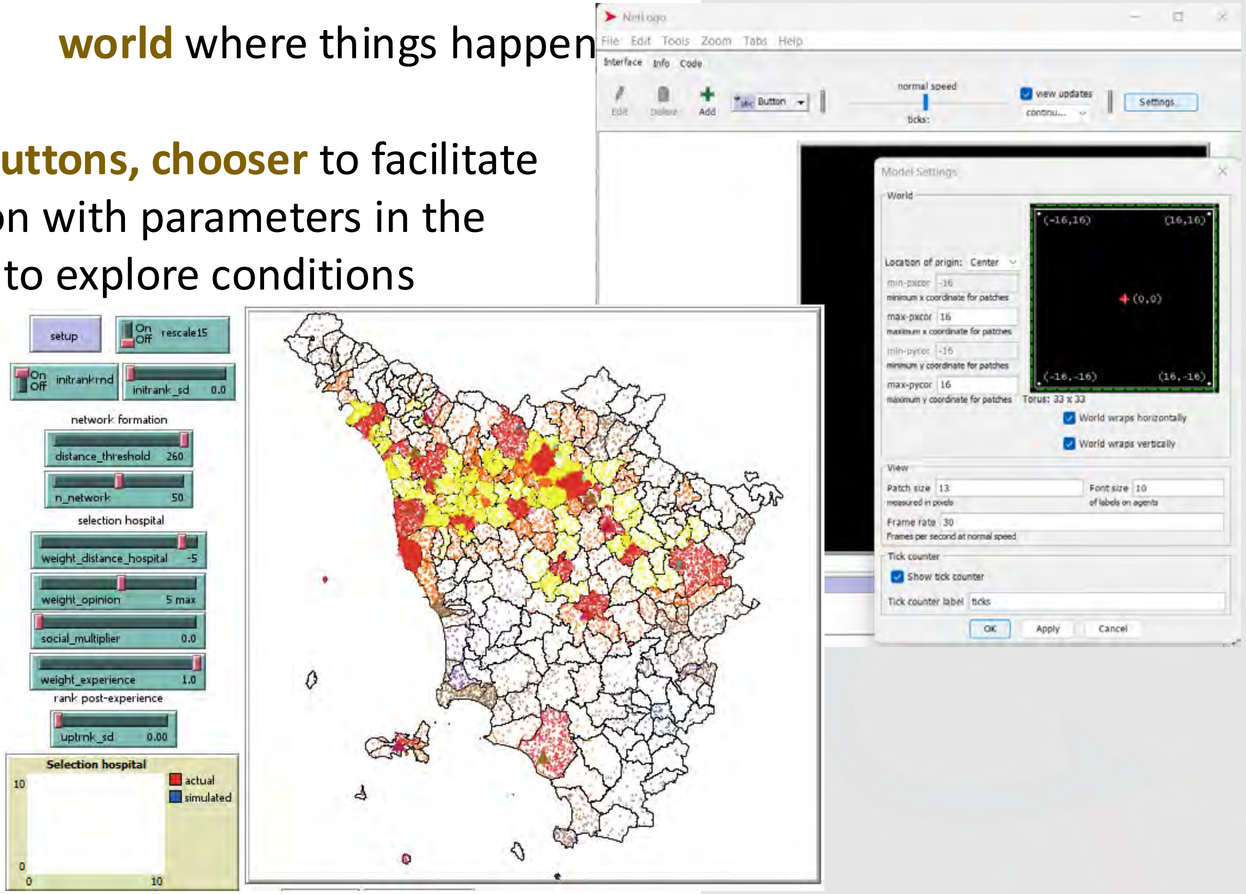
Interface tab to interact
Info tab tab to document
Code tab to build the model
(also in conjunction with interface)



A command line to interact *on the fly*

world where things happen

Sliders, buttons, chooser to facilitate interaction with parameters in the interface to explore conditions

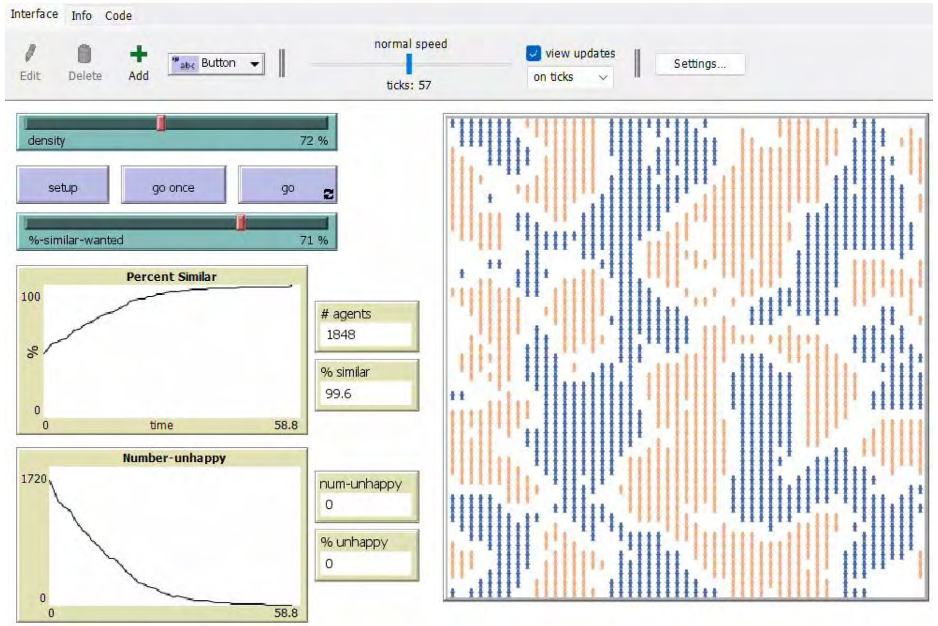


initrank_sd 0.0
distance_threshold 260
n_network 50
weight_distance_hospital -5
weight_opinion 5 max
social_multiplier 0.0
weight_experience 1.0
uptrnk_sd 0.00

Location of origin: Center
min-pxcor -16
minimum x coordinate for patches
max-pxcor 16
maximum x coordinate for patches
min-pycor -16
minimum y coordinate for patches
max-pycor 16
maximum y coordinate for patches
Torus: 33 x 33
☑ World wraps horizontally
☑ World wraps vertically
Patch size 13
measured in pixels
Font size 10
of labels on agents
Frame rate 30
Frames per second at normal speed
☑ Show tick counter
Tick counter label ticks

Plots, monitors to detect how the phenomenon is emerging

NetLogo



density 72 %

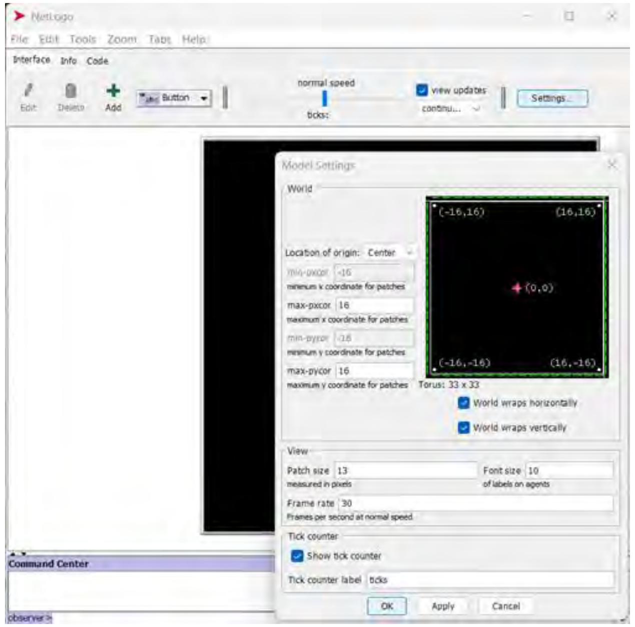
%-similar-wanted 71 %

agents 1848

% similar 99.6

num-unhappy 0

% unhappy 0



World

Location of origin: Center

min-pxcor -16
minimum x coordinate for patches

max-pxcor 16
maximum x coordinate for patches

min-pycor -16
minimum y coordinate for patches

max-pycor 16
maximum y coordinate for patches

Torus: 33 x 33

☑ World wraps horizontally

☑ World wraps vertically

View

Patch size 13
measured in pixels

Font size 10
of labels on agents

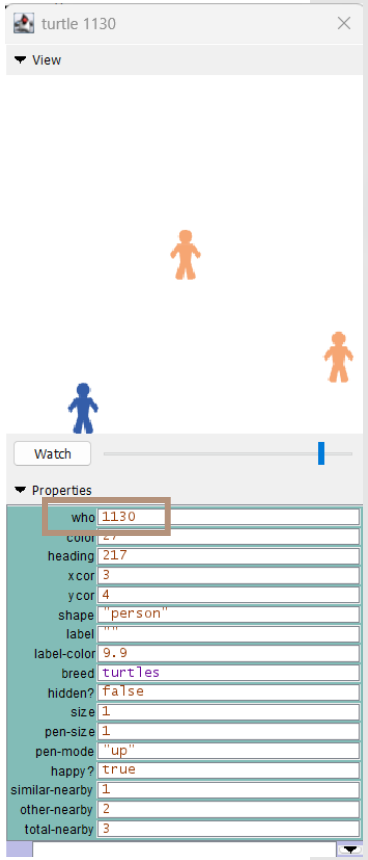
Frame rate 30
Frames per second at normal speed

Tick counter

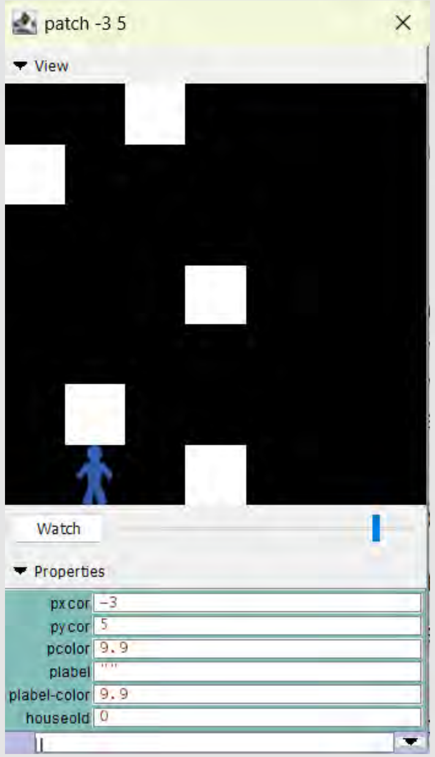
☑ Show tick counter

Tick counter label ticks

grid space world



who	1130
color	27
heading	217
xcor	3
ycor	4
shape	"person"
label	""
label-color	9.9
breed	turtles
hidden?	false
size	1
pen-size	1
pen-mode	"up"
happy?	true
similar-nearby	1
other-nearby	2
total-nearby	3



pxcor	-3
pycor	5
pcolor	9.9
plabel	""
plabel-color	9.9
houseold	0

0-indexed

`list [a b c]`

show item 1 (list 1 2 4) > 2

first agent to appear has who 0

Agents are called **turtles** identified by **who** ID

Grid cells are called **patches** and can interact as agents

NetLogo

simulated synchrony: agents execute commands asynchronously, but every agent acts knowing update in the model, based on coding

```
to ask dosomething
  ask turtles [do_A]
  ask turtles [do_B]
  ask turtles [do_C]
end
```

An agent in random order does A, then another agent does A. When all have done A, one random agent does B, then another does B. When all have done B, dosomething is executed

```
to ask dosomething
  ask turtles [
    do_A
    do_B]
end
```

An agent in random order does A then B, then another agent does A then B. When all have done A then B, dosomething is executed

```
to-report sumall [a b]
  report a + b
end
```

report 5 6 > 11

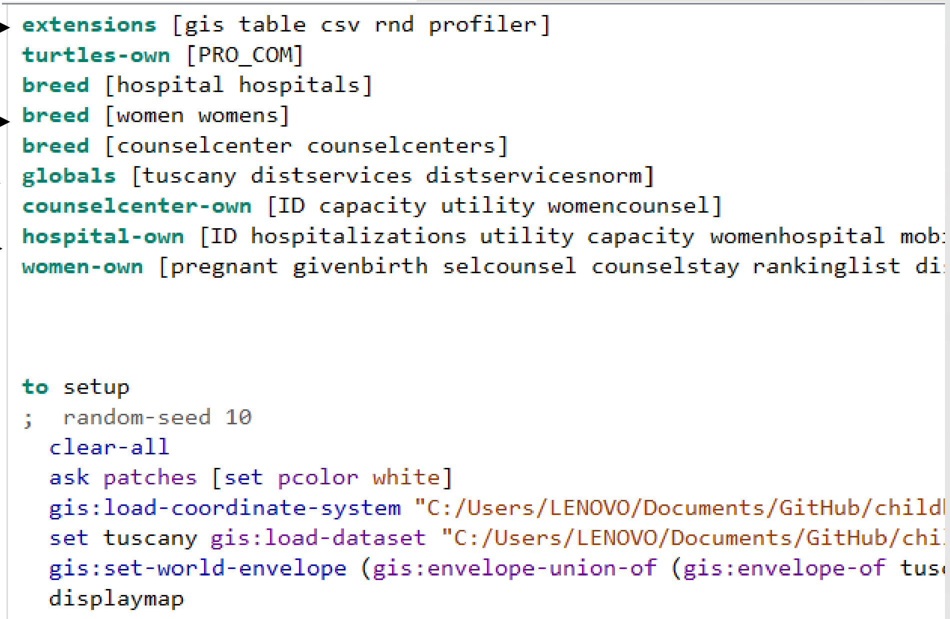
activate native extensions

agent-class (breed)

global variable

agent-class level variable

command block that translates model components to be run



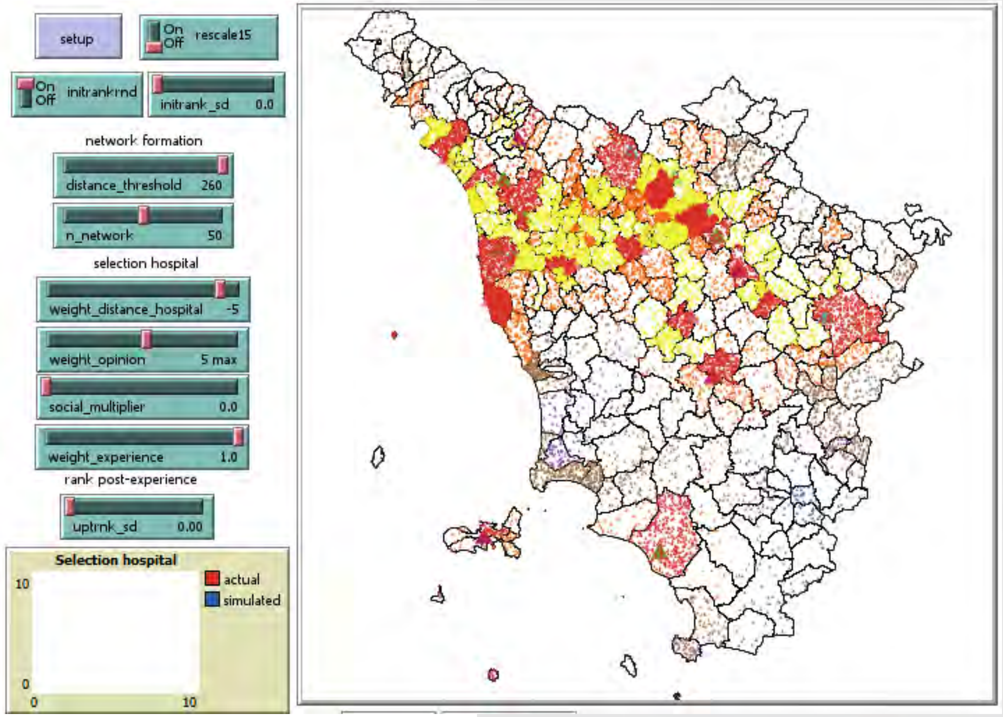
```
extensions [gis table csv rnd profiler]
turtles-own [PRO_COM]
breed [hospital hospitals]
breed [women womens]
breed [counselcenter counselcenters]
globals [tuscany distservices distservicesnorm]
counselcenter-own [ID capacity utility womencounsel]
hospital-own [ID hospitalizations utility capacity womenhospital mob
women-own [pregnant givenbirth selcounsel counselstay rankinglist di


to setup
;  random-seed 10
  clear-all
  ask patches [set pcolor white]
  gis:load-coordinate-system "C:/Users/LENOVO/Documents/GitHub/child
  set tuscany gis:load-dataset "C:/Users/LENOVO/Documents/GitHub/chi
  gis:set-world-envelope (gis:envelope-union-of (gis:envelope-of tus
  displaymap
```

`mean, count, sort`	primitive reporters
`ask, set, forward`	primitive commands
`let h who`	local variable existing within a command block (to alleviate memory)

Let's implement in a case study...

Childbirth Mobilities: a Geo-Spatial Simulation Approach

- **Context**: While some determinants of hospital maternity selection are identified in the literature, the individual decisional processes, and social influence processes underlying the choice are unknown, neither specific data available.
- **Why ABM**: we can model the weights of preferences for hospital attributes at agents' micro-level, compare different social influence processes and compare how they replicate the data
- **Data available**: Mobility patterns in Tuscany 2023:
  - municipality residencies of women who gave birth (aggregated and anonymous)
  - municipality hospital where they gave birth
  - ranking of hospital (PNE performance indicator)
  - matrix of ditances
  - shapefile to map geographies to data

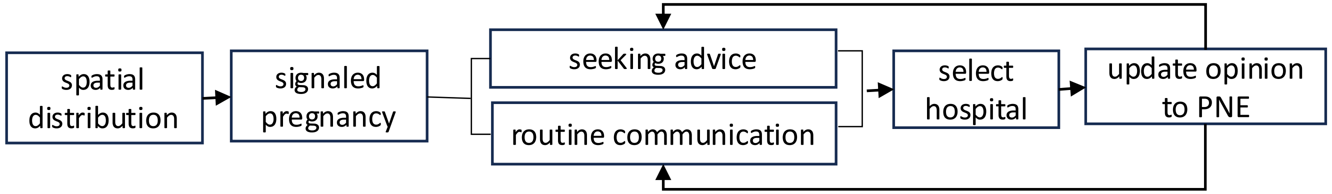


rescale15: Off
initrankrnd: On
initrank_sd 0.0
distance_threshold 260
n_network 50
weight_distance_hospital -5
weight_opinion 5 max
social_multiplier 0.0
weight_experience 1.0
uptrnk_sd 0.00

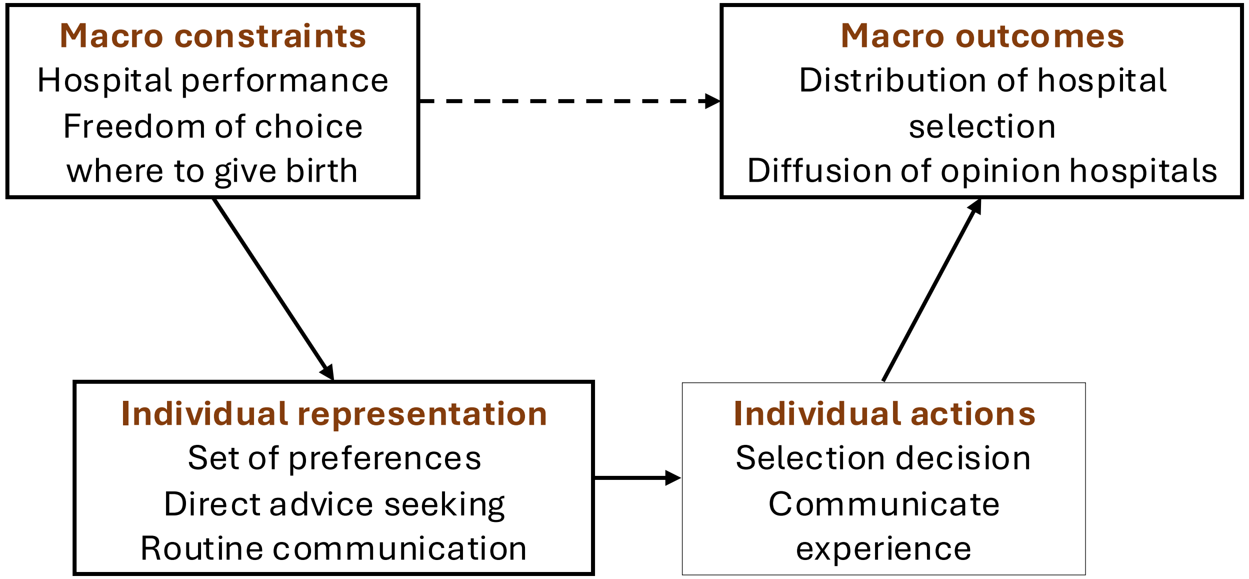
Goal

- Model a combination of individual decisional processes and social influence processes that can underline the selection of maternity hospital
- Which condition can best replicate the mobilities we observe?

Paolillo, Accordino, Pecoraro, https://github.com/RoccoPaolillo/childbirthnet/tree/MIE



spatial distribution → signaled pregnancy → seeking advice / routine communication → select hospital → update opinion to PNE

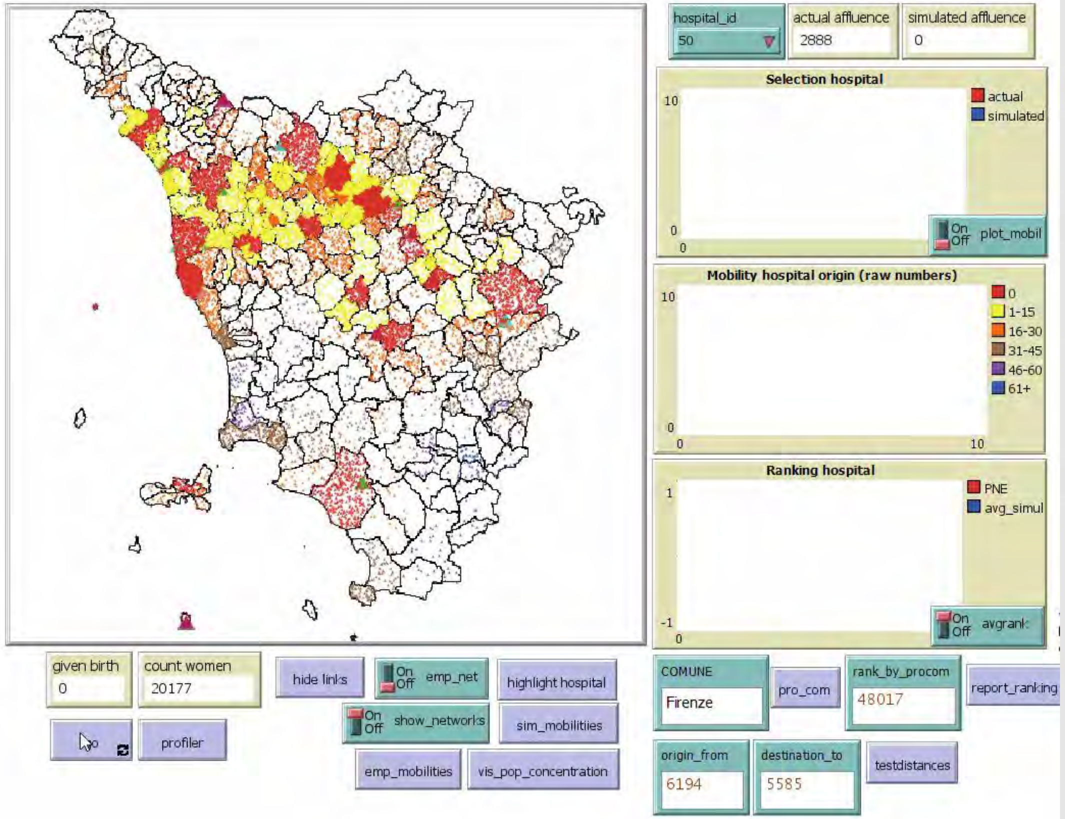


Macro constraints
Hospital performance
Freedom of choice where to give birth

Macro outcomes
Distribution of hospital selection
Diffusion of opinion hospitals

Individual representation
Set of preferences
Direct advice seeking
Routine communication

Individual actions
Selection decision
Communicate experience



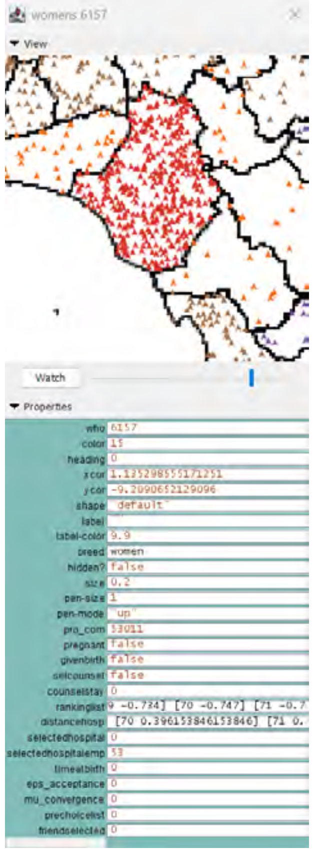
hospital_id 50
actual affluence 2888
simulated affluence 0
given birth 0
count women 20177
COMUNE Firenze
rank_by_procom 48017
origin_from 6194
destination_to 5585

Let's compare with concepts presented in part one...

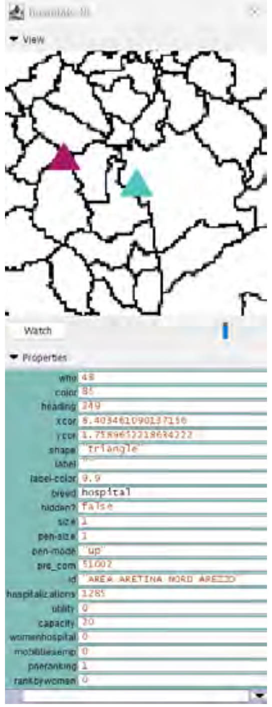
Agents – Who are the actors involved in the phenomenon?

women:

- they hold an initial random distribution of ranking opinion for each hospital in the region, when they become pregnant activate for choosing one hospital.
- they can expressely ask advice to friends in their municipality or base on common opinion of hospitals from routine communication
- after selecting one hospital, they can vehiculate the opinion of actual performance of hospital (PNE)



```
               who 6157
             color 15
           heading 0
              xcor 1.135298555171251
              ycor -9.2090652129096
             shape "default"
             label ""
       label-color 9.9
             breed women
           hidden? false
              size 0.2
          pen-size 1
          pen-mode "up"
           pro_com 53011
          pregnant false
        givenbirth false
        selcounsel false
       counselstay 0
       rankinglist 9 -0.734] [70 -0.747] [71 -0.7
      distancehosp  [70 0.396153846153846] [71 0.
  selectedhospital 0
selectedhospitalemp 53
       timeatbirth 0
    eps_acceptance 0
    mu_convergence 0
     precholcelist 0
    friendselected 0
```



```
            who 48
          color 85
        heading 249
           xcor 8.403461090137156
           ycor 1.7589612218684222
          shape "triangle"
          label ""
    label-color 9.9
          breed hospital
        hidden? false
           size 1
       pen-size 1
       pen-mode "up"
        pre_com 51002
             id "AREA ARETINA NORD AREZZO"
hospitalizations 1285
        utility 0
       capacity 20
 womenhospital 0
  mobilitaemp 0
    pneranking 1
  rankbywomen 0
```

hospital:

- PNE empirical ranking

Parameters – What are the conditions affecting the phenomenon?

```
to setup

  clear-all
  ask patches [set pcolor white]
  gis:load-coordinate-system "C:/../comuni_consultori_2019.prj"
  set tuscany gis:load-dataset "C:/../comuni_consultori_2019.shp"
  gis:set-world-envelope (gis:envelope-union-of (gis:envelope-of tuscany))
  displaymap

  set distservices csv:from-file "C:/../matrice_distanze_consultori.csv"
  set distservicesnorm csv:from-file "C:/../normalized_distance.csv«

  create-counselcenters
  create-hospitals
  create-womens

  let sorted-hospitals sort-by [[a b] -> [hospitalizations] of a >
[hospitalizations] of b] hospital

ask women [options_hospital]
  plot-hospitals

  reset-timer
  reset-ticks
end
```

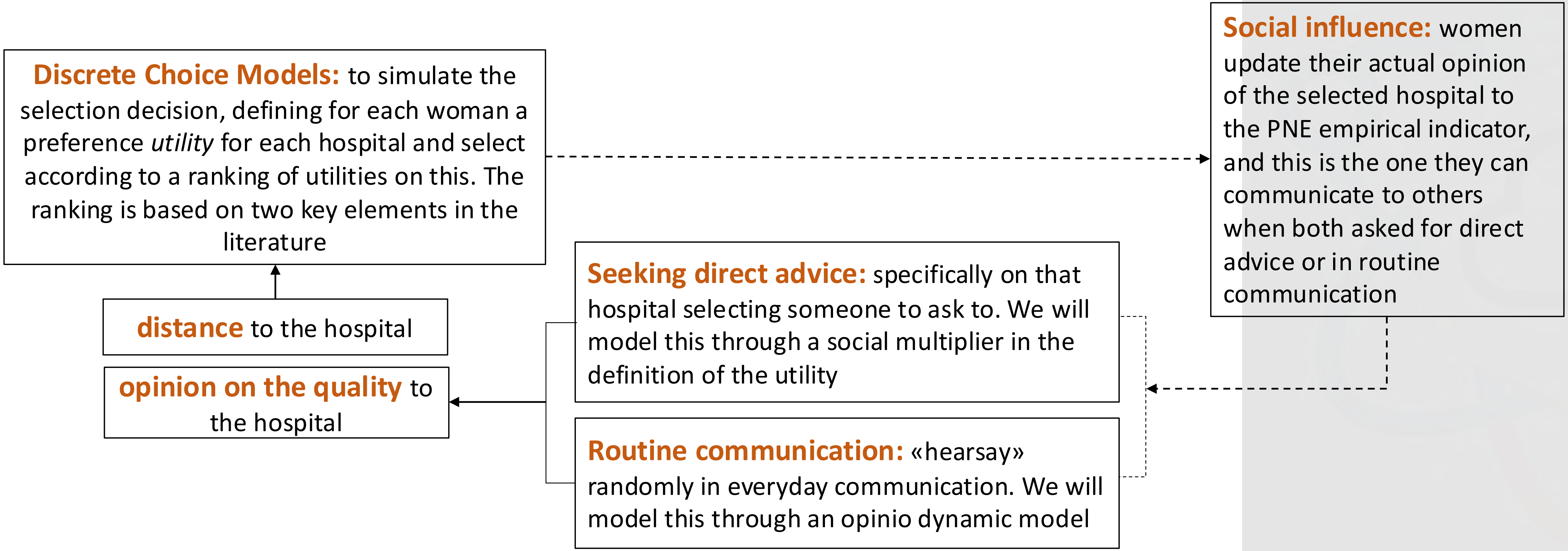
initialize the model with shapefiles

fetch data available

create agent classes

setup time schedule

Evolution - How does the order of agents' action influence each other?



Discrete Choice Models: to simulate the selection decision, defining for each woman a preference *utility* for each hospital and select according to a ranking of utilities on this. The ranking is based on two key elements in the literature

distance to the hospital

opinion on the quality to the hospital

Seeking direct advice: specifically on that hospital selecting someone to ask to. We will model this through a social multiplier in the definition of the utility

Routine communication: «hearsay» randomly in everyday communication. We will model this through an opinio dynamic model

Social influence: women update their actual opinion of the selected hospital to the PNE empirical indicator, and this is the one they can communicate to others when both asked for direct advice or in routine communication

Parameters – What are the conditions affecting the phenomenon?

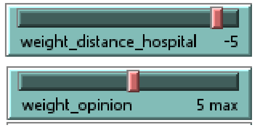
Discrete choice modeling: modeling the selection decision of agents, defining a utility (U) for each hospital h, based on a weight (parameter β) of how two characteristics of each hospital are relevant to the agent:
D_h: distance from the agent to the hospital
O_h: opinion on quality

The utility is used to define a probability to select that hospital *h* over the others hospital *k*

The higher is β, the more deterministic the selection is based on differences for attribute, the closer β is to 0, the more the selection is random ε

stable softmax to avoid numerical overflow

* 10 to harmonize ß of ranking and distance due to different scales


weight_distance_hospital -5
weight_opinion 5 max

We can input and manipulate the weight of each characteristic in the mind of agents

- equivalent to coefficients from regressions (clogit), not available
- test the consequences of combining different weights

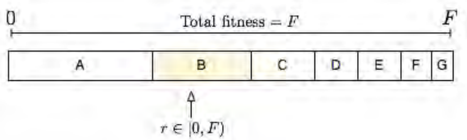
$$U_h = -\beta(D_h) + \beta(O_h) + \varepsilon$$

```
set utility ((weight_distance_hospital *
(distancefrom*10)) +  (weight_opinion * opinionquality)  )
```

$$P_h = \frac{e^{(Uh\ - \max(Uk))}}{\sum e^{(U_k - \max(U_k))}}$$

```
set selectedhospital [who] of rnd:weighted-one-of hospital
[exp(utility - max [utility] of hospital)]
```

rnd:weighted-one-of *agentset reporter*



Parameters – What are the conditions affecting the phenomenon?

Social multiplier: A weight$\theta[0,1]$ in the definition of opinion quality O_h at the moment of decision.

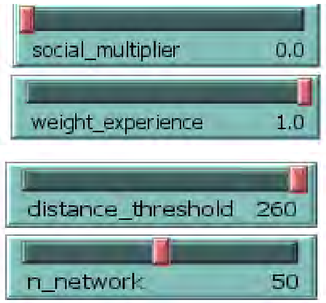
$\theta = 1$: the opinion quality completly aligns to that of people to whom asked for advice

θ =0: the advice of others is not taken into consideration

We also included a weighted average to allocate different weights to friend who gave birth to that hospital (*p*), and whose opinion *o* is based on actual experience, and those who speak for hearsay (*a*)

a = 1 – w

w = 1, only those who gave birth influence



social_multiplier 0.0

weight_experience 1.0

distance_threshold 260

n_network 50

We can manipulate how influenced people will be by those they seek advice to

We can manipulate how many people advice is searched for advice and how far from hometown

$$\left(\frac{o_w + o_a + o_w + \cdots}{w + a + w + \cdots}\right)$$

```
foreach sort friends   [ z ->
  let weightfriend ifelse-value ([selectedhospital] of z = [who] of self)
  [weight_experience][(1 - weight_experience)]
  set totweightfriend lput weightfriend totweightfriend
  set ranking_othweight lput (table:get [rankinglist] of z [who] of self * weightfriend)
ranking_othweight]
```

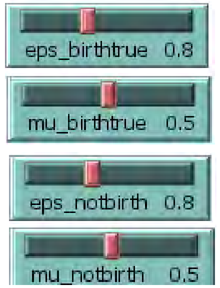
$$O_h = \ OwnOpinion_h + \theta\left(\left(\frac{o_w + o_a + o_w + \cdots}{w + a + w + \cdots}\right) - \ OwnOpinion_h\right)$$

```
set opinionquality [( opinionquality + social_multiplier *
((reduce + ranking_othweight / reduce + totweightfriend) - opinionquality ) )]
```

Parameters – What are the conditions affecting the phenomenon?

Opinion dynamics: a method to model the routine communication from women who gave birth to others in their municipality, spreading the own (updated) opinion a of that hospital. The receiver agent i accepts to listen if the distance between the own opinion of the hospital and that of the sender falls below a latitude of acceptance $|o_t^i - o_t^a| \leq \varepsilon$.
If so, the receiver aligns to the sender with convergence μ

ε = 0, not communication occurs
ε = 1, everyone is listened
μ = 0, not influence occurs
μ = 1, complete alignment occurs



eps_birthtrue 0.8
mu_birthtrue 0.5
eps_notbirth 0.8
mu_notbirth 0.5

We can manipulate how available to listen to those who gave birth and to what degree they will be influenced already in the *hearsay* routine communication

$$\begin{aligned} &if\ |o_t^i - o_t^a| \leq \ \varepsilon \\ &o_t^i = o_{t-1}^i + \mu(o_{t-1}^a - o_{t-1}^i) \end{aligned}$$

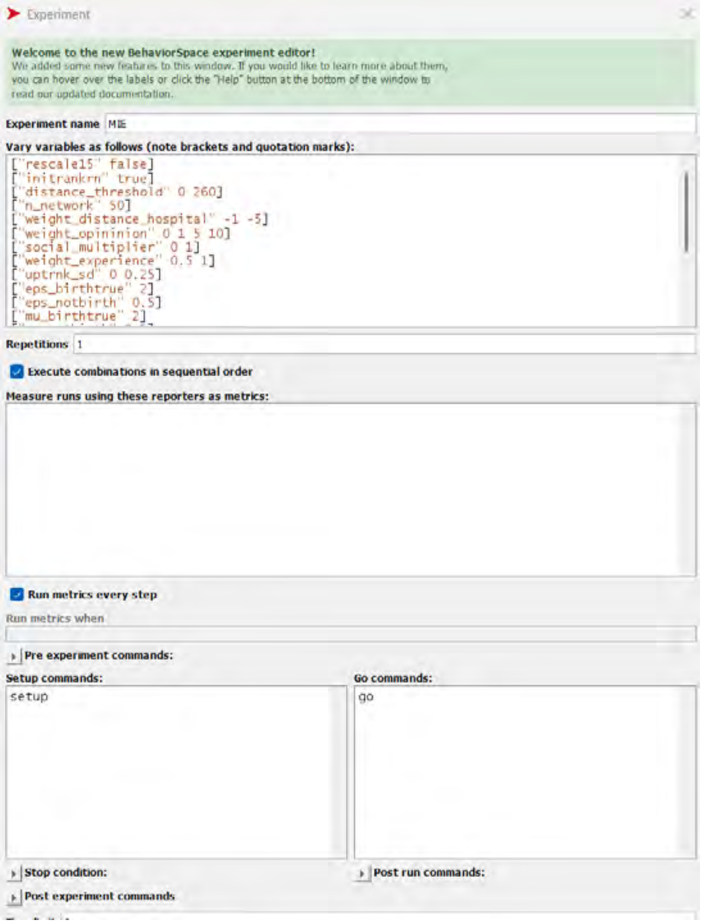
```
ask alter [
if abs(table:get rankinglist topic -
table:get [rankinglist] of myself topic) <= eps_acceptance
[table:put rankinglist topic
( table:get rankinglist topic +
(mu_convergence * (table:get [rankinglist] of myself topic -
table:get rankinglist topic)))]
```

We set to communicate every 80 time steps and to random 10% of women in municipality

Outcome – How can I identify how the conditions affect the phenomenon?

BehaviorSpace: a tool provided by NetLogo to set many experiments to run independently, setting the conditions for each parameter, define specific report measures, how many repetition wanted, and collect data in csv file

- Tools > BehaviorSpace
- Supports batch mode (headless)
- Better with a Server! For computational power, can run on laptop anyway



Experiment

Welcome to the new BehaviorSpace experiment editor!
We added some new features to this window. If you would like to learn more about them, you can hover over the labels or click the "Help" button at the bottom of the window to read our updated documentation.

Experiment name MIE

Vary variables as follows (note brackets and quotation marks):

```
["rescale15" false]
["initrankrn" true]
["distance_threshold" 0 260]
["n_network" 50]
["weight_distance_hospital" -1 -5]
["weight_opininion" 0 1 5 10]
["social_multiplier" 0 1]
["weight_experience" 0.5 1]
["uptrnk_sd" 0 0.25]
["eps_birthtrue" 2]
["eps_notbirth" 0.5]
["mu_birthtrue" 2]
```

Repetitions 1

☑ Execute combinations in sequential order

Measure runs using these reporters as metrics:

☑ Run metrics every step

Run metrics when

▸ Pre experiment commands:

Setup commands:

```
setup
```

Go commands:

```
go
```

▸ Stop condition:

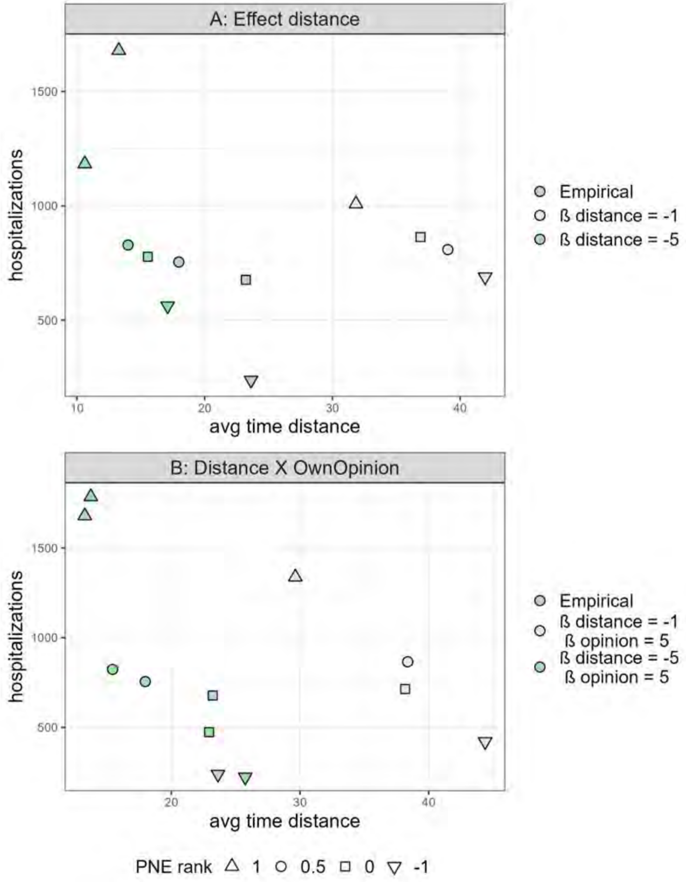
▸ Post run commands:

▸ Post experiment commands

['eps_birthtrue' 0 2] will run the conditions with the variable set 0 and 2

['eps_birthtrue' [0 0.1 2]] will run all the conditions with the variable set from 0 to 2 in increments 0.1 (e.g. 0 0.1 0.2...1.9 2)

Outcome – How can I identify how the conditions affect the phenomenon?

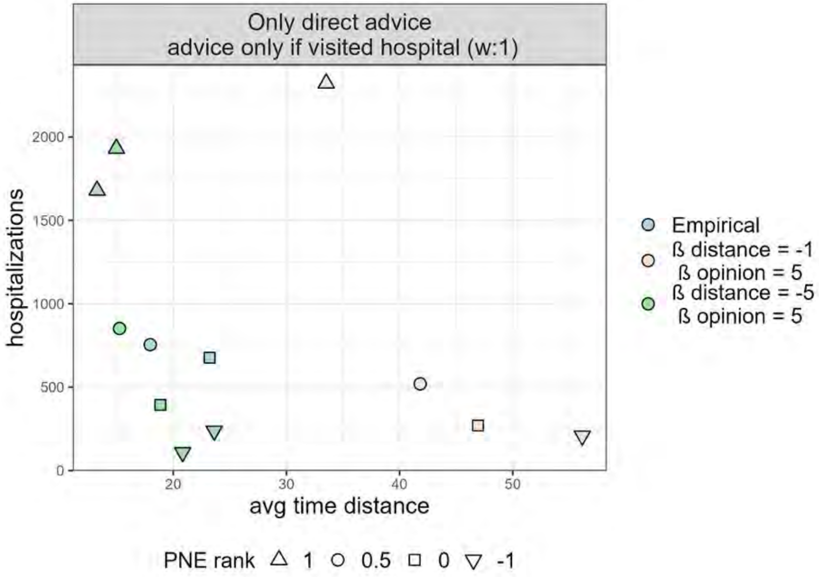


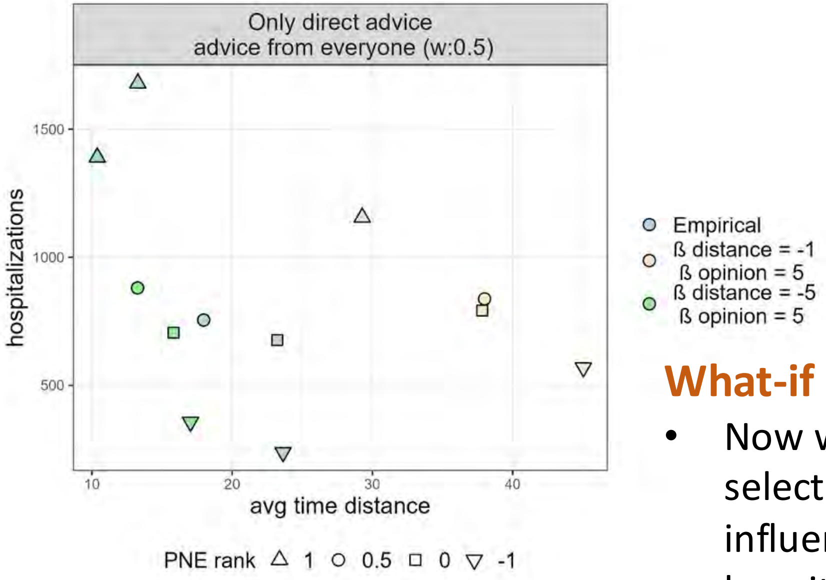
What-if scenarios

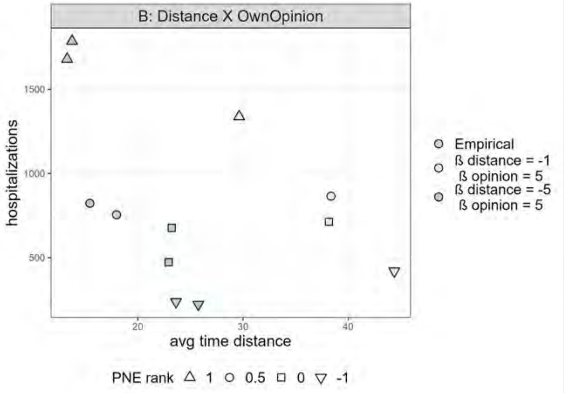
- Women communicate with everyone in routine communication (ε = 2), updating their opinion to the actual experience of those who gave birth, but they do not seek for advice (social multiplier θ = 0).
- We manipulate the weights for distance [0 -1,-5] and opinion quality updated via routine communication [0 1 5]
  - **condition A**: only distance matters (opinion weight 0)
    - with minimal weight of ditance (β = -1), women select hospitals more sparsely and difference between rankings do not emerge. Increasing weight distance (β = -5), the simulated distribution overestimates proximity of selected hospitals, and still sort by ranking not appearing
  - **condition B**: when we include also high opinion weight (β = 5), simulation results better approximate empirical data when coupled with high weight of distance (green condition β = -5, β = 5), both distance-wise and ranking-wise

What would be the effect of seeking advice then?

Outcome – How can I identify how the conditions affect the phenomenon?







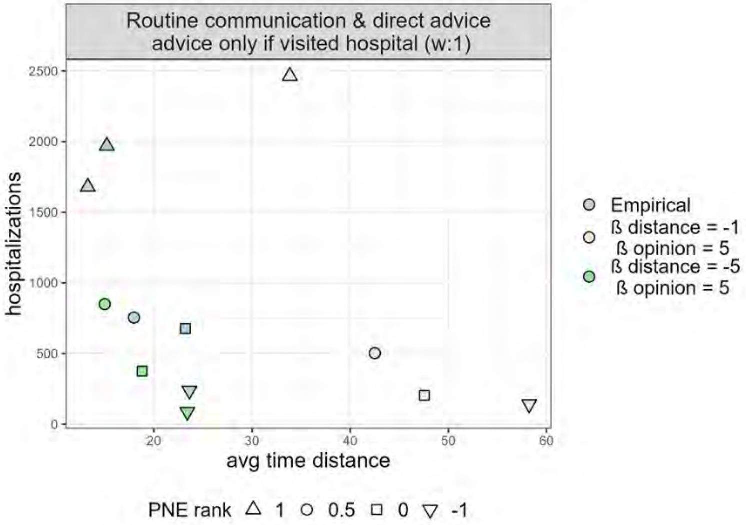
What-if scenarios

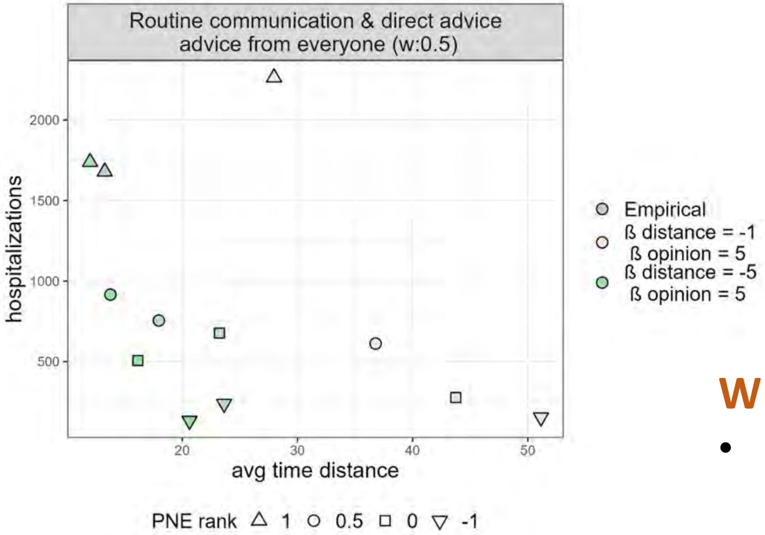
- Now women only seek for advice at the moment of selection (latitude opinion dynamic ε = 0), they can be influenced only by those who actually experience the hospital (*w* = 1) or by everyone equally (*w* = 0.5)

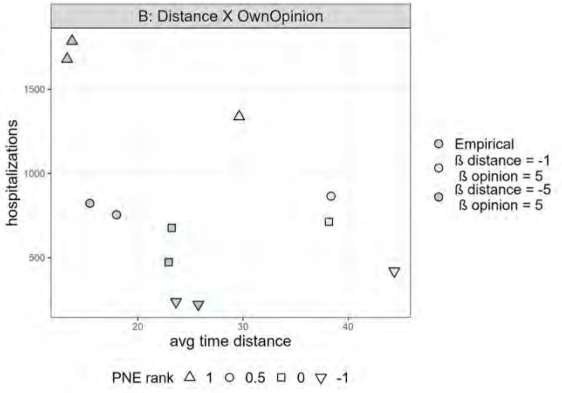
- Being influenced by those who gave birth (*w* = 1), hospitals with high PNE are overestimated, and more when the selection is more random by distance (β = -1)
- Being influenced with equal weight by everyone, those with actual experience and those with random opinion, underestimates the match with empirical data instead

What if we combine the two types of social influence?

Outcome – How can I identify how the conditions affect the phenomenon?







What-if scenarios

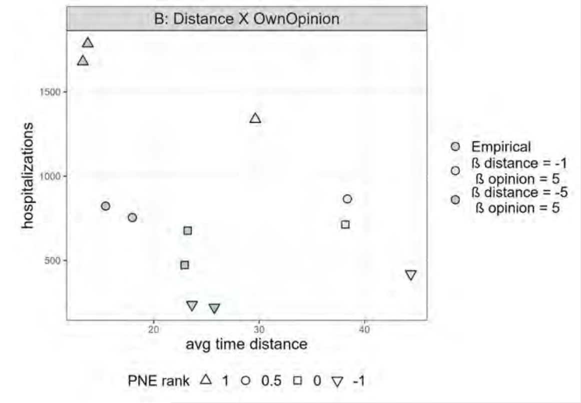
- Now women undergo both types of influence. When seeking advice, they can be influenced only by those who actually experience the hospital (*w* = 1) or by everyone equally (*w* = 0.5)

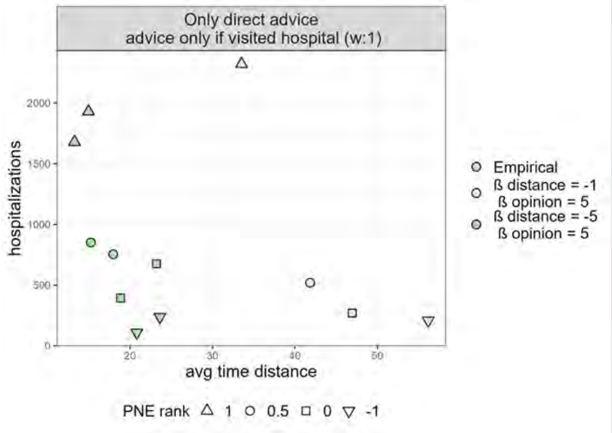
- Being influenced by those who gave birth (*w* = 1), still shows overestimation of hospitals with high, and more when the selection is more random by distance (β = -1)
- Being influenced with equal weight by everyone, those with actual experience and those who updated opinion by *hearsay* in common routine, better approximates the empirical data and reduces overestimation (sligthly)

Outcome – How can I read the evolution of the phenomenon?

Conclusions

- The best approximation to empirical mobilities is due to a combination of preference for shorter distance and high opinion quality. But high opinion quality with different weight of distance doesn’t produce the same effect. So, distance seems more relevant, and conditioning the diffusion of opinion updates
- Concerning the two modalities of social influence, seeking for advice would overestimate the effect of PNE ranking of hospitals, since the difference in opinion quality becomes more salient. The effect is higher if agents relocate randomly in space, probably because more likely to find high PNE hospitals, that are more and in more populated areas.
- Being exposed to different opinions when seeking advice and in combination with routine communication ameliorates the polarization effect of ranking coming closer to the empirical data, and routine communication seems to suffice





Outcome - How the phenomenon emerges through agents' interaction?

Limits and Next Steps

- To better understand the actual differentiation between seeking advice modality and routine communication, looking at the evolution through time and wider parameter space
- To differentiate action of women and measures considering the actual microspace within the region

BUT

- Overall, we had quite amount of data here
- Sometimes information on sociodemographic population might be missing, how could we do?
- Synthetic Populations

Let's see what synthetic populations are...

Thank you! Questions?

rocco.paolillo@cnr.it

@roccopaolillo.bsky.social

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union– NextGenerationEU under NRRP Grant agreement n. MUR IR0000008

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Part Three

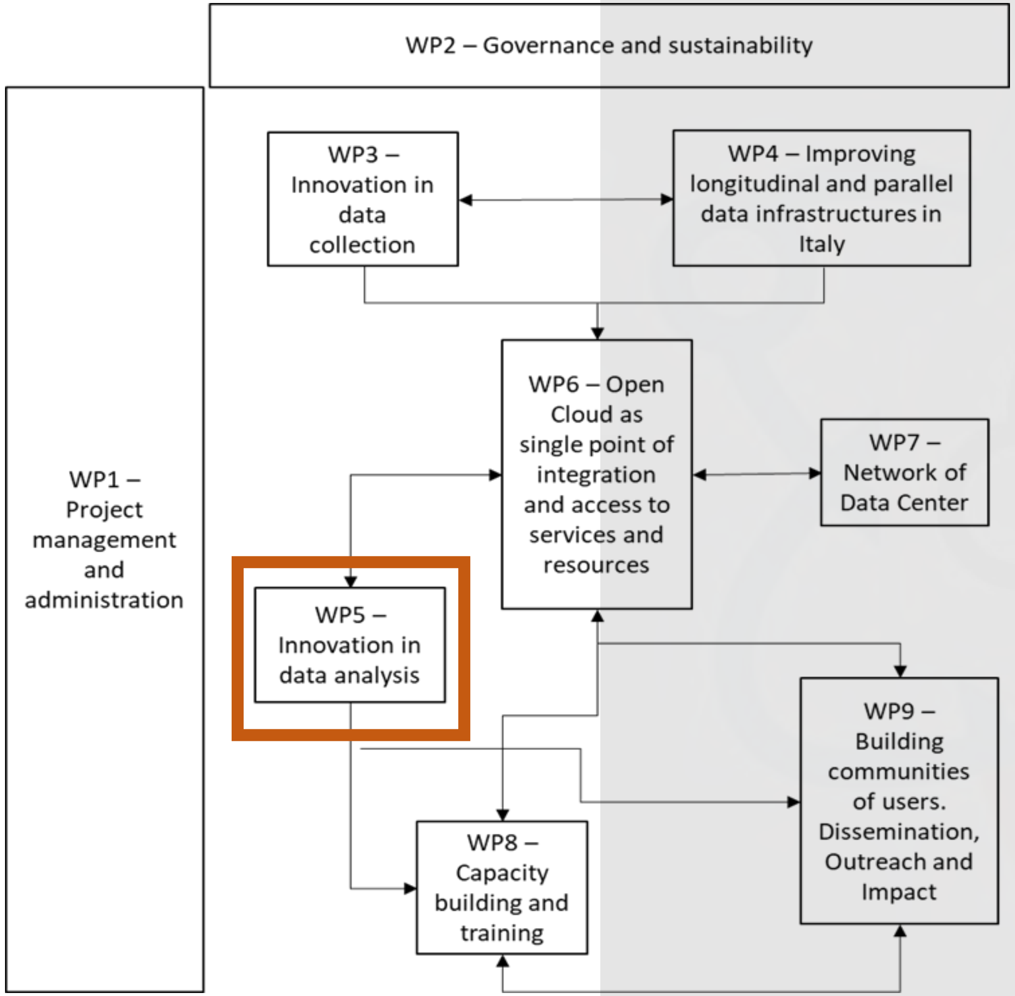
A service for synthetic populations extraction...

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Creation of an ***Italian Open Science Cloud for the Social Sciences*** guided by ***Open Science*** principles

which shall provide **innovative tools and services** to investigate issues related to the **economic and societal change of contemporary societies** through the enhancement of **research infrastructures**

https://www.fossr.eu/


WP2 – Governance and sustainability
WP1 – Project management and administration
WP3 – Innovation in data collection
WP4 – Improving longitudinal and parallel data infrastructures in Italy
WP6 – Open Cloud as single point of integration and access to services and resources
WP7 – Network of Data Center
WP5 – Innovation in data analysis
WP8 – Capacity building and training
WP9 – Building communities of users. Dissemination, Outreach and Impact

Luciana Taddei
Mario Paolucci *Editors*

Longitudinal Data Infrastructures in Europe

Tools for Open Science in Social Science Research

OPEN ACCESS

Springer

Chapter 12
Synthetic Populations in Research Infrastructures

Rocco Paolillo, Nicholas Roxburgh, Alice Sbrana, Gary Polhill, Evelina Carmen Sabatella, and Mario Paolucci

12.1 Collective Phenomena and Social Complexity

Many issues of interest to social and economic policies are complex phenomena, not derivable from the individuality of citizens. Phenomena such as opinion polarization or spatial segregation need a theoretical framework that embraces their complexity as an aggregated outcome of the collective dynamics of social actors and institutions interacting within contextualized spaces (Hedström & Bearman, 2009; Cioffi-Revilla, 2014). Following what is known as the metaphor of the society as a *common boat* by James S. Coleman (1994) (see Fig. 12.1), the methodological challenge to understand such phenomena is to address the transformative mechanisms that link the micro-level of individual actors, driven by their representations and motives, to the macro-level of the emerged observable phenomenon (Hedström & Ylikoski, 2011). Agent-based modeling is a methodology to this aim, building artificial societies in simulation scenarios used to study the emergence of collective phenomena in a dynamic and reproducible manner (Railsback & Grimm, 2019). The transformative mechanisms mentioned above are addressed through the interactive processes of virtual agents representing social actors and institutions, provided with dynamic and stable attributes to mimic their empirical counterparts (Macy & Willer, 2002; Grimm et al., 2006, 2010, 2020). Researchers can experiment how

R. Paolillo (✉) · M. Paolucci
Institute for Research on Population and Social Policies (CNR-IRPPS), Rome, Italy
e-mail: rocco.paolillo@cnr.it; mario.paolucci@cnr.it

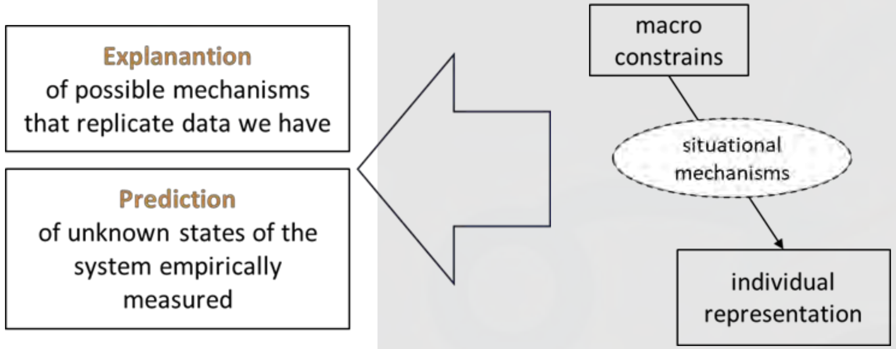
N. Roxburgh · G. Polhill
The James Hutton Institute, Aberdeen, Scotland, UK
e-mail: nick.roxburgh@hutton.ac.uk; gary.polhill@hutton.ac.uk

A. Sbrana · E. C. Sabatella
Institute for Research on Population and Social Policies (CNR-IRPPS), Fisciano, SA, Italy
e-mail: alice.sbrana@cnr.it; evelinacarmen.sabatella@cnr.it

© The Author(s) 2026 153
L. Taddei, M. Paolucci (eds.), *Longitudinal Data Infrastructures in Europe*,
https://doi.org/10.1007/978-3-032-07005-0_12

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

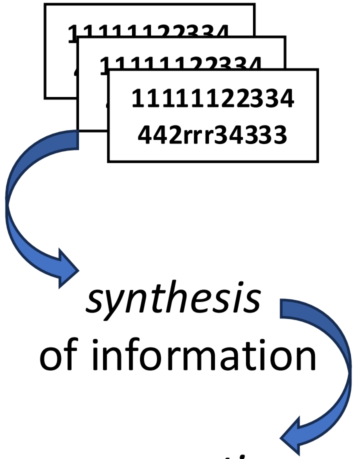
Ex-ante policy analysis evaluation
Post-ante policy analysis evaluation
Conterfactual policy scenarios


Explanantion
of possible mechanisms that replicate data we have
Prediction
of unknown states of the system empirically measured
macro constrains
situational mechanisms
individual representation

- An artificial society is a stylized social system where to study the mechanisms of the phenomenon
  - Especially when agent-based modeling is used for policy purposes, the mechanisms observed need to be bounded to the conditions of the system they want to operate in, e.g. Digital Twin systems
  - The system needs to be a synthesis of the information available of the target society

- Challenge to identify attributes at micro-level:
  - data not collected
  - separate datasets
  - privacy issues

Synthetic populations: a series of techniques to handle available data and replicate attributes of the target population

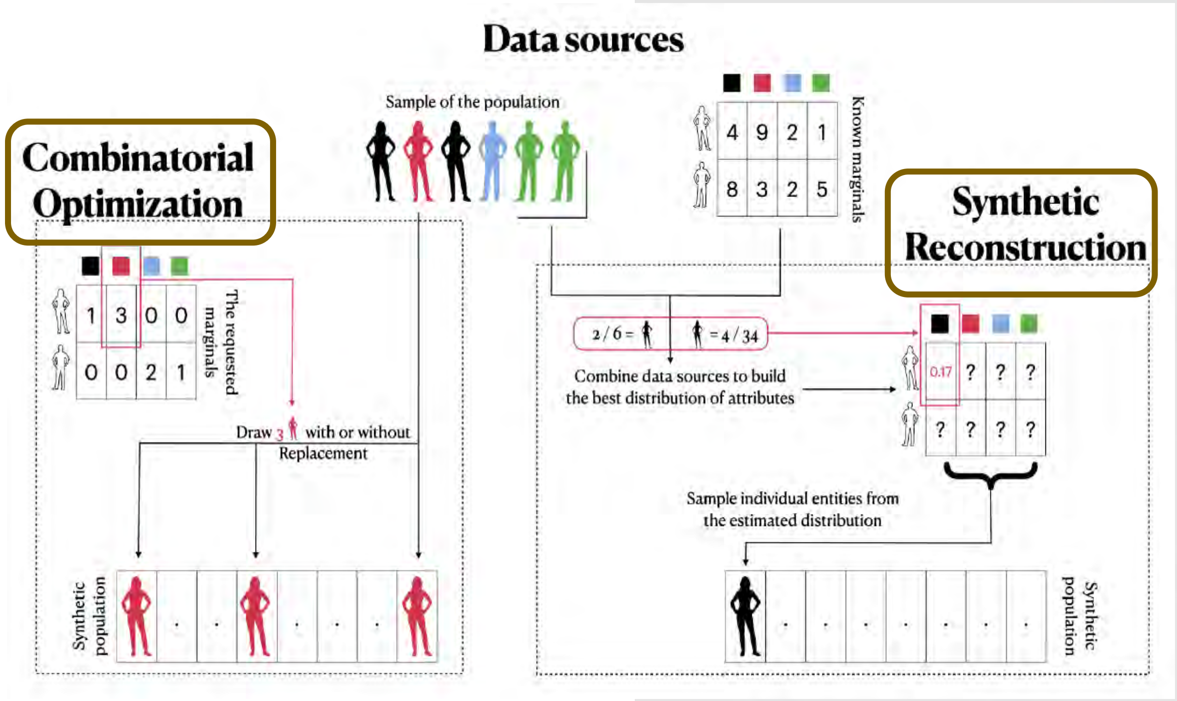


synthesis
of information

generation
of synthetic data

Bigi et al. (2022)

comparison
with available data



Chapuis et al. (2022)

Machine Learning

‘While a synthetic population is implicitly an artificial population, an artificial population is not necessarily a synthetic population’

Every method has its peculiarities and boundaries not set in stone

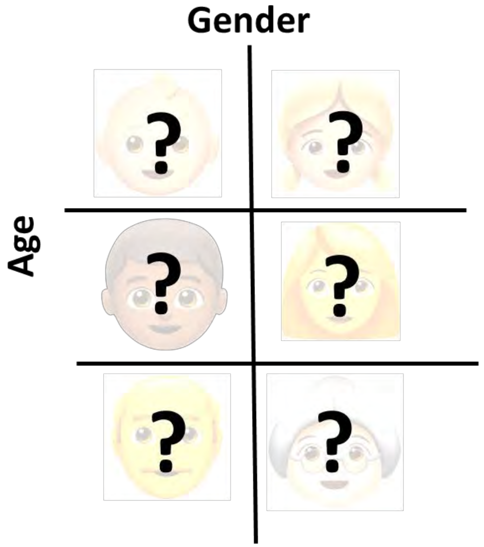
Synthetic Reconstruction

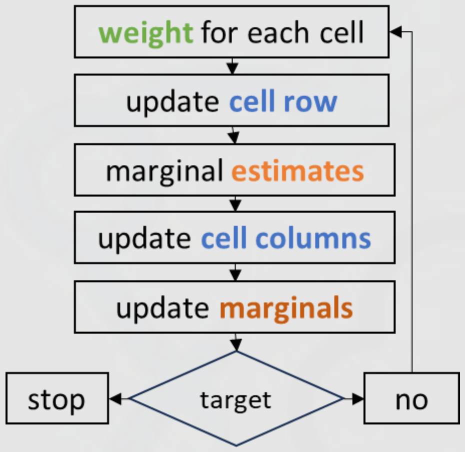
- Focus on estimate of unknown joint distributions from data available from marginal distribution
- macro to micro

Extensions

- **Multiple Iterative Proportional Fitting (MIPF)**
  taking the estimated joints as marginal to next step dimension
- **Hierarchical Iterative Proportional Fitting (HIPF)**
  nested data fitting marginals from one level (e.g. household) to the narrow (e.g. citizens) (Yamaego et al., 2021)
- **Iterative Proportional Updating (IPU)**: from micro data finds weight for cross-category multiplied by marginals and correct backwards -> combinatorial optmization

- Archetype in the **Iterative Proportional Fitting** (IPF, raking)





$$\text{weight} = \frac{\text{observed marginal}}{\text{fitted marginal}}$$

✓ mathematically transparent, robust, focus on weights

🟥 require ad hoc setting of algorithms, zero-cell problem

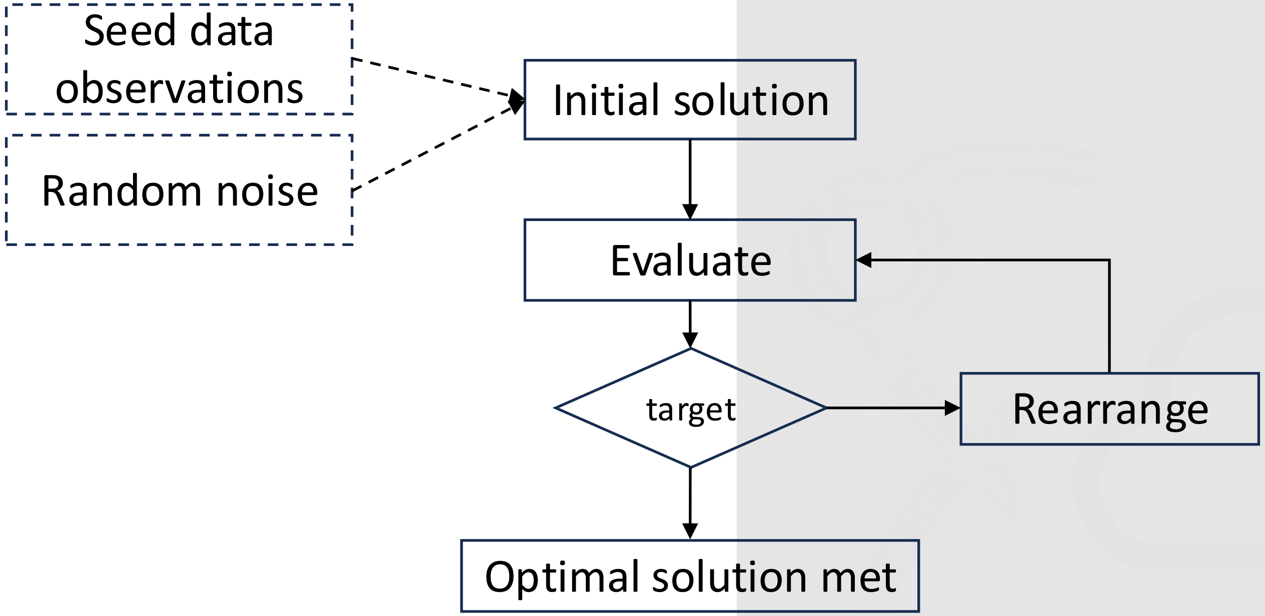
FOSSR:SPG

R::synthpop (MIPF)

Combinatorial Optimization

- Scaling of the synthetic population
- Draw random data and optimize against marginals observed

  - **Simulated Annealing (SA)**
    from the micro-data identifies some seed numbers and compares synthetic marginas to empirical marginals to correct backwards
  - **Markov Chain Monte Carlo (MCMC)**
    random number generations for all intersection and correct backwards



✓ overcome some setbacks of IPF family

✓ better with complex intersections (handled *at once*)

🟥 focus on the outcome rather than weights & inner conditions

🟥 computationally more demanding

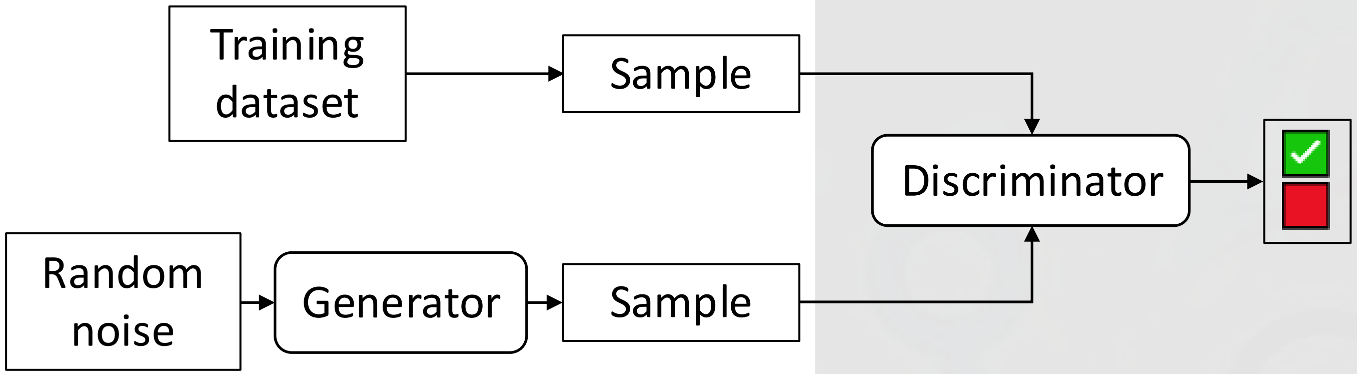
py::simmaneal

R::MCMCpack py::PyMC

Machine Learning Approach

- Most recent in time
- Statistical learning
- micro to macro

- **Generative Adversarial Networks (GAN)**
  - originated from images AI, applied to data
  - two competing (neural) networks:
    - **Generator** who produces random data
    - **Discriminator** that discriminate realistic data from not realistic
    - Goal of Generator is to get better to *deceive* the Discriminator who gets better in discriminate, meaning that synthetic data are very realistic



✅ promising because they integrate the performance of combinatorial methods with transparency, multidimensionally compared to IPF family

🟥 they require a training dataset from which the learning process occurs, where underrepresented groups are likely to be ignored in estimates

- Pre-adjust the training dataset with ad hoc weights to marginals (Falck, 2025)
- Post-adjust the synthetic outcome with weights to marginals

py::PyTorch

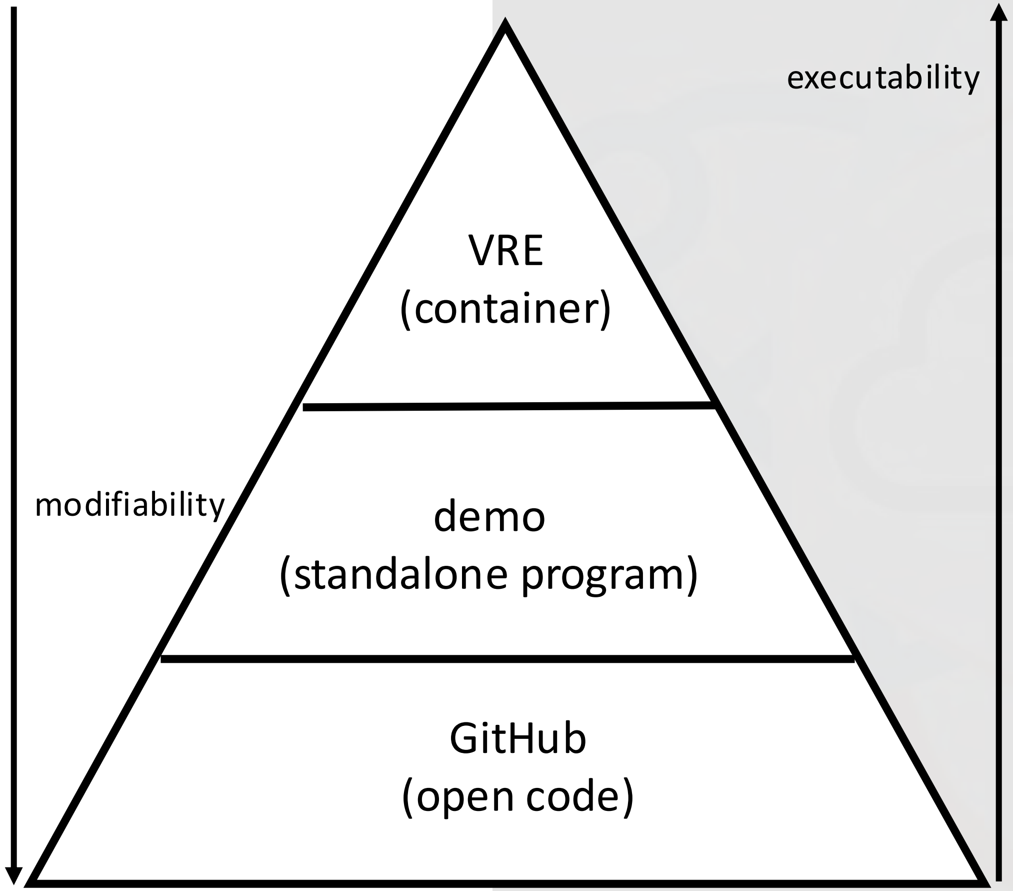
Synthetic Populations Generator (SPG)

Service to enable researchers and policy makers to extract synthetic populations at desired level of information from input dataset

- agent-based modeling
- spatial analysis
- conditional model
- ...

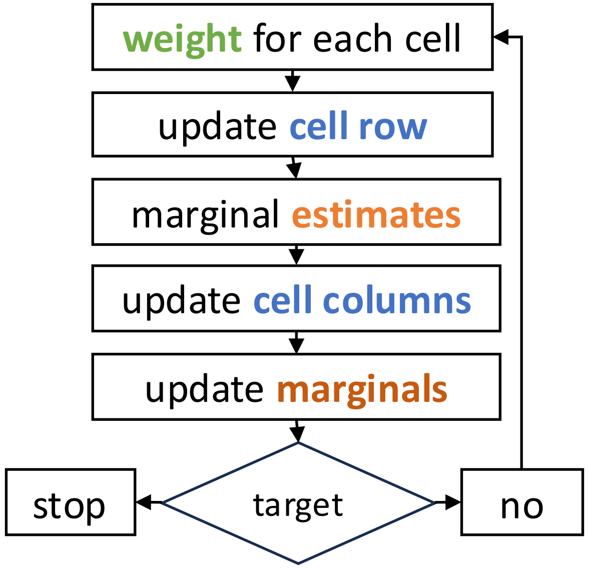
Open Science & Source Software

https://github.com/RoccoPaolillo/IPF_multidim.git > synthpopgen.py



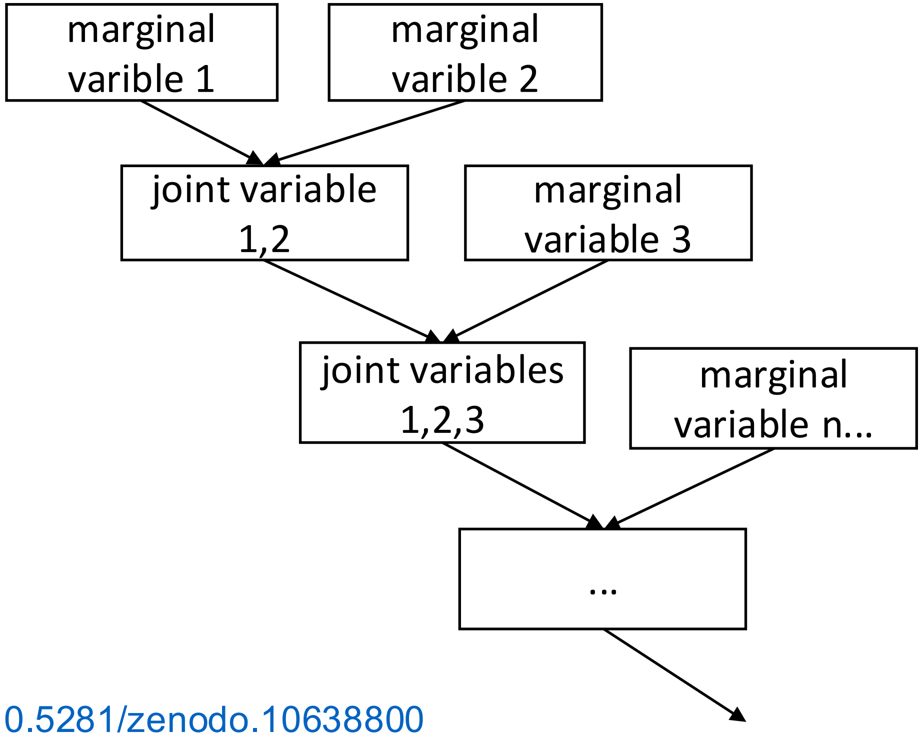
(Jimenez et al., 2017; Hong et al., 2022)

Iterative Proportional Fitting (IPF)



$$\text{weight} = \frac{\text{observed marginal}}{\text{fitted marginal}}$$

Multiple Iterative Proportional Fitting (MIPF)



Selected for higher transparency, robustness, light computing, core mechanism common to other method (somehow), but provided more service-oriented direction is guaranteed

First version released (ISTAT gender X age for validation) 10.5281/zenodo.10638800

Aims of Synthetic Populations Generator (SPG)

- Include **multidimensionality**
- Increase **generalizability** of variable handling
- Enable **automation** input-execution-output
- Customize **filtering** selection

- leverage estimate of joint and conditional probability over in-cell weight iteration

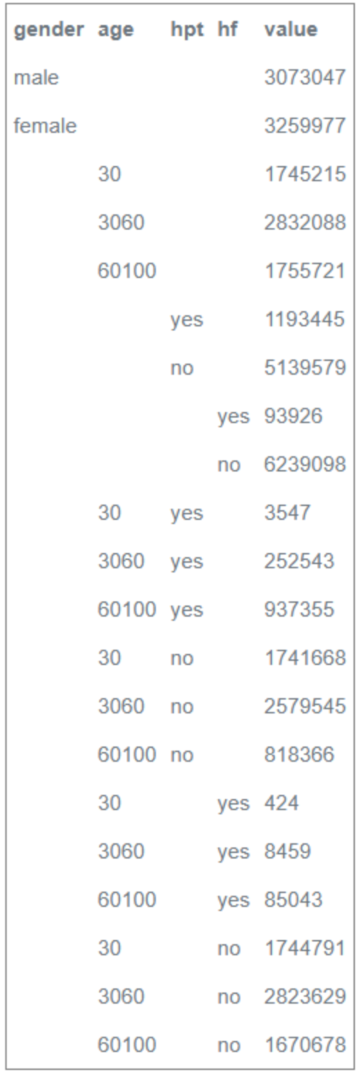
Tested with **opensalute Lazio** health data:

- gender
- age
- hyptertension (hpt)
- heart failure (hf)

known joints:

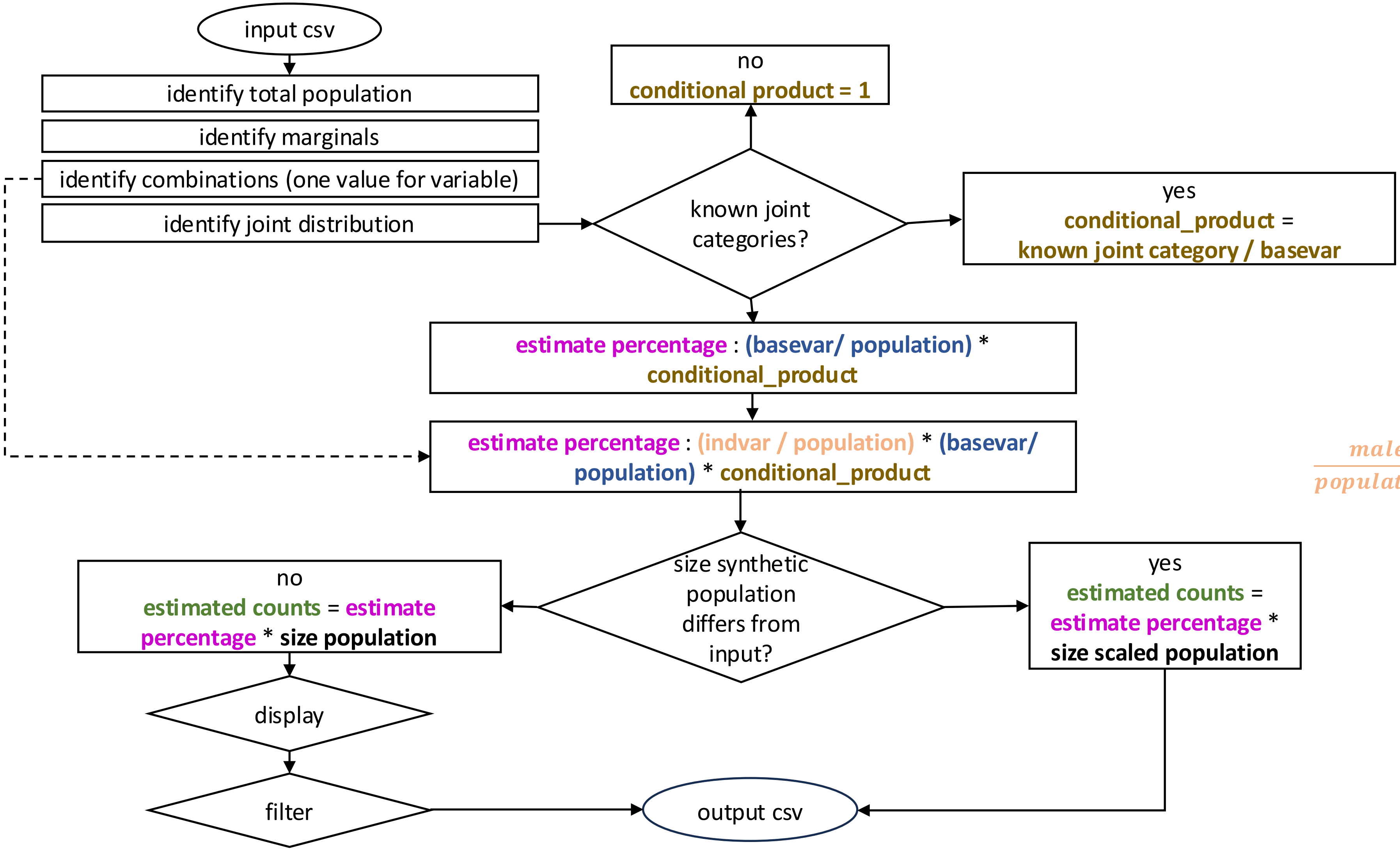
- hypertension over age
- heart failure over age

goal of service: identify joint distribution for all combinations



gender	age	hpt	hf	value
male				3073047
female				3259977
	30			1745215
	3060			2832088
	60100			1755721
		yes		1193445
		no		5139579
			yes	93926
			no	6239098
	30	yes		3547
	3060	yes		252543
	60100	yes		937355
	30	no		1741668
	3060	no		2579545
	60100	no		818366
	30		yes	424
	3060		yes	8459
	60100		yes	85043
	30		no	1744791
	3060		no	2823629
	60100		no	1670678

The algorithm



input csv

identify total population

identify marginals

identify combinations (one value for variable)

identify joint distribution

known joint categories?

no
conditional product = 1

yes
conditional_product = known joint category / basevar

estimate percentage : (basevar/ population) * conditional_product

estimate percentage : (indvar / population) * (basevar/ population) * conditional_product

size synthetic population differs from input?

no
estimated counts = estimate percentage * size population

yes
estimated counts = estimate percentage * size scaled population

display

filter

output csv

Estimate percentageM30HPTHF

gender male, age 30, hptyes, hf yes

$$\frac{male}{population} * \left(\frac{age30}{population} * \frac{hptyes, age30}{age30} * \frac{hfno, age30}{age30}\right)$$

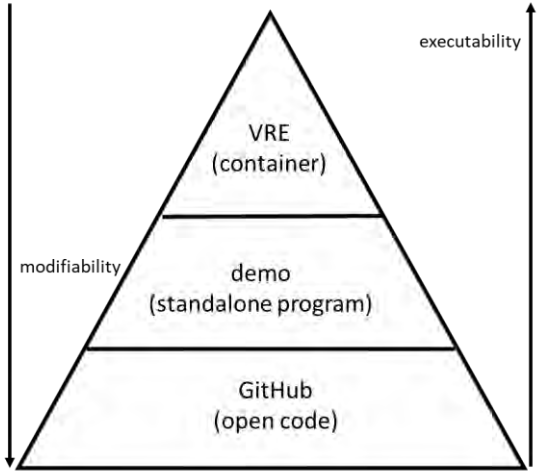
Estimate countM30HPTHF

Estimate percentageM30HPTHF * size synthetic population

Open code

https://github.com/RoccoPaolillo/IPF_multidim.git > synthpopgen.py

cmd line



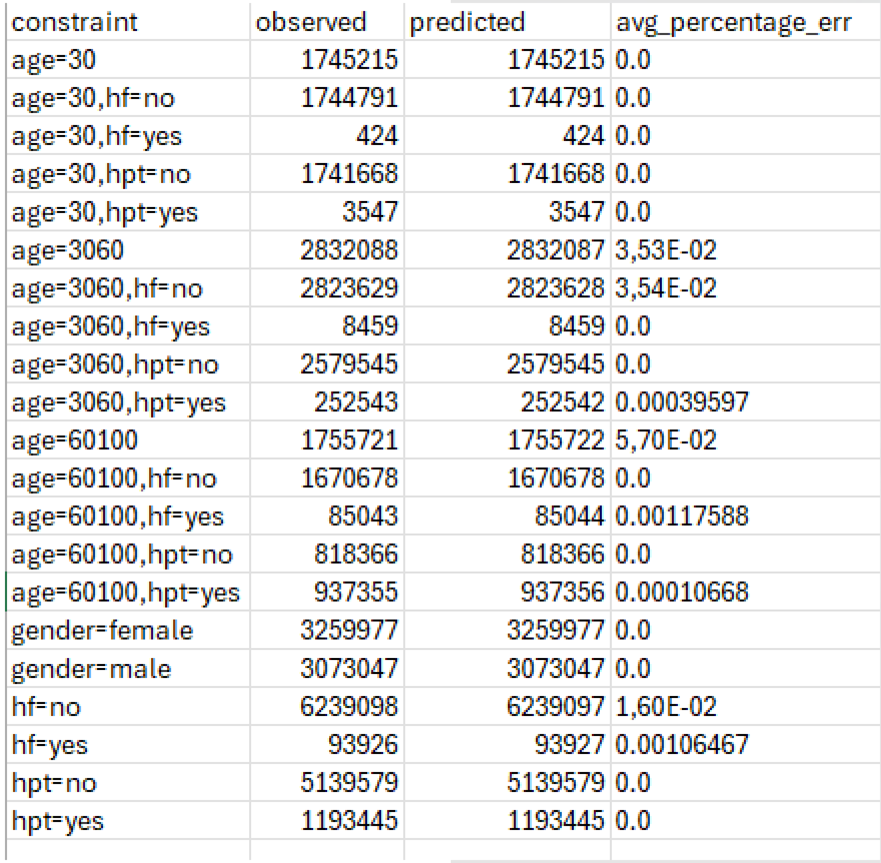
```
python synthpopgen.py -i input_file_tuples.csv \
 -f (filter)
   'all'
   'gender:female,age:3060' \
-d (display)
   'split'
   'aggregate'\
- v (validate)*
--synth-total 20303*
-o results.csv
```

→ pro: high modifiability for users(rewrite, retest...)

→ vs: knowledge coding, dependencies

measure of validation

Average percentage error between input marginals (and joints) data and stimated marginals, **RMSE 0.6**

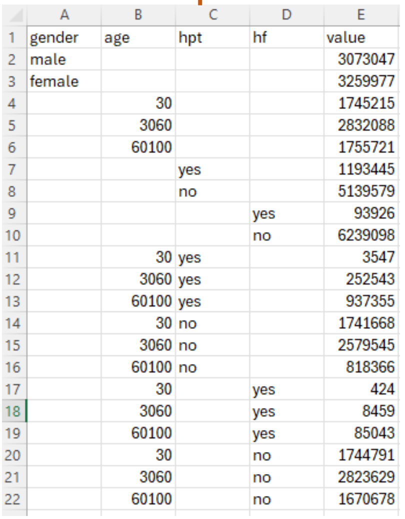


constraint	observed	predicted	avg_percentage_err
age=30	1745215	1745215	0.0
age=30,hf=no	1744791	1744791	0.0
age=30,hf=yes	424	424	0.0
age=30,hpt=no	1741668	1741668	0.0
age=30,hpt=yes	3547	3547	0.0
age=3060	2832088	2832087	3,53E-02
age=3060,hf=no	2823629	2823628	3,54E-02
age=3060,hf=yes	8459	8459	0.0
age=3060,hpt=no	2579545	2579545	0.0
age=3060,hpt=yes	252543	252542	0.00039597
age=60100	1755721	1755722	5,70E-02
age=60100,hf=no	1670678	1670678	0.0
age=60100,hf=yes	85043	85044	0.00117588
age=60100,hpt=no	818366	818366	0.0
age=60100,hpt=yes	937355	937356	0.00010668
gender=female	3259977	3259977	0.0
gender=male	3073047	3073047	0.0
hf=no	6239098	6239097	1,60E-02
hf=yes	93926	93927	0.00106467
hpt=no	5139579	5139579	0.0
hpt=yes	1193445	1193445	0.0

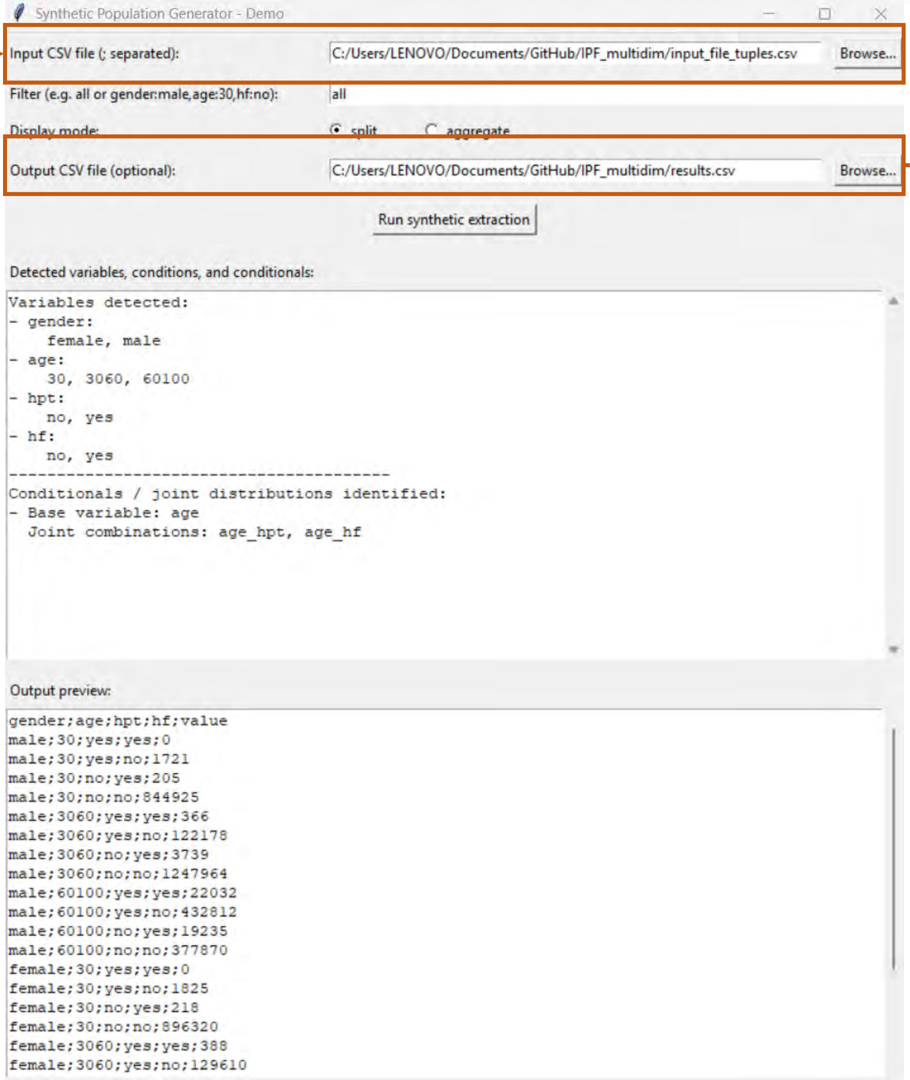
* only if whole population combinations are stimated (-f all)

Standalone program

.exe local file
py:: tkinter



	A	B	C	D	E
1	gender	age	hpt	hf	value
2	male				3073047
3	female				3259977
4		30			1745215
5		3060			2832088
6		60100			1755721
7			yes		1193445
8			no		5139579
9				yes	93926
10				no	6239098
11		30	yes		3547
12		3060	yes		252543
13		60100	yes		937355
14		30	no		1741668
15		3060	no		2579545
16		60100	no		818366
17		30		yes	424
18		3060		yes	8459
19		60100		yes	85043
20		30		no	1744791
21		3060		no	2823629
22		60100		no	1670678



Input CSV file (; separated): C:/Users/LENOVO/Documents/GitHub/IPF_multidim/input_file_tuples.csv

Filter (e.g. all or gender:male,age:30,hf:no): all

Display mode: ● split ○ aggregate

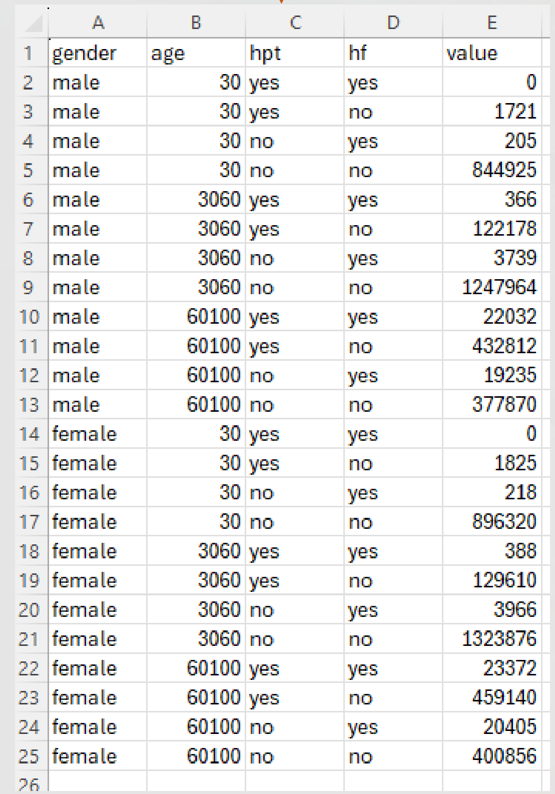
Output CSV file (optional): C:/Users/LENOVO/Documents/GitHub/IPF_multidim/results.csv

Detected variables, conditions, and conditionals:

```
Variables detected:
- gender:
    female, male
- age:
    30, 3060, 60100
- hpt:
    no, yes
- hf:
    no, yes
----------------------------------------
Conditionals / joint distributions identified:
- Base variable: age
  Joint combinations: age_hpt, age_hf
```

Output preview:

```
gender;age;hpt;hf;value
male;30;yes;yes;0
male;30;yes;no;1721
male;30;no;yes;205
male;30;no;no;844925
male;3060;yes;yes;366
male;3060;yes;no;122178
male;3060;no;yes;3739
male;3060;no;no;1247964
male;60100;yes;yes;22032
male;60100;yes;no;432812
male;60100;no;yes;19235
male;60100;no;no;377870
female;30;yes;yes;0
female;30;yes;no;1825
female;30;no;yes;218
female;30;no;no;896320
female;3060;yes;yes;388
female;3060;yes;no;129610
```



	A	B	C	D	E
1	gender	age	hpt	hf	value
2	male	30	yes	yes	0
3	male	30	yes	no	1721
4	male	30	no	yes	205
5	male	30	no	no	844925
6	male	3060	yes	yes	366
7	male	3060	yes	no	122178
8	male	3060	no	yes	3739
9	male	3060	no	no	1247964
10	male	60100	yes	yes	22032
11	male	60100	yes	no	432812
12	male	60100	no	yes	19235
13	male	60100	no	no	377870
14	female	30	yes	yes	0
15	female	30	yes	no	1825
16	female	30	no	yes	218
17	female	30	no	no	896320
18	female	3060	yes	yes	388
19	female	3060	yes	no	129610
20	female	3060	no	yes	3966
21	female	3060	no	no	1323876
22	female	60100	yes	yes	23372
23	female	60100	yes	no	459140
24	female	60100	no	yes	20405
25	female	60100	no	no	400856
26					

→ pro: not coding needed, no dependencies

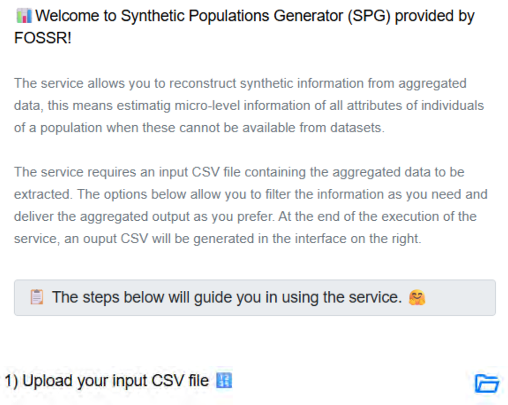
→ vs: local CPU, no modifiability

FOSSR VRE Container

enhance collaboration and tools of researchers through digital platform

Finanziato dall'Unione europea NextGenerationEU
FOSSR VRE
FOSSR VRE Gateway
Collaboration and innovation to enhance social research through accessible data sharing, development of innovative tools, and investment in advanced training
Sign in

https://fossr.d4science.org



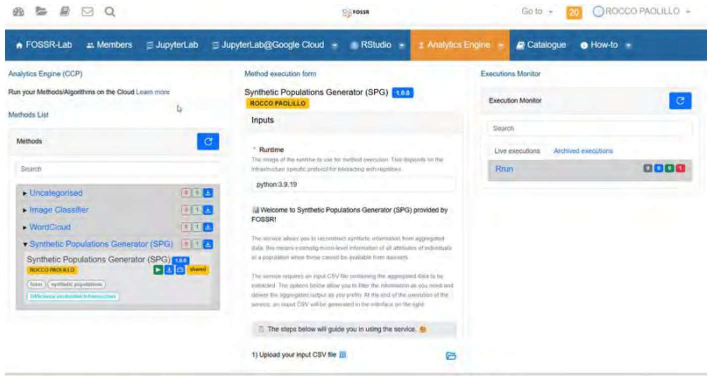
📊 Welcome to Synthetic Populations Generator (SPG) provided by FOSSR!

The service allows you to reconstruct synthetic information from aggregated data, this means estimatig micro-level information of all attributes of individuals of a population when these cannot be available from datasets.

The service requires an input CSV file containing the aggregated data to be extracted. The options below allow you to filter the information as you need and deliver the aggregated output as you prefer. At the end of the execution of the service, an ouput CSV will be generated in the interface on the right.

📋 The steps below will guide you in using the service. 🤗

1) Upload your input CSV file



Analytics Engine (CCP)

Run your Methods/Algorithms on the Cloud Learn more

Methods List

Methods

Search

- Uncategorised
- Image Classifier
- WordCloud
- Synthetic Populations Generator (SPG)
  - Synthetic Populations Generator (SPG) 1.0.0 ROCCO PAOLILLO shared

Method execution form

Synthetic Populations Generator (SPG) 1.0.0 ROCCO PAOLILLO

Inputs

* Runtime

The image of the runtime to use for method execution. This depends on the infrastructure specific protocol for interacting with registries.

python:3.9.19

📊 Welcome to Synthetic Populations Generator (SPG) provided by FOSSR!

📋 The steps below will guide you in using the service. 🤗

1) Upload your input CSV file

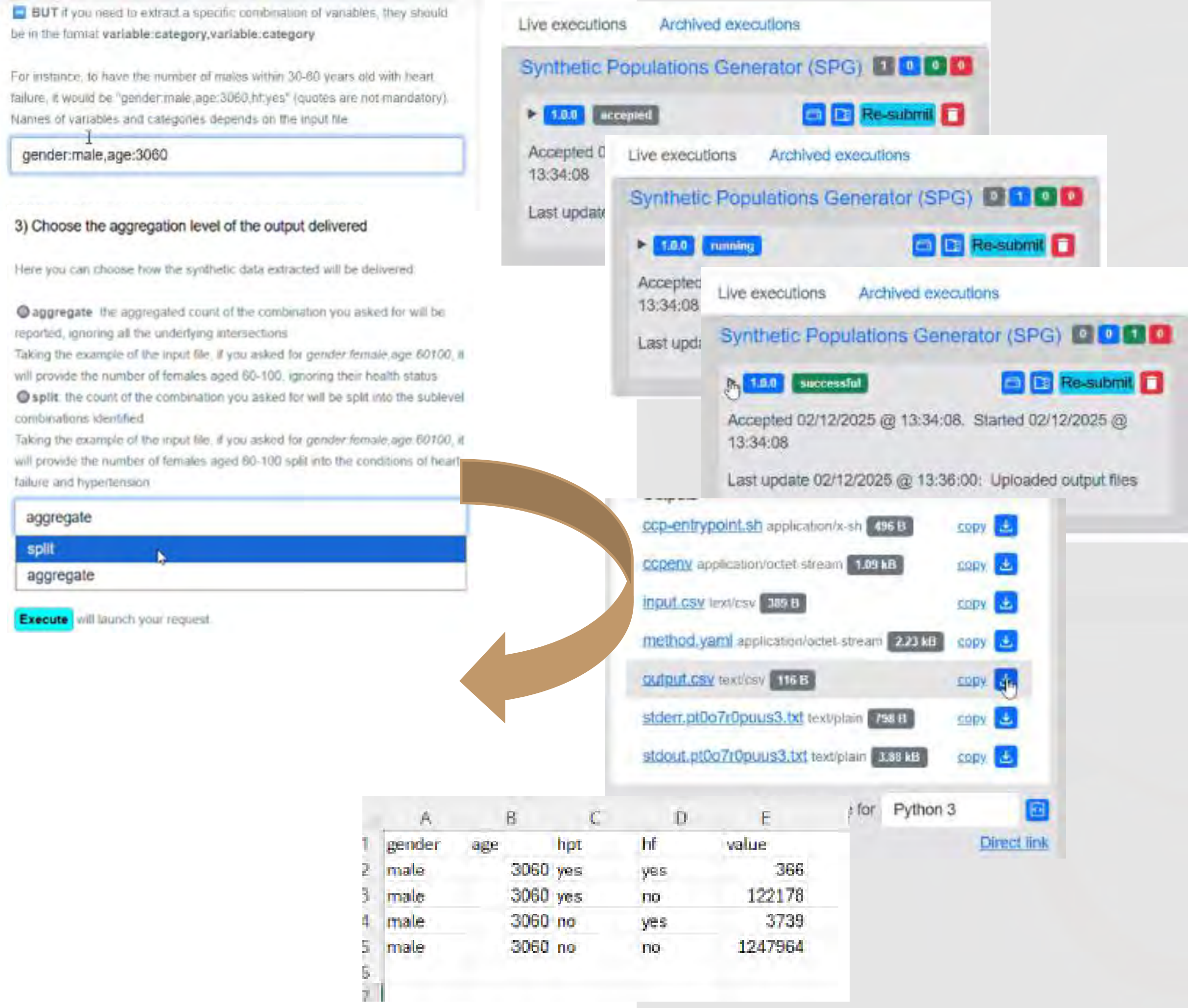
Executions Monitor

Execution Monitor

Search

Live executions | Archived executions

Rrun 0 0 0 1



🟦 **BUT** if you need to extract a specific combination of variables, they should be in the format **variable:category,variable:category**

For instance, to have the number of males within 30-60 years old with heart failure, it would be "gender:male,age:3060,hf:yes" (quotes are not mandatory). Names of variables and categories depends on the input file

gender:male,age:3060

3) Choose the aggregation level of the output delivered

Here you can choose how the synthetic data extracted will be delivered

● **aggregate** the aggregated count of the combination you asked for will be reported, ignoring all the underlying intersections
Taking the example of the input file, if you asked for *gender:female,age:60100*, it will provide the number of females aged 60-100, ignoring their health status

● **split**: the count of the combination you asked for will be split into the sublevel combinations identified
Taking the example of the input file, if you asked for *gender:female,age:60100*, it will provide the number of females aged 60-100 split into the conditions of heart failure and hypertension

aggregate
split
aggregate

Execute will launch your request

Live executions | Archived executions

Synthetic Populations Generator (SPG) 1 0 0 0
▸ 1.0.0 accepted Re-submit
Accepted 0… 13:34:08
Last update…

Live executions | Archived executions

Synthetic Populations Generator (SPG) 0 1 0 0
▸ 1.0.0 running Re-submit
Accepted… 13:34:08
Last upd…

Live executions | Archived executions

Synthetic Populations Generator (SPG) 0 0 1 0
▸ 1.0.0 successful Re-submit
Accepted 02/12/2025 @ 13:34:08. Started 02/12/2025 @ 13:34:08
Last update 02/12/2025 @ 13:36:00: Uploaded output files

ccp-entrypoint.sh	application/x-sh	496 B	copy
ccpenv	application/octet-stream	1.09 kB	copy
input.csv	text/csv	389 B	copy
method.yaml	application/octet-stream	2.23 kB	copy
output.csv	text/csv	116 B	copy
stderr.pt0o7r0puus3.txt	text/plain	798 B	copy
stdout.pt0o7r0puus3.txt	text/plain	3.88 kB	copy

…for Python 3

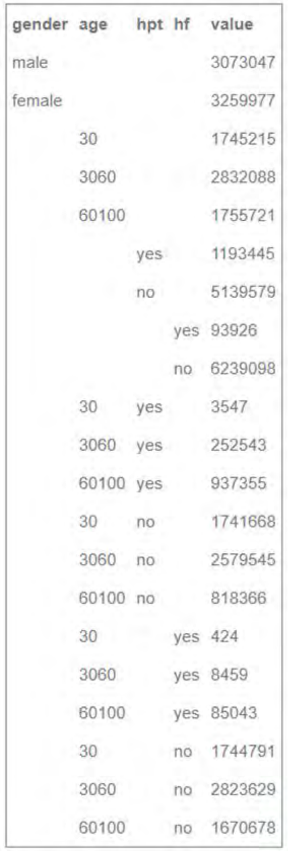
Direct link

	A	B	C	D	E
1	gender	age	hpt	hf	value
2	male	3060	yes	yes	366
3	male	3060	yes	no	122178
4	male	3060	no	yes	3739
5	male	3060	no	no	1247964
6					
7					

→ **pro**: uses D4Science servers, web-app
→ **vs**: no modifiable, internet-dependent

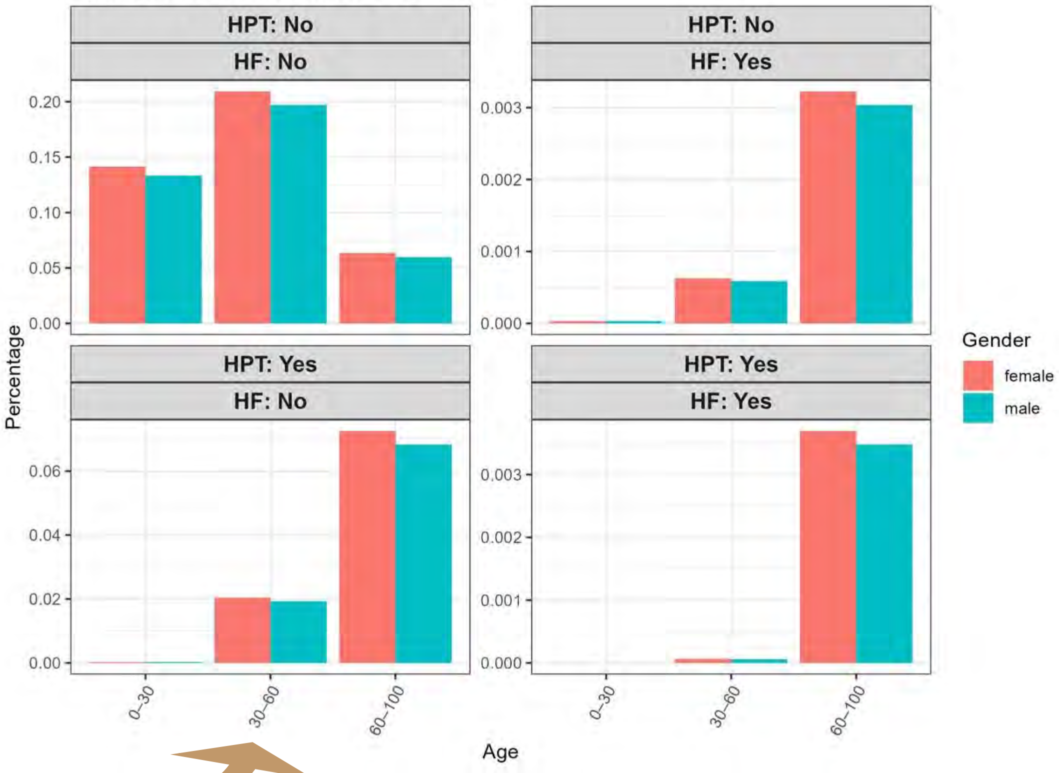
VRE > CCP (Cloud Computing Platform)> Synthetic Populations Generator (SPG)

Output & applications



gender	age	hpt	hf	value
male				3073047
female				3259977
	30			1745215
	3060			2832088
	60100			1755721
		yes		1193445
		no		5139579
			yes	93926
			no	6239098
	30	yes		3547
	3060	yes		252543
	60100	yes		937355
	30	no		1741668
	3060	no		2579545
	60100	no		818366
	30		yes	424
	3060		yes	8459
	60100		yes	85043
	30		no	1744791
	3060		no	2823629
	60100		no	1670678

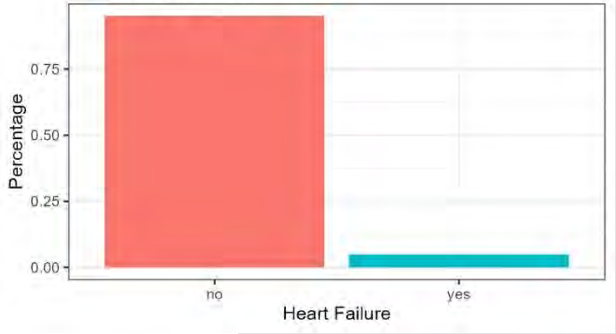
Synthetic population, whole population



gender:male,age:60100, hpt:yes -d aggregate: 454844

gender:male,age:60100, hpt:yes -d split

Males, age 60-100

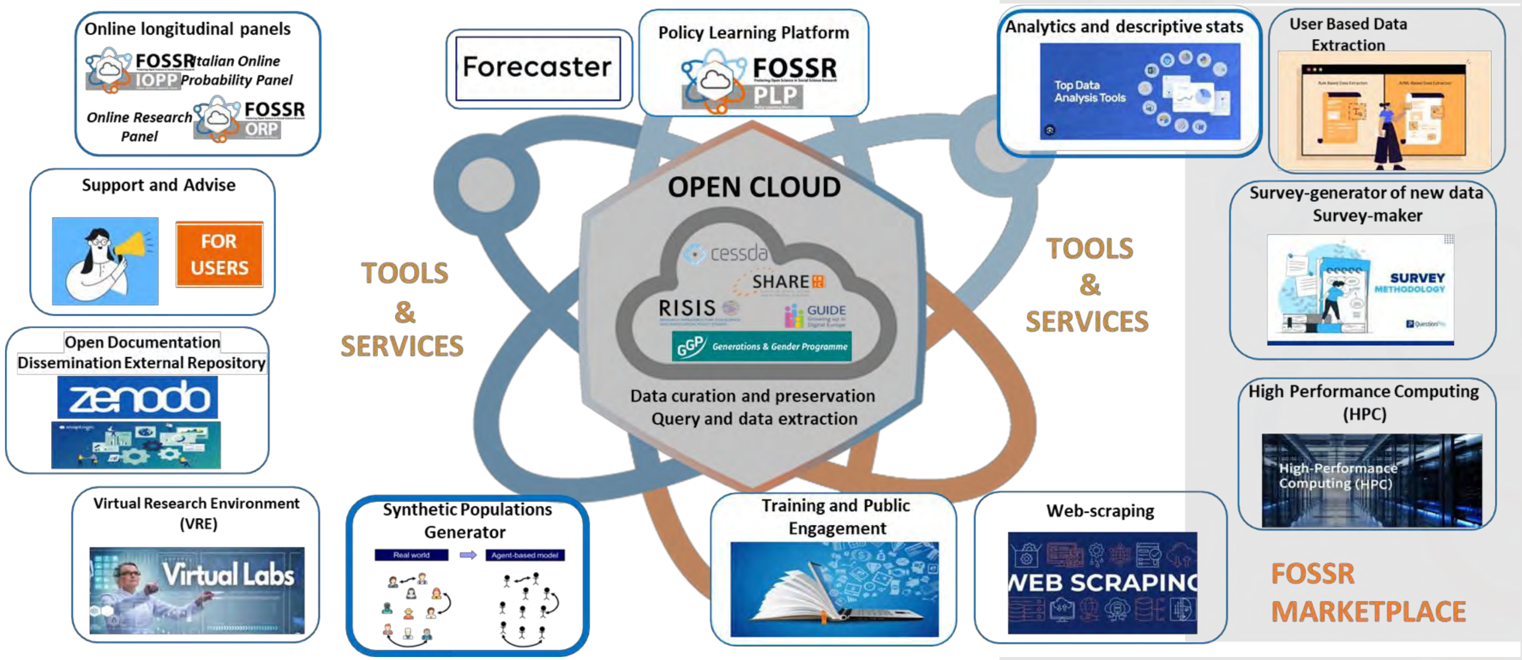


- individual-level attributes as per request
- preserving original data and restrictions
- identify intersections and heterogeneity in the population
- actor-based models initialization (agent-based models, social network analysis,...)

Current and future steps

- Assumes marginals from the same total population
- looking at machine learning as micro-to-macro approach
- Automated iteration over spatial units (census tract)
- Automated multi-source input
- Enhance UX experience and assitance
  - LLM

Deployment into the **FOSSR market place**


Online longitudinal panels
FOSSR IOPP Italian Online Probability Panel
Online Research Panel FOSSR ORP
Support and Advise
FOR USERS
Open Documentation Dissemination External Repository
zenodo
Virtual Research Environment (VRE)
Virtual Labs
Forecaster
Policy Learning Platform
FOSSR PLP
OPEN CLOUD
cessda
SHARE
RISIS
GUIDE
GGP Generations & Gender Programme
Data curation and preservation
Query and data extraction
TOOLS & SERVICES
TOOLS & SERVICES
Synthetic Populations Generator
Real world
Agent-based model
Training and Public Engagement
Web-scraping
WEB SCRAPING
Analytics and descriptive stats
Top Data Analysis Tools
User Based Data Extraction
Survey-generator of new data Survey-maker
SURVEY METHODOLOGY
High Performance Computing (HPC)
High-Performance Computing (HPC)
FOSSR MARKETPLACE

https://github.com/RoccoPaolillo/IPF_multidim.git

References

Paolillo, R., Roxburgh, N., Sbrana, A., Polhill, G., Sabatella, E. C., & Paolucci, M. (2025). Synthetic Populations in Research Infrastructures. In Longitudinal Data Infrastructures in Europe: Tools for Open Science in Social Science Research (pp. 153-164). Cham: Springer Nature Switzerland.

Chapuis, K., Taillandier, P., & Drogoul, A. (2022). Generation of synthetic populations in social simulations: a review of methods and practices. Journal of Artificial Societies and Social Simulation, 25(2).

Bigi, F., Rashidi, T. H., & Viti, F. (2024). Synthetic population: A reliable framework for analysis for agent-based modeling inmobility.Transportation Research Record, 2678(11),1–15.

Yameogo, B. F., Vandanjon, P. O., Gastineau, P., & Hankach, P. (2021). Generating a two-layered synthetic population for French municipalities: Results and evaluation of four synthetic reconstruction methods. Journal of Artificial Societies and Social Simulation, 24,27.

Falck, V. (2025). Generating spatial synthetic populations using Wasserstein generative adver-sarial network: A case study withEU-SILC data for Helsinki and Thessaloniki. Preprint.arXiv:2501.16080.

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., ... & Crouch, S. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research*, *6*, ELIXIR-876.

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A. L., Martinez, C., Psomopoulos, F. E., ... & Yehudi, Y. (2022). FAIR principles for research software (FAIR4RS principles). *Zenodo*.

THANK YOU!

fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

fossr-eu

@fossr

zenodo zenodo.org/communities/fossr

Thank you! Questions?

FOSSR DAYS 2026, 4-5-6 February

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | Consiglio Nazionale delle Ricerche

FOSSR

FOSSR DAYS 2026

policy-makers session
general conference
demonstration event

Feb.4 | Feb. 5 | Feb. 6
online and hybrid events

Feb. 4
POLICY-MAKERS SESSION
HIGH-QUALITY, HIGH-FREQUENCY INDIVIDUAL DATA PRODUCTION IN ITALY: INNOVATION INTRODUCED BY FOSSR
14.00-16.00 | online
CHECK THE UPDATED AGENDA

Feb. 5
FINAL GENERAL CONFERENCE
ACHIEVEMENTS, TOOLS, AND FUTURE PERSPECTIVES FOR OPEN SCIENCE IN THE SOCIAL SCIENCES
9.00-17.30 | CNR and hybrid
Aula Convegni, Piazzale Aldo Moro, 7, Rome
CHECK THE UPDATED AGENDA

Feb. 6
DEMONSTRATION EVENT
THE FOSSR CLOUD INFRASTRUCTURE AND MARKETPLACE
11.30-13.00 | online
CHECK THE UPDATED AGENDA

Organisational details

Scientific/organising committee: Alessandra Maria Stilo (CNR-IRCrES), Serena Fabrizio (CNR-IRCrES), Rita Giuffredi (CNR-IRCrES), Elisa Storace (CNR-IIVR), Emanuela Reale (CNR-IRCrES), Giovanni Cerulli (CNR-IRCrES), Marco Sposato (CNR-IRCrES).

The events will all take place online, except the General conference which will be organised in hybrid modality and will be shared online from the Aula Convegni at CNR headquarter premises: the link will be provided to participants in due time.

This event is supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union - NextGenerationEU under NPRR Grant agreement n. MUR IR0000008

ICAR CNR

How to participate

REGISTRATION
To join one or more events, please register at the link:
l.cnr.it/fossr-days-2026-registration-form
info: communication@fossr.eu
https://www.fossr.eu/eventi/fossr-days-2026/

www.fossr.eu | fossrproject | fossr.eu | fossr-eu

l.cnr.it/fossr-days-2026-registration-form

rocco.paolillo@cnr.it
@roccopaolillo.bsky.social

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NRRP Grant agreement n. MUR IR0000008

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

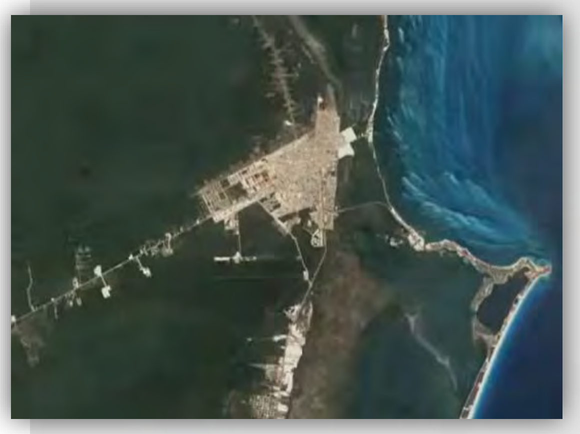
CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Part One

Social complexity and social mechanisms



- Most of social phenomena are inherently collective phenomena
- They assume the definition of a system made of interacting components (e.g. market place, urban landscape, welfare state) where the phenomenon unfolds
- Common scope of different disciplines is to unfold the mechanisms that unfold the dynamics of the phenomenon

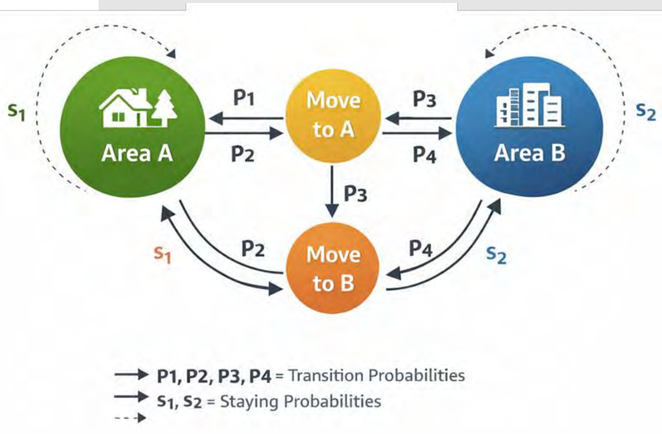
If the concepts of system and the goal to identify mechanisms to disclose the phenomenon is common to many disciplines, different perspectives apply

Different perspectives on the concept of *emergence* and *mechanisms*

Macro

Macro to Macro

- Markov Chain
- Evolution of processes depending on the previous state of the system as unit of analysis

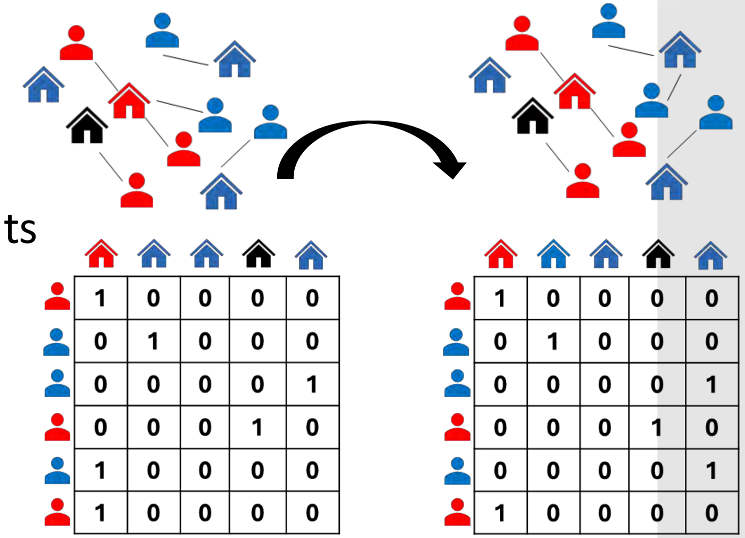


P1, P2, P3, P4 = Transition Probabilities
S1, S2 = Staying Probabilities

Meso

Meso to Macro

- Random graph models
- Evolution of networks as units of analysis



	🟥	🟦	🟦	⬛	🟦
🔴	1	0	0	0	0
🔵	0	1	0	0	0
🔵	0	0	0	0	1
🔴	0	0	0	1	0
🔵	1	0	0	0	0
🔴	1	0	0	0	0

	🟥	🟦	🟦	⬛	🟦
🔴	1	0	0	0	0
🔵	0	1	0	0	0
🔵	0	0	0	0	1
🔴	0	0	0	1	0
🔵	0	0	0	0	1
🔴	1	0	0	0	0

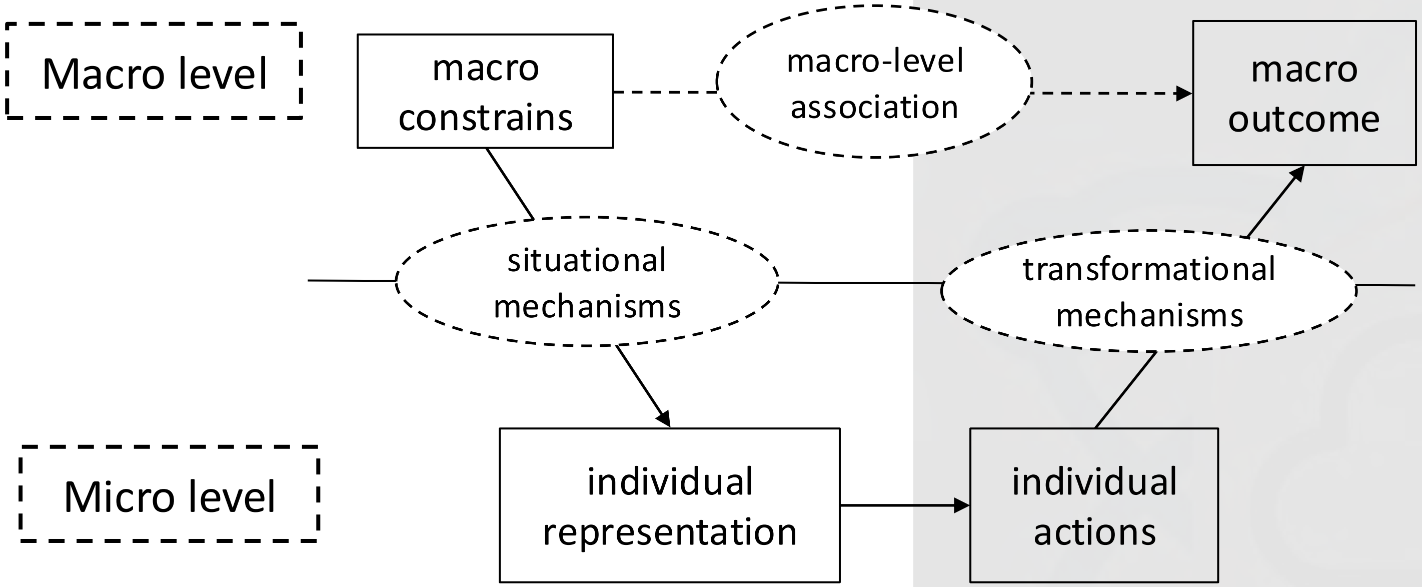
Micro

- Interactions of individuals and institution as the interacting components of the system

Micro to Macro

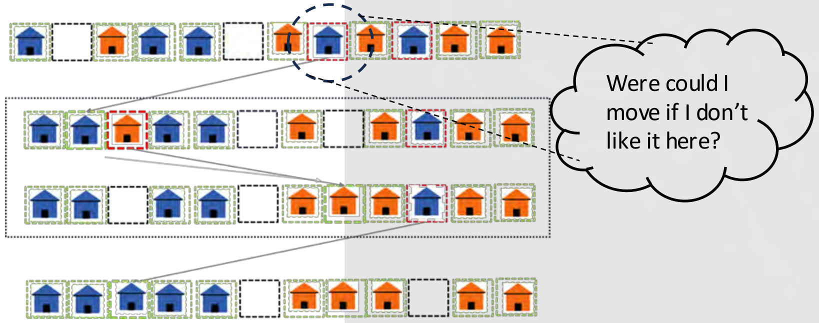
- But sometimes, we might be interested in those mechanisms that move from the micro level, e.g. citizens/institutions with their attitudes, motivations and course of action, but get **outside of the individual agency and inglobe the interaction** of individuals as explicative mechanism of emergence

- The phenomenon is an aggregated, mutual adaptation of individuals rather than the sum of individual action



v

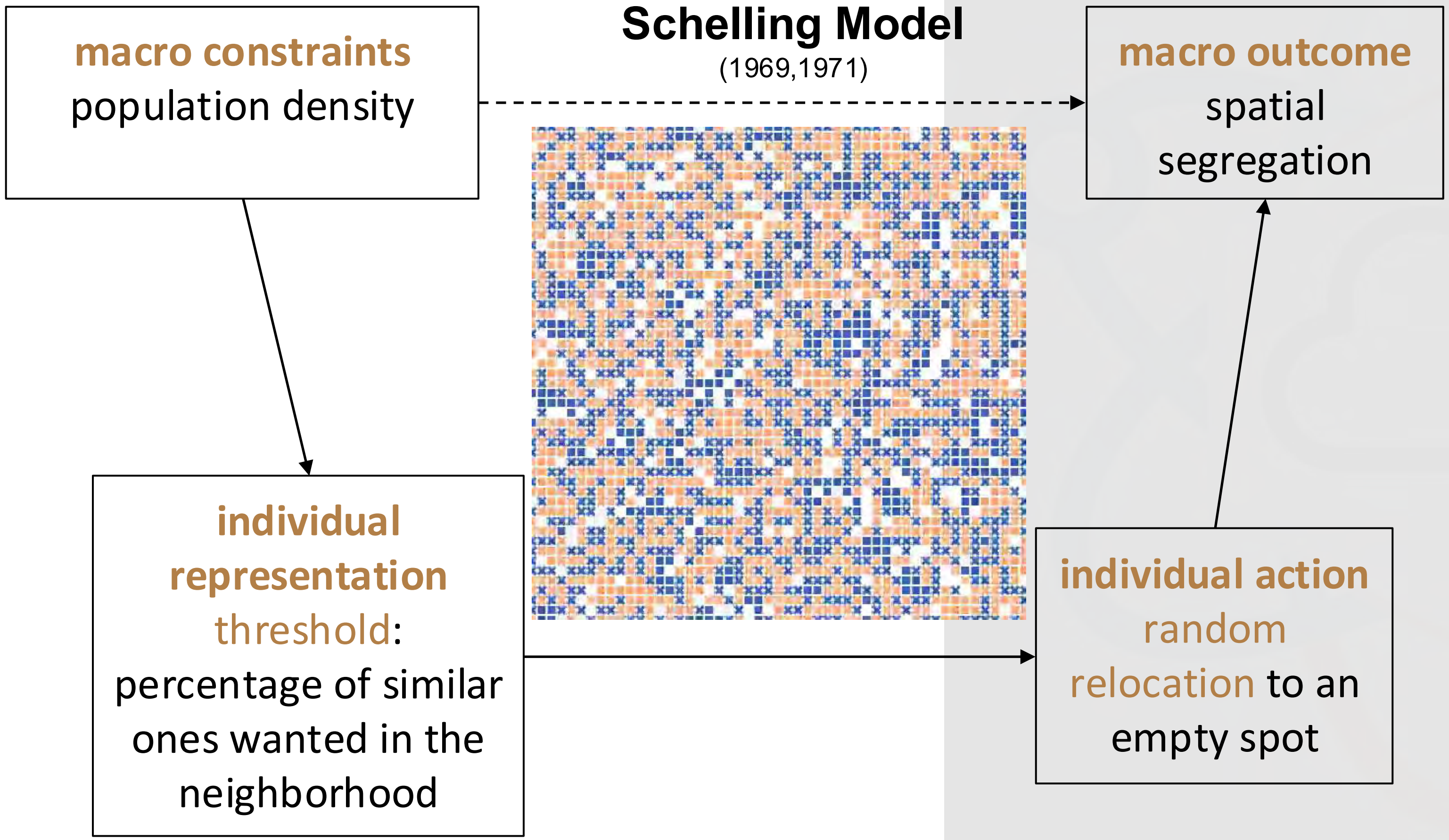
Analytical Sociology
Agent-based Modeling



Coleman Boat (1994), additions by Hedström and Ylikoski (2010), adapted

Example of spatial segregation

- **actors:** households evaluating their neighborhood
- **individual behavior:** preference for a percentage (threshold) of similar ones in their neighborhood (homophily)
- **mechanism of emergence:** cascade effects where the behavior of one household influences the composition of neighborhood and preference of others
- **unexpected outcome:** high levels of spatial segregation, also for mild threshold preferences



Schelling Model

(1969,1971)

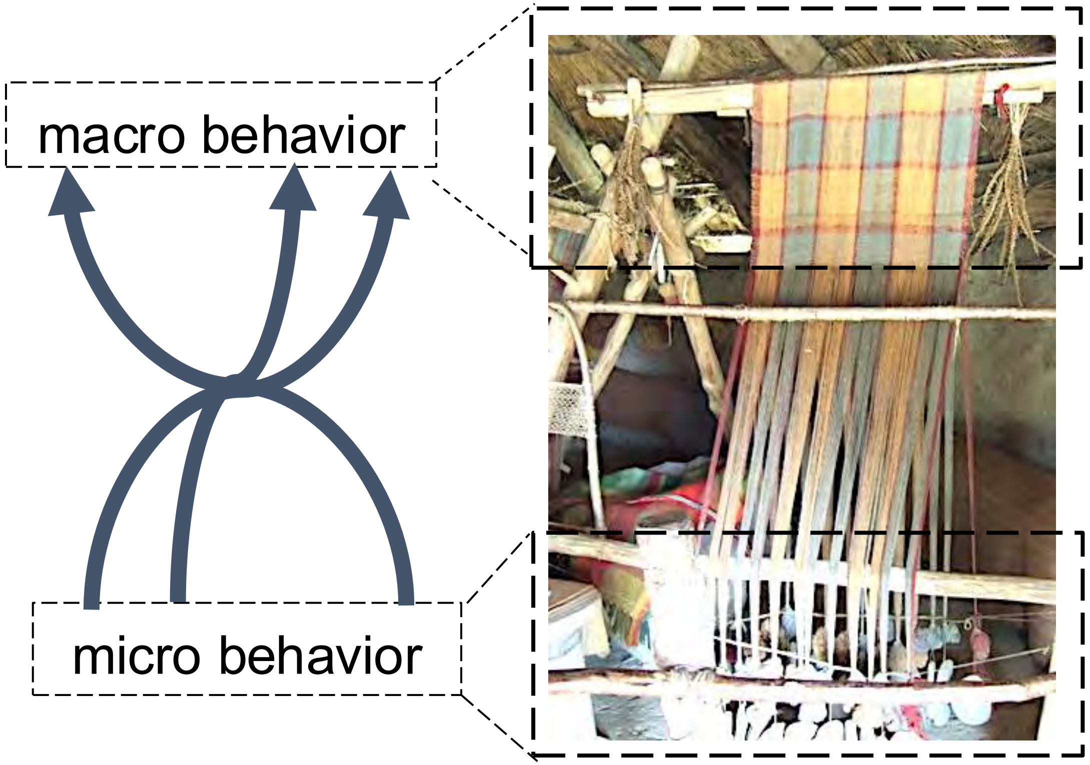
macro constraints
population density

individual representation
threshold: percentage of similar ones wanted in the neighborhood

individual action
random relocation to an empty spot

macro outcome
spatial segregation

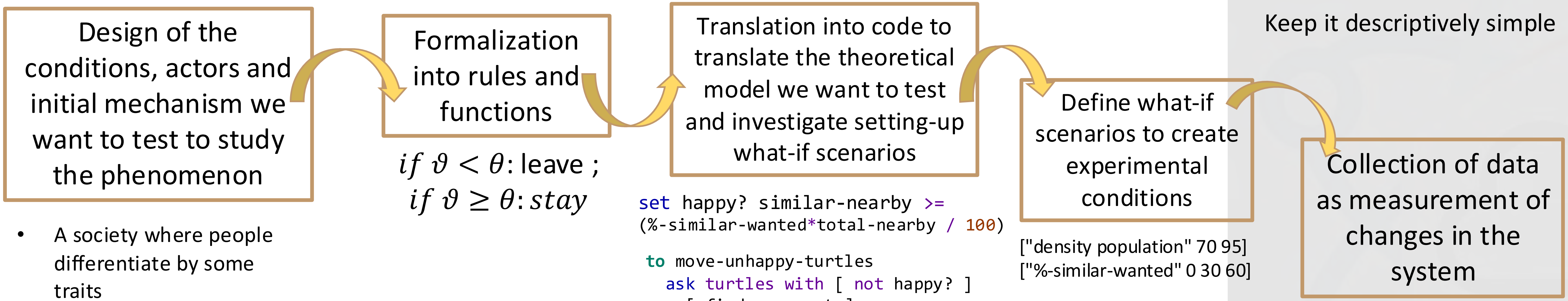
Agent-based Modeling



- **Simulation method** tailored to model the interacting components that constitute the system, e.g. agents representing citizens in an **artificial society**
- We can manipulate both attributes and plan of actions of agents and observe the consequences of interaction of agents executing their plans.
- By manipulating plans and conditions where the agents interact and adapt, we can experiment on and formalize the dynamics of emergence of the collective phenomenon

Social computing with agent-based modeling

KISS
Keep it short simple, stupid
KIDS
Keep it descriptively simple



Design of the conditions, actors and initial mechanism we want to test to study the phenomenon

Formalization into rules and functions

$if\ \vartheta < \theta$: leave ;
$if\ \vartheta \geq \theta: stay$

Translation into code to translate the theoretical model we want to test and investigate setting-up what-if scenarios

```
set happy? similar-nearby >=
(%-similar-wanted*total-nearby / 100)

 to move-unhappy-turtles
   ask turtles with [ not happy? ]
     [ find-new-spot ]
 end
```

Define what-if scenarios to create experimental conditions

["density population" 70 95]
["%-similar-wanted" 0 30 60]

Collection of data as measurement of changes in the system

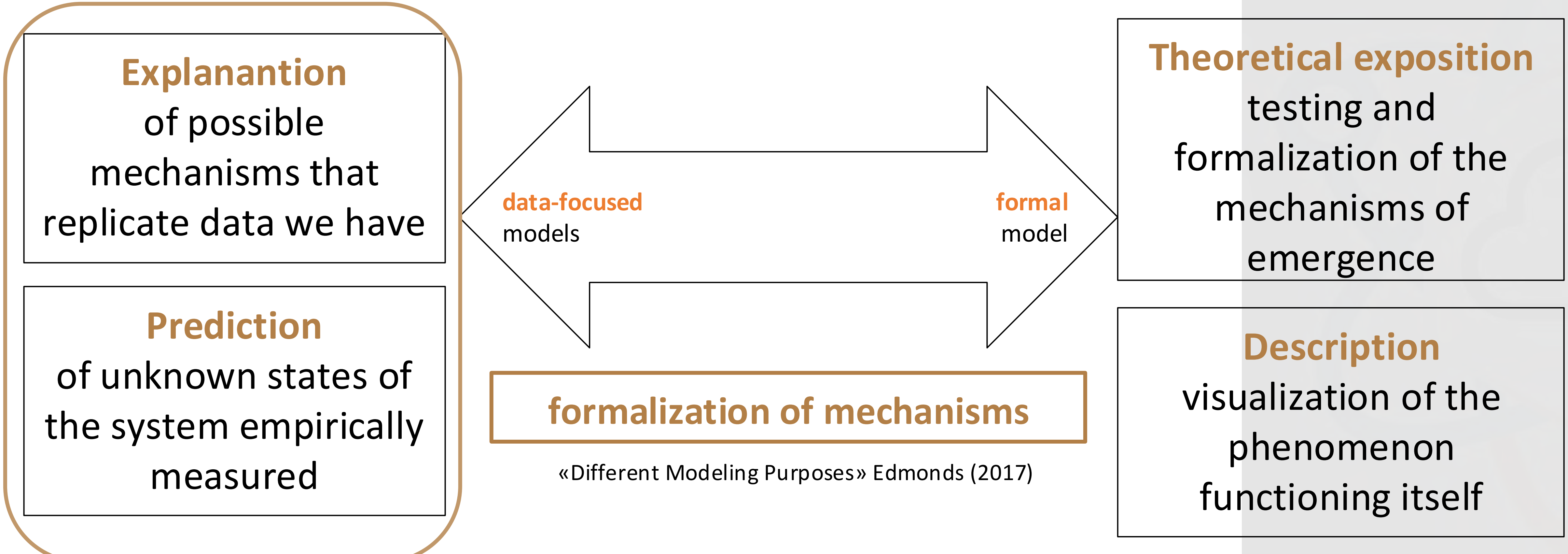
mean exposure to similars in the neighborhood of agents when no one relocates anymore

- A society where people differentiate by some traits
- They stay in a neighborhood if certain threshold of similarity are satisfied
- Can segregation emerge even if the threshold is not that high?

‘If you can grow it, you have explained it’ (Epstein, 2006)

‘If you don’t know how you grew it, you didn’t explain it.’ (Macy & Flache, 2009, p.263)

What is agent-based modeling useful for?



Explanantion of possible mechanisms that replicate data we have

Prediction of unknown states of the system empirically measured

data-focused models

formal model

Theoretical exposition testing and formalization of the mechanisms of emergence

Description visualization of the phenomenon functioning itself

formalization of mechanisms

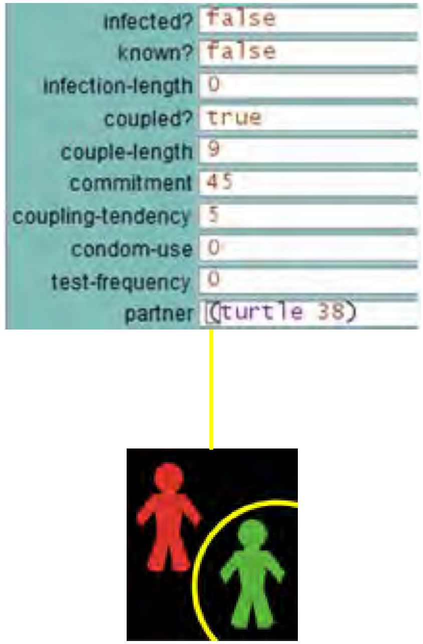
«Different Modeling Purposes» Edmonds (2017)

Let's see some basic concepts and guiding questions

Agents – Who are the actors involved in the phenomenon?

Agents: a virtual object capable of elaborating information and able to execute an action (individuals, institutions, households...)

- **Intentionality**: acting based on goals or plans
- **Proactivity**: initiating actions rather than waiting
- **Reactivity**: responding to external stimuli or changes
- **Prosociality**: acting in coordination with others (social agents)



infected?	false
known?	false
infection-length	0
coupled?	true
couple-length	9
commitment	45
coupling-tendency	5
condom-use	0
test-frequency	0
partner	(turtle 38)

State variable: what characteristics cannot change through time? Ethnicity, gender

Dynamic variable: what characteristics can change through time? Opinions, preferences

Global variable: shared by all agents

Local variable: shared by specific agents or class of agents

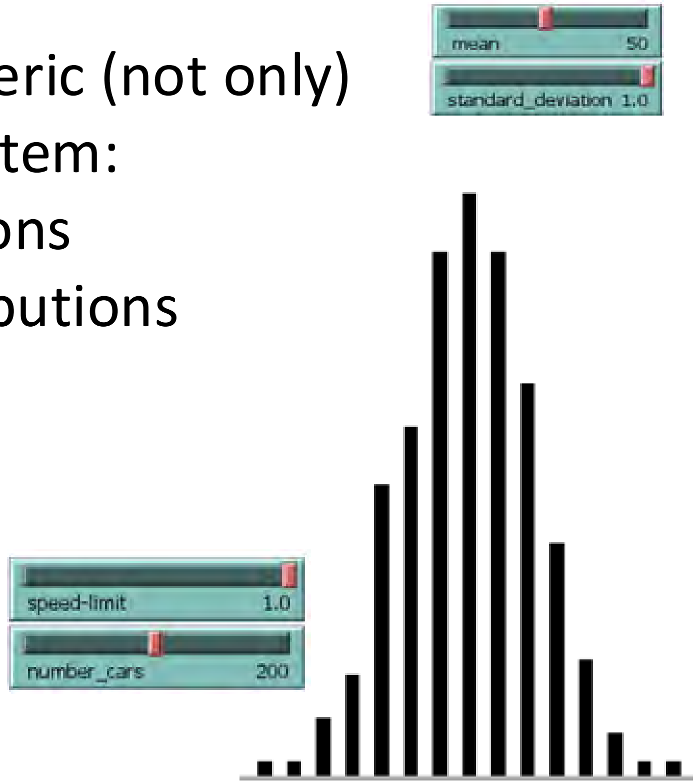
Heterogeneous vs **Homogeneous** (attributes distribution)

Attributes, Beliefs, Desire, Intentions (A+BDI)

Parameters – What are the conditions affecting the phenomenon?

Parameters: tunable numeric (not only) variables to modify the system:

- calibrate initial conditions
- agents' variables distributions



mean 50

standard_deviation 1.0

speed-limit 1.0

number_cars 200

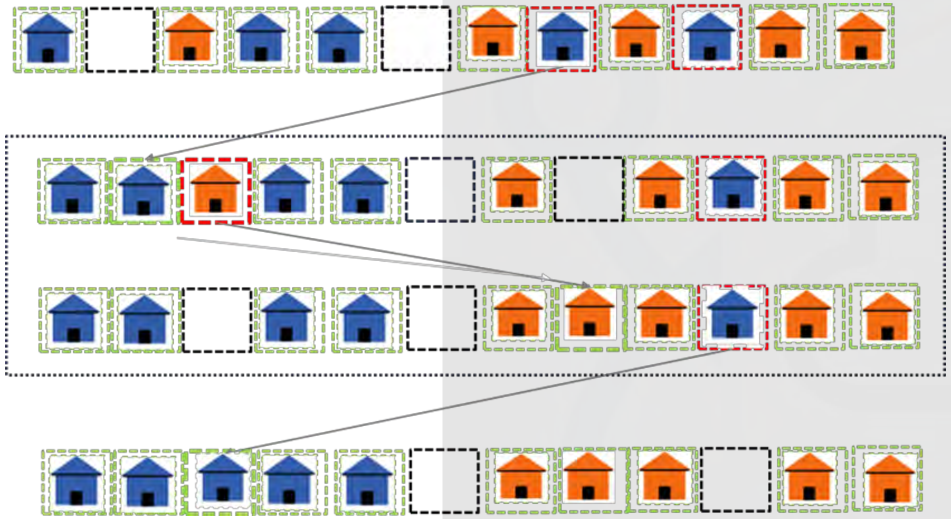
Global parameter: variable affecting all agents, and every agent can interact with
Belief shared by all agents

Local variable: accessible only to some specific agents
Norm specific to a class of agents

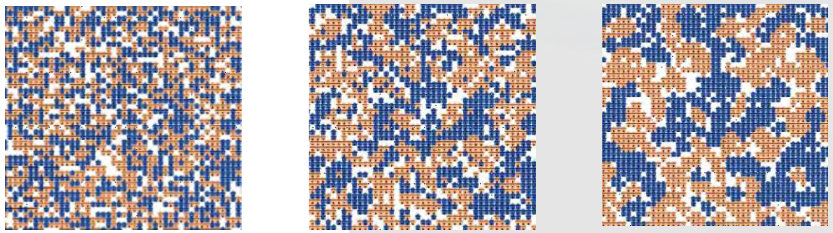
Evolution (I) – How the phenomen emerges through agents' interaction?

It is not much **time** as a continuous variable, rather the evolution of the system along two interconnected concepts:

- **micro level: schedule** of activation of agents' behavior
- **macro level transition phase** of the system changing due to mutual adaptation of the agents



Cascade effect of the behavior of one agent on the neighborhood composition to other agents, affecting segregation at macro level



Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

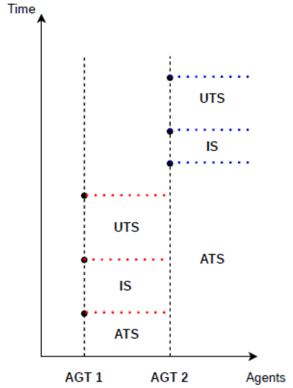
Consiglio Nazionale delle Ricerche

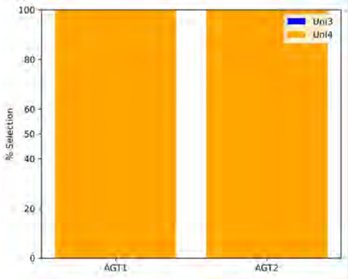
Evolution (II) – How does the order of agents’ action influence each other?

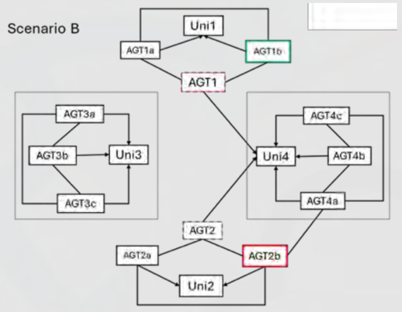
Parallelization: how the behavior is executed
Synchronous behavior: agents act together in parallel
Asynchronous behavior: agents act sequentially
(physical threads)

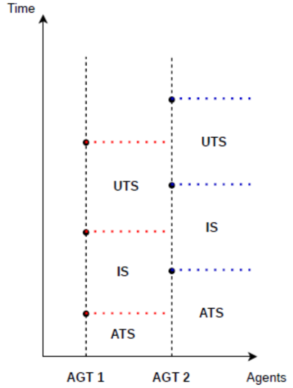
Synchrony: when the behavior is executed
Synchronization: agents decide based on the same knowledge of the world, including effects of actions of others
(they act «in parallel»)

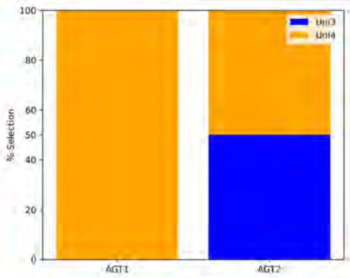
Example: scholar agents choose between two universities based on the chance to be introduced to elective authors based on potential shared connections. The sequential order of agents can affect the decision of those who select after

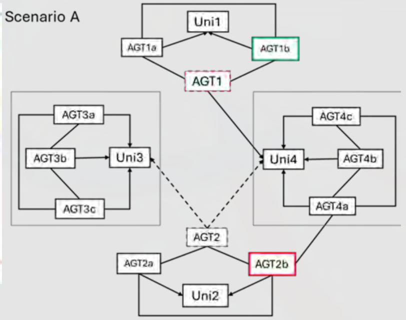










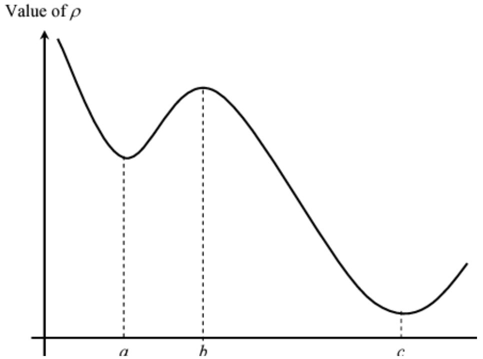


UTS > ATS: update shared knowledge

Longo, Paolillo, Ceriani (2026, forthcoming)

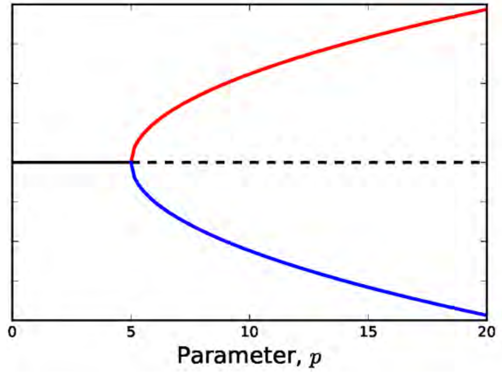
Outcome (I) – How can I read the evolution of the phenomenon?

Tipping points

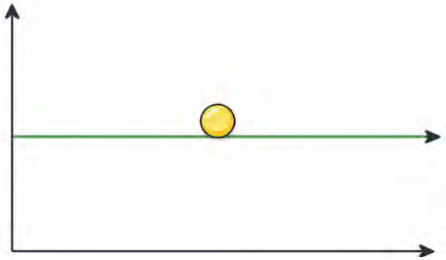


A sudden transition of the system is narrowed to one direction
e.g. in Schelling model the local level of segregation triggers relocation so that segregation becomes steady

Bifurcation

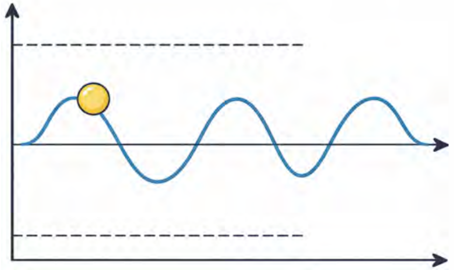


A moment where the phenomenon can diverge in two opposite directions with equal probability
e.g. in case of political polarization



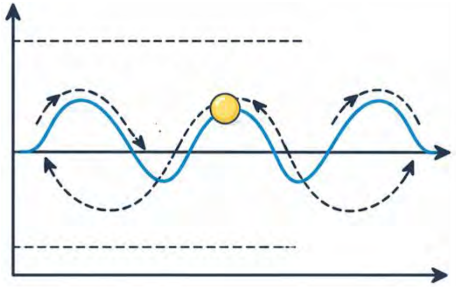
stable equilibrium

not mutable system
(Schelling, consensu)



cyclic equilibrium

system follows a trend of sequential cycles
(grass & sheeps, gentrification)

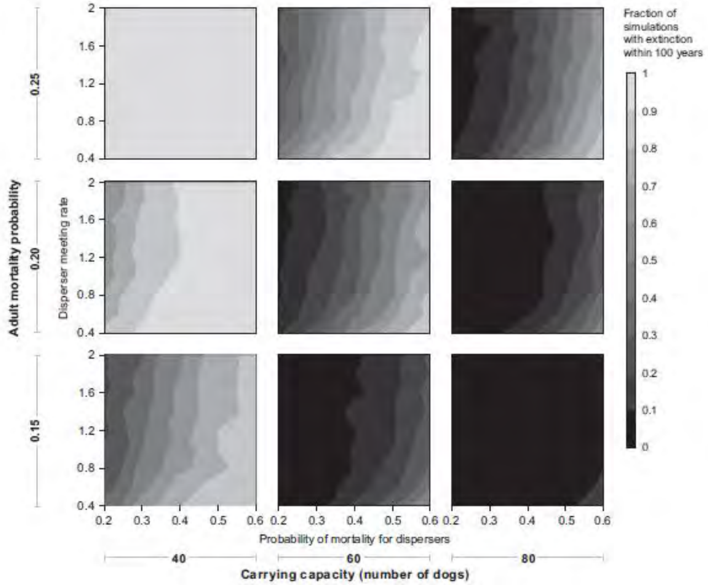


dynamic equilibrium

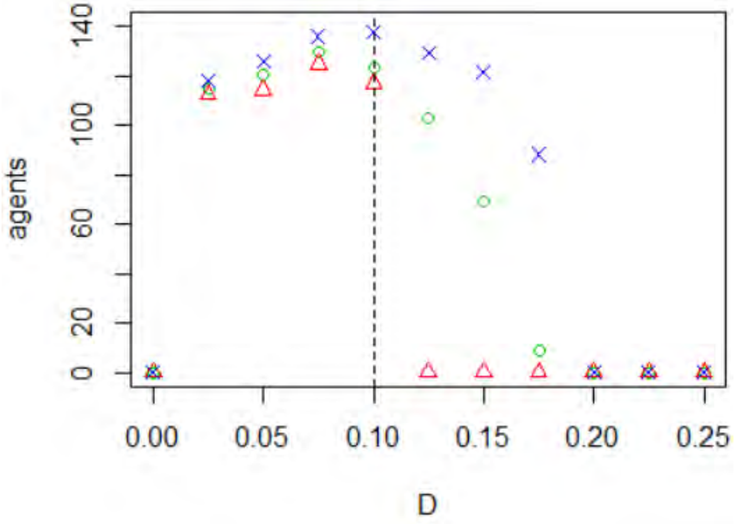
oscillations/inflows/outflows causing the system to apparently remain in balance
(e.g. supply & demand)

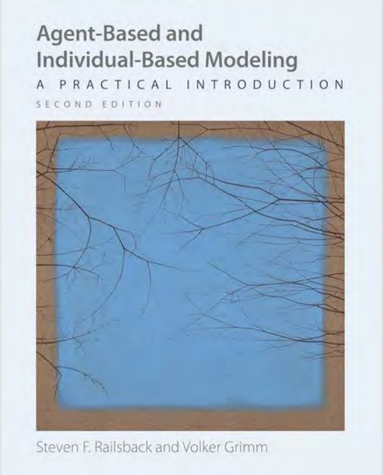
Outcome (II) – How can I identify how the conditions affect the phenomenon?

Global sensitivity analysis: pattern-oriented-modeling (POM) interaction between parameters (space of the model) to understand the overall mechanisms of the model

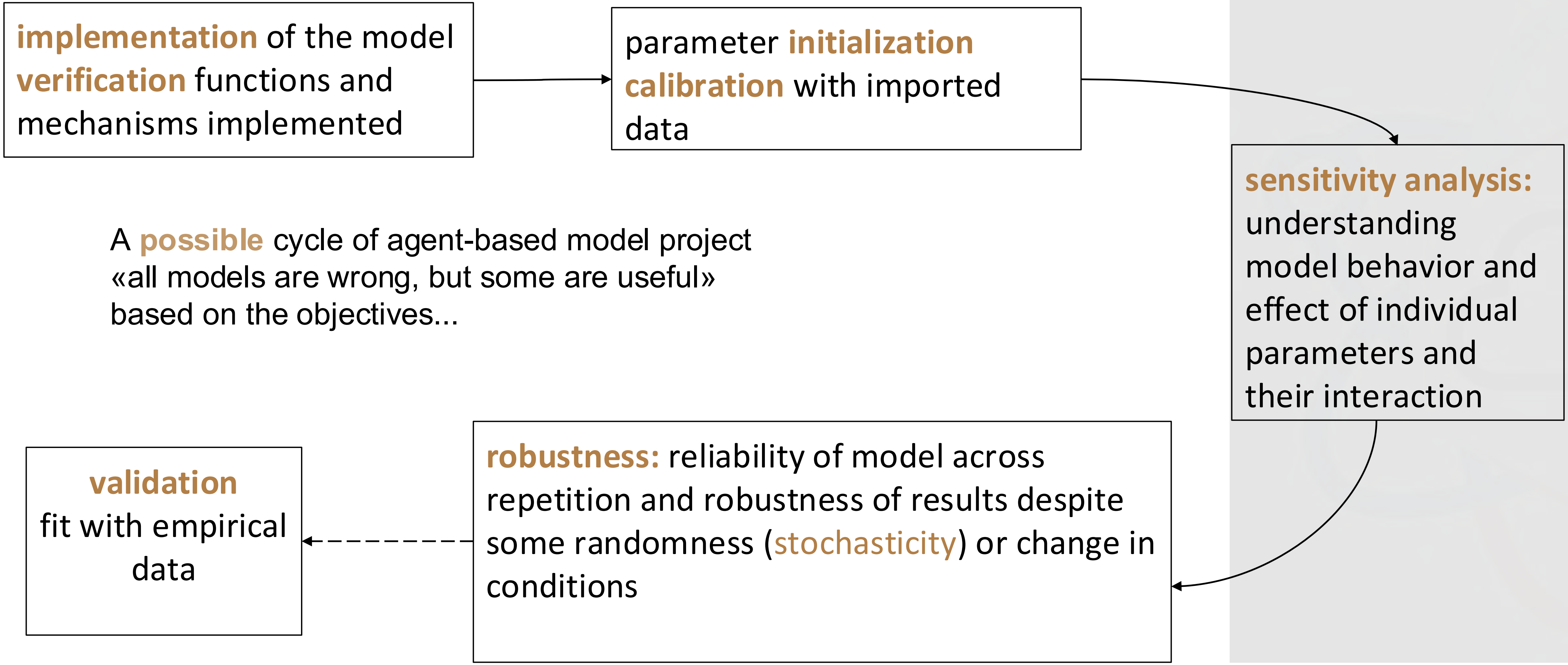


Local sensitivity analysis: one-factor-at-time (OFAT), focus on the effect of one specific parameter (nominal value) over the others




Agent-Based and Individual-Based Modeling
A PRACTICAL INTRODUCTION
SECOND EDITION
Steven F. Railsback and Volker Grimm

Compare what-if scenarios and measures



implementation of the model
verification functions and mechanisms implemented

parameter **initialization**
calibration with imported data

sensitivity analysis: understanding model behavior and effect of individual parameters and their interaction

robustness: reliability of model across repetition and robustness of results despite some randomness (stochasticity) or change in conditions

validation
fit with empirical data

A **possible** cycle of agent-based model project
«all models are wrong, but some are useful»
based on the objectives...

Parameters
What are the conditions affecting the phenomenon?

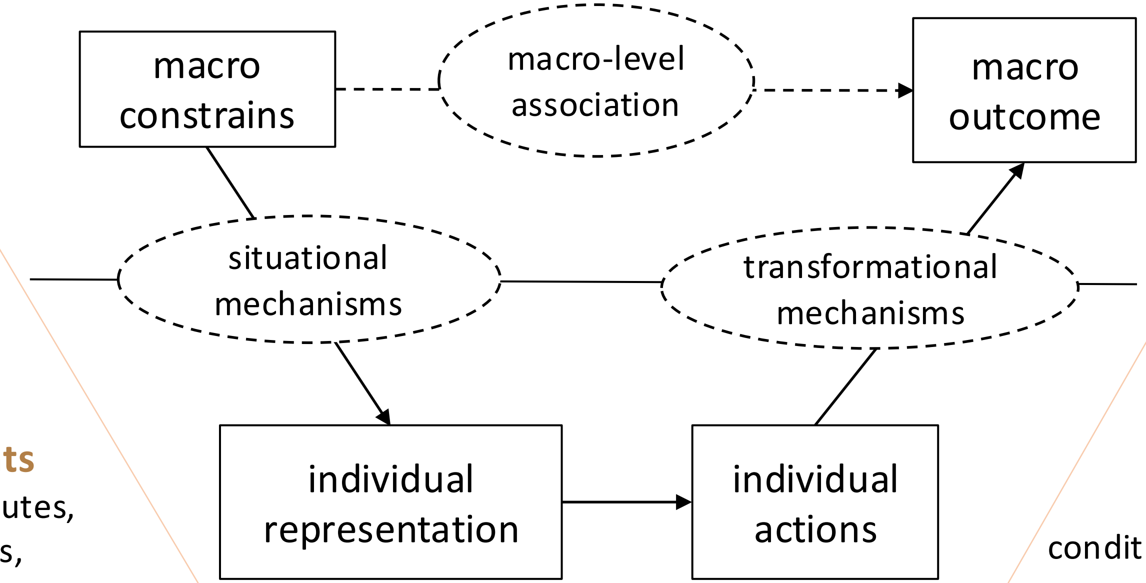
Design

Agents
Who are the actors involved in the phenomenon?

Formalization

Agents
Attributes,
Beliefs,
Desires,
Intentions

Code



macro constrains

macro-level association

macro outcome

situational mechanisms

transformational mechanisms

individual representation

individual actions

Evolution
How does the order of agents' action influence each other

Outcome
How can I read the evolution of the phenomenon?

Evolution
How the phenomenon emerges through agents' interaction?

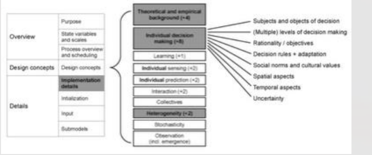
Outcome
How can I identify how the conditions affect the phenomenon?

What-if scenarios

Adapt your model to the grammars of a programming language

Let's implement to a case study and one programming tool (NetLogo)

ODD+D protocol to clarify ideas


Overview
Purpose
State variables and scales
Process overview and scheduling
Design concepts
Design concepts
Details
Implementation details
Initialization
Input
Submodels
Theoretical and empirical background (+4)
Individual decision making (+8)
Learning (+1)
Individual sensing (+2)
Individual prediction (+2)
Interaction (+2)
Collectives
Heterogeneity (+2)
Stochasticity
Observation (incl. emergence)
Subjects and objects of decision
(Multiple) levels of decision making
Rationality / objectives
Decision rules + adaptation
Social norms and cultural values
Spatial aspects
Temporal aspects
Uncertainty

Müller et al., 2013

References, suggested readings and tools

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. Annual review of sociology, 28(1), 143-166.

Edmonds, B. (2017). Different modelling purposes. Simulating social complexity: A handbook, 39-58.

Edmonds, B., & Moss, S. (2004, July). From KISS to KIDS–an 'anti-simplistic'modelling approach. In International workshop on multi-agent systems and agent-based simulation (pp. 130-144). Berlin, Heidelberg: Springer Berlin Heidelberg.

Epstein, J. M., & Axtell, R. (1996). Growing artificial societies: social science from the bottom up. Brookings Institution Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. Annual review of sociology, 36, 49-67.

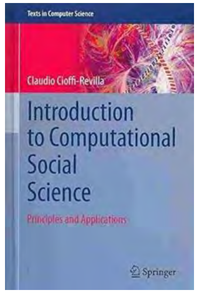
Macy, Michael W., and Andreas Flache. 2009. "Social Dynamics from the Bottom Up: Agent-based Models of Social Interaction." In Hedström, P. and Bearman, P. (Eds.) The Oxford Handbook of Analytical Sociology. Oxford, UK: Oxford University Press.

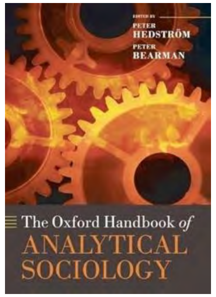
Schelling, T. C. (1969). Models of segregation. The American economic review, 59(2), 488-493.

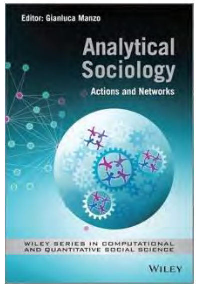
Schelling, T. C. (1971). Dynamic models of segregation. Journal of mathematical sociology, 1(2), 143-186.

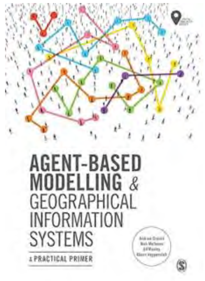
Longo, C.F., Paolillo, R., & Ceriani, M. (Forthcoming). T2B2T: The Ontology for adaptive Agent-driven seamless integration with the Semantic Web. In Proceedings of the Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), To appear on CEUR

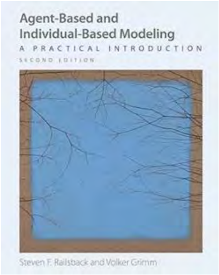
Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., ... & Schwarz, N. (2013). Describing human decisions in agent-based models–ODD+ D, an extension of the ODD protocol. Environmental Modelling & Software, 48, 37-48.


Texts in Computer Science
Claudio Cioffi-Revilla
Introduction to Computational Social Science
Principles and Applications
Springer


PETER HEDSTRÖM
PETER BEARMAN
The Oxford Handbook of ANALYTICAL SOCIOLOGY


Editor: Gianluca Manzo
Analytical Sociology
Actions and Networks
WILEY SERIES IN COMPUTATIONAL AND QUANTITATIVE SOCIAL SCIENCE
WILEY


AGENT-BASED MODELLING & GEOGRAPHICAL INFORMATION SYSTEMS
A PRACTICAL PRIMER


Agent-Based and Individual-Based Modeling
A PRACTICAL INTRODUCTION
SECOND EDITION
Steven F. Railsback and Volker Grimm

COMPLEXITY EXPLORER
SANTA FE INSTITUTE

JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

essa
https://essa.eu.org/

MAILING LIST
SIMSOC@jiscmail.ac.uk

li.u
Institute for Analytical Sociology
Linköping University

Softwares free

NetLogo
https://www.netlogo.org/

MESA python
https://mesa.readthedocs.io/latest/

GAMA Platform
https://gama-platform.org/

Thank you! Questions?

rocco.paolillo@cnr.it

@roccopaolillo.bsky.social

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NRRP Grant agreement n. MUR IR0000008

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Part Two

Parameters
What are the conditions affecting the phenomenon?

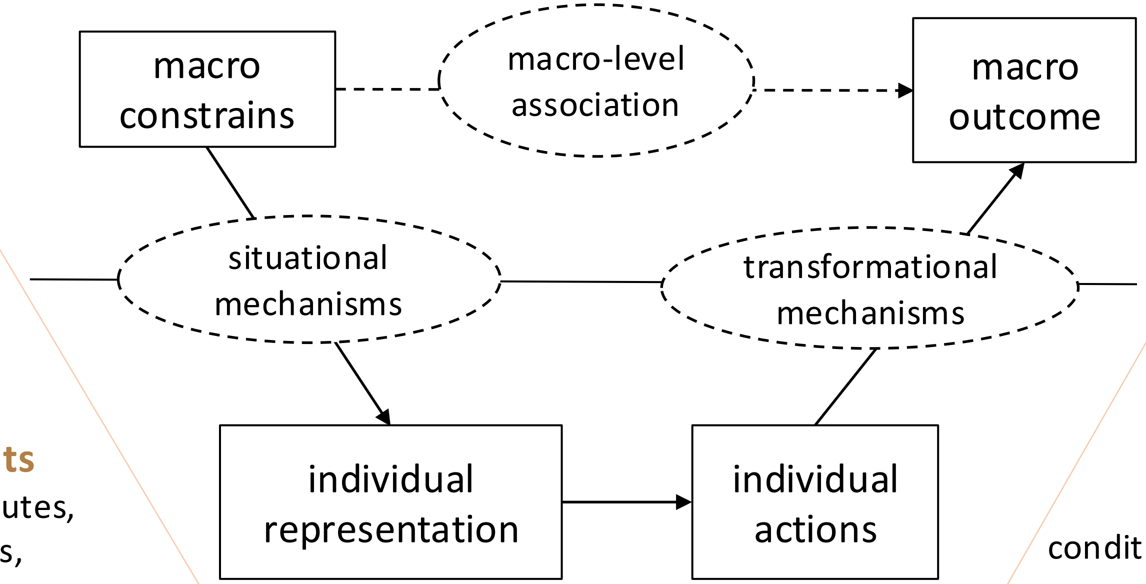
Design

Agents
Who are the actors involved in the phenomenon?

Formalization

Agents
Attributes,
Beliefs,
Desires,
Intentions

Code



Outcome
How can I read the evolution of the phenomenon?

Evolution
How the phenomenon emerges through agents’ interaction?

Outcome
How can I identify how the conditions affect the phenomenon?

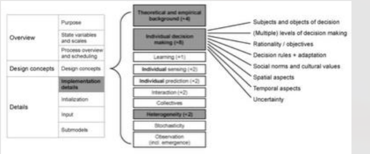
What-if scenarios

Evolution
How does the order of agents’ action influence each other

Adapt your model to the grammars of a programming language

Let’s implement to a case study and one programming tool (NetLogo)

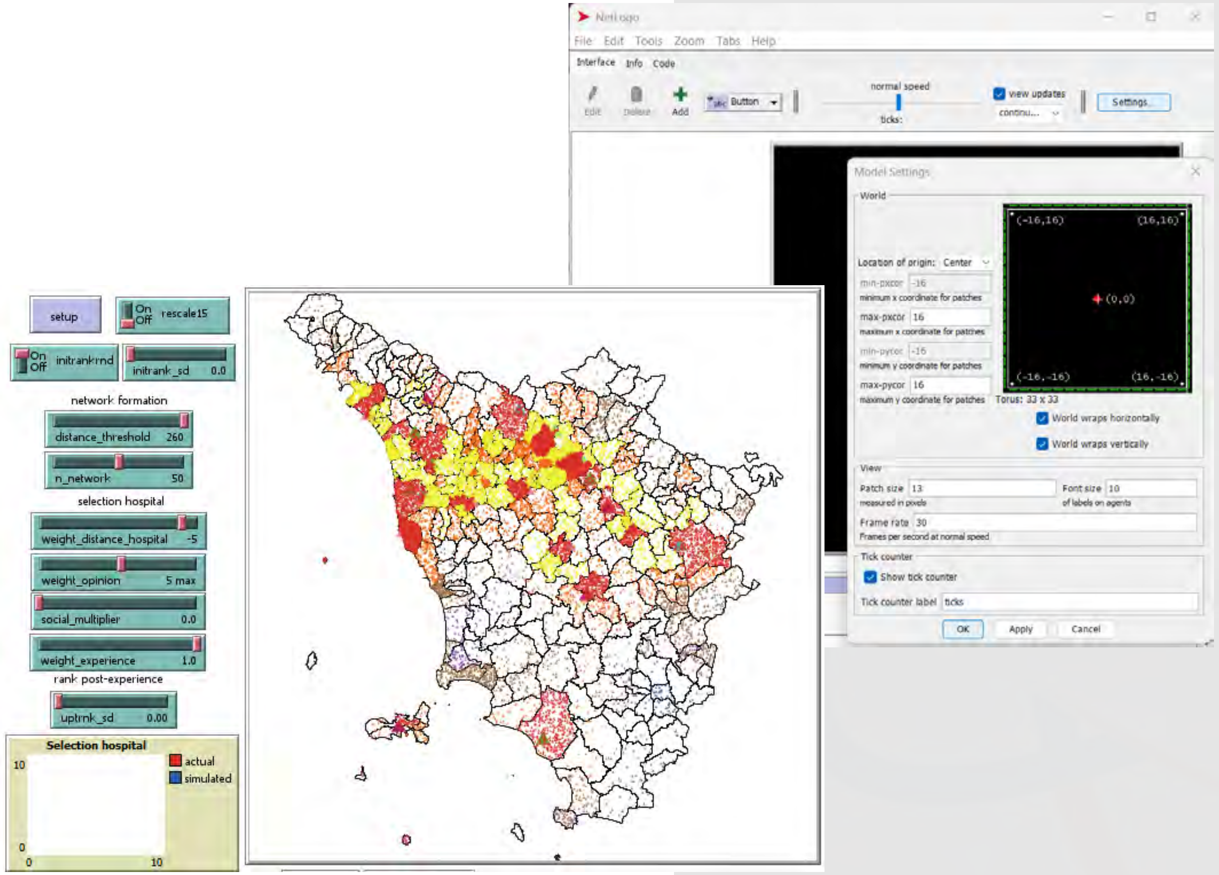
ODD+D protocol to clarify ideas


Overview
Purpose
State variables and scales
Process overview and scheduling
Design concepts
Design concepts
Details
Implementation details
Initialization
Input
Submodels
Theoretical and empirical background (+4)
Individual decision making (+8)
Learning (+1)
Individual sensing (+2)
Individual prediction (+2)
Interaction (+2)
Collectives
Heterogeneity (+2)
Stochasticity
Observation (incl. emergence)
Subjects and objects of decision
(Multiple) levels of decision making
Rationality / objectives
Decision rules + adaptation
Social norms and cultural values
Spatial aspects
Temporal aspects
Uncertainty

Müller et al., 2013

NetLogo

- Platform IDE (interfaccia) e programming language (java + starlogo) specific to agent-based modeling and experiments
- Open Source & User-friendly
- Allows many extensions (shapefile, csv import, random-wheel selection…)
- Widely used in the social science community and continuously maintained (7+ version)
- Programming language tailored to be as intuitive as possible and ready functions
- Supported by documentation (and Chat-GPT)
  - https://docs.netlogo.org/dictionary



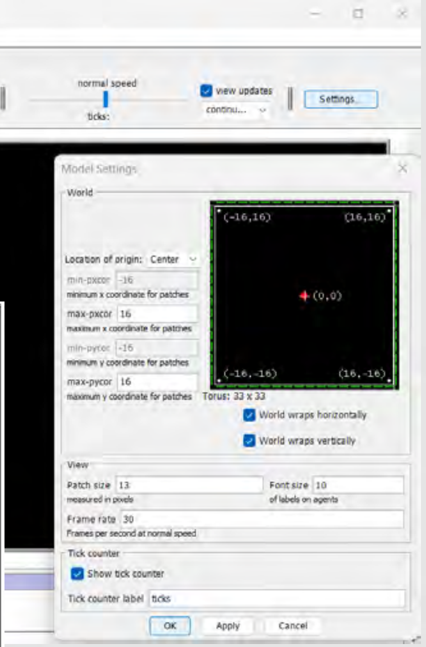
On Off rescale15
On Off initrankrnd
initrank_sd 0.0
network formation
distance_threshold 260
n_network 50
selection hospital
weight_distance_hospital -5
weight_opinion 5 max
social_multiplier 0.0
weight_experience 1.0
rank post-experience
uptrnk_sd 0.00

Model Settings
World
Location of origin: Center
min-pxcor -16
minimum x coordinate for patches
max-pxcor 16
maximum x coordinate for patches
min-pycor -16
minimum y coordinate for patches
max-pycor 16
maximum y coordinate for patches
Torus: 33 x 33
☑ World wraps horizontally
☑ World wraps vertically
View
Patch size 13
measured in pixels
Font size 10
of labels on agents
Frame rate 30
Frames per second at normal speed
Tick counter
☑ Show tick counter
Tick counter label ticks

NetLogo Simulate, Explore, Understand https://www.netlogo.org/

NetLogo

world where things happen



Model Settings

World

Location of origin: Center

min-pxcor -16 minimum x coordinate for patches

max-pxcor 16 maximum x coordinate for patches

min-pycor -16 minimum y coordinate for patches

max-pycor 16 maximum y coordinate for patches

(-16,16) (16,16) (0,0) (-16,-16) (16,-16)

Torus: 33 x 33

☑ World wraps horizontally

☑ World wraps vertically

View

Patch size 13 measured in pixels

Font size 10 of labels on agents

Frame rate 30 Frames per second at normal speed

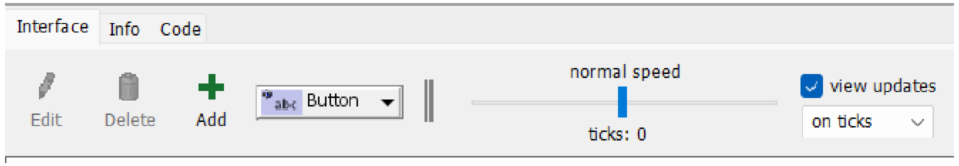
Tick counter

☑ Show tick counter

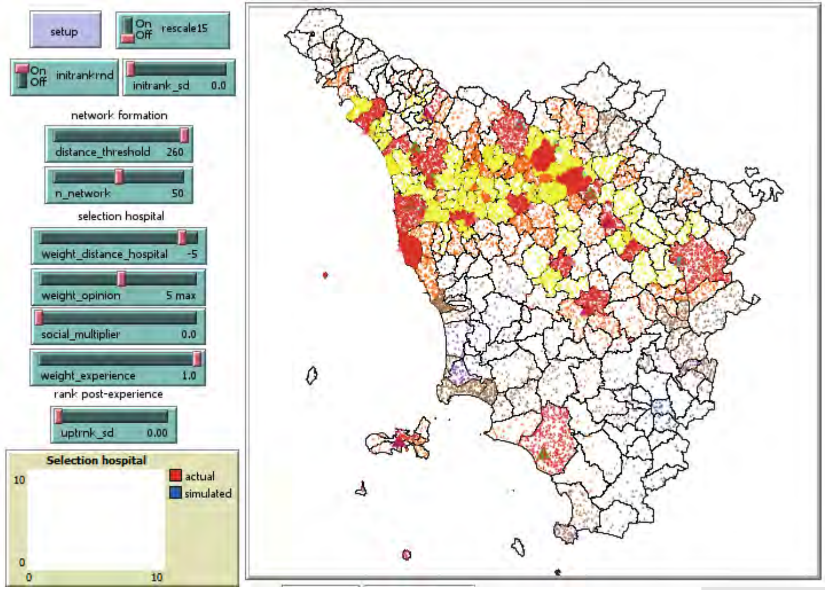
Tick counter label ticks

OK Apply Cancel

Sliders, buttons, chooser to facilitate interaction with parameters in the interface to explore conditions



Interface tab to interact
Info tab tab to document
Code tab to build the model
(also in conjunction with interface)



On Off rescale15

On Off initrankrnd

initrank_sd 0.0

network formation

distance_threshold 260

n_network 50

selection hospital

weight_distance_hospital -5

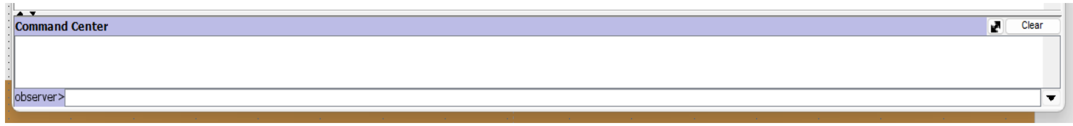
weight_opinion 5 max

social_multiplier 0.0

weight_experience 1.0

rank post-experience

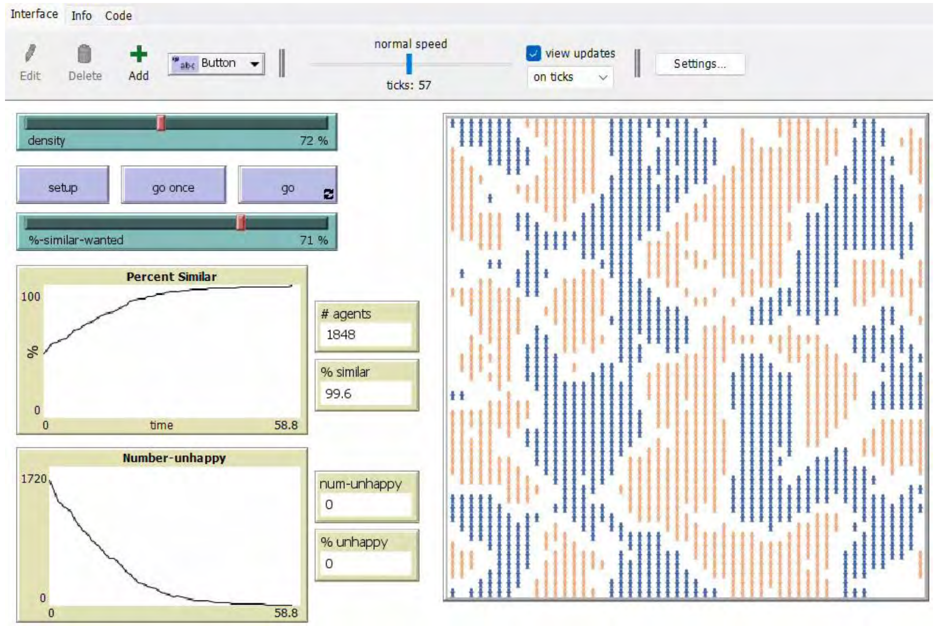
uptrnk_sd 0.00



A command line to interact *on the fly*

Plots, monitors to detect how the phenomenon is emerging

NetLogo



ticks: 57

density 72 %

%-similar-wanted 71 %

agents 1848

% similar 99.6

num-unhappy 0

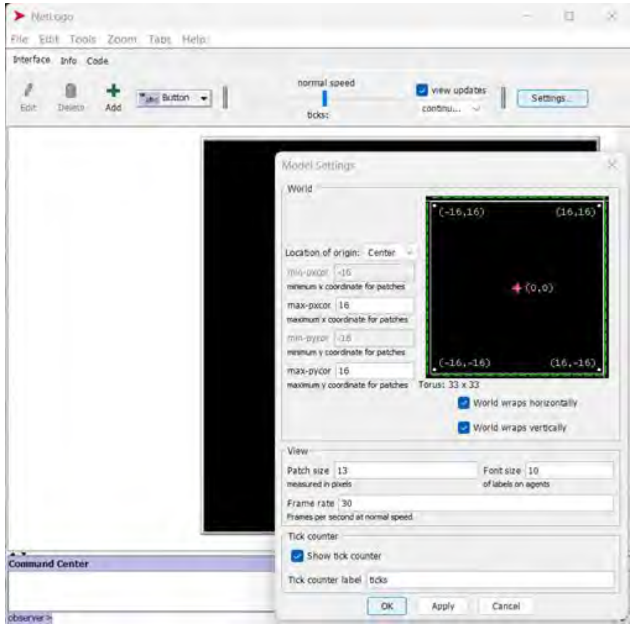
% unhappy 0

0-indexed

`list [a b c]`

show item 1 (list 1 2 4) > 2

first agent to appear has who 0



Location of origin: Center

min-pxcor -16
minimum x coordinate for patches

max-pxcor 16
maximum x coordinate for patches

min-pycor -16
minimum y coordinate for patches

max-pycor 16
maximum y coordinate for patches

Torus: 33 x 33

☑ World wraps horizontally

☑ World wraps vertically

Patch size 13
measured in pixels

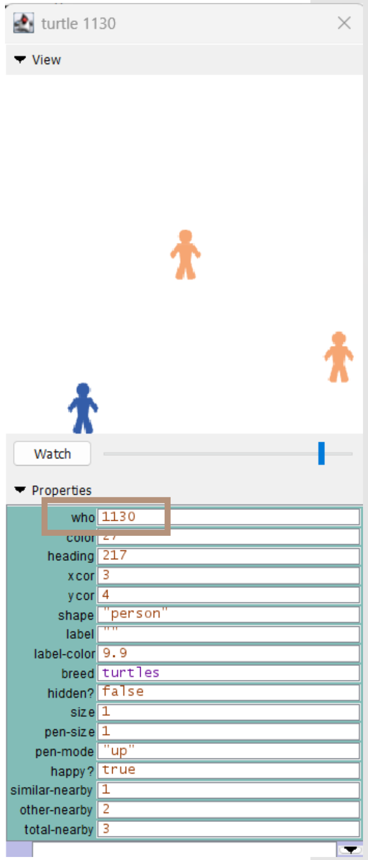
Font size 10
of labels on agents

Frame rate 30
Frames per second at normal speed

☑ Show tick counter

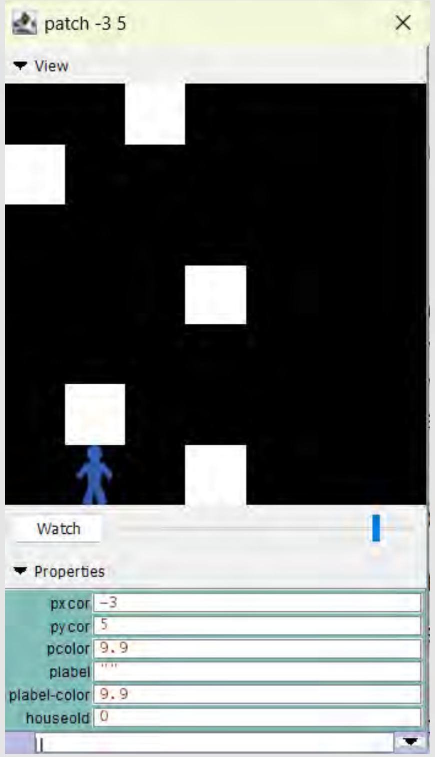
Tick counter label ticks

grid space world



who	1130
color	27
heading	217
xcor	3
ycor	4
shape	"person"
label	""
label-color	9.9
breed	turtles
hidden?	false
size	1
pen-size	1
pen-mode	"up"
happy?	true
similar-nearby	1
other-nearby	2
total-nearby	3

Agents are called **turtles** identified by **who** ID



pxcor	-3
pycor	5
pcolor	9.9
plabel	""
plabel-color	9.9
houseold	0

Grid cells are called **patches** and can interact as agents

NetLogo

simulated synchrony: agents execute commands asynchronously, but every agent acts knowing update in the model, based on coding

```
to ask dosomething
  ask turtles [do_A]
  ask turtles [do_B]
  ask turtles [do_C]
end
```

An agent in random order does A, then another agent does A. When all have done A, one random agent does B, then another does B. When all have done B, dosomething is executed

```
to ask dosomething
  ask turtles [
    do_A
    do_B]
end
```

An agent in random order does A then B, then another agent does A then B. When all have done A then B, dosomething is executed

```
to-report sumall [a b]
  report a + b
end
```

report 5 6 > 11

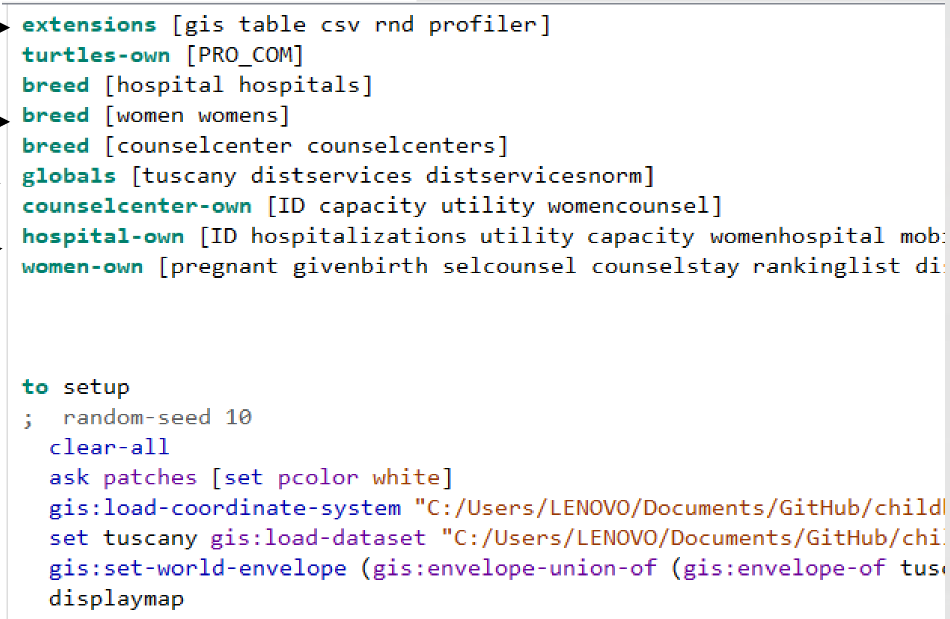
activate native extensions →

agent-class (breed) →

global variable →

agent-class level variable →

command block → that translates model components to be run



```
extensions [gis table csv rnd profiler]
turtles-own [PRO_COM]
breed [hospital hospitals]
breed [women womens]
breed [counselcenter counselcenters]
globals [tuscany distservices distservicesnorm]
counselcenter-own [ID capacity utility womencounsel]
hospital-own [ID hospitalizations utility capacity womenhospital mob
women-own [pregnant givenbirth selcounsel counselstay rankinglist di


to setup
;  random-seed 10
  clear-all
  ask patches [set pcolor white]
  gis:load-coordinate-system "C:/Users/LENOVO/Documents/GitHub/childl
  set tuscany gis:load-dataset "C:/Users/LENOVO/Documents/GitHub/chi
  gis:set-world-envelope (gis:envelope-union-of (gis:envelope-of tus
  displaymap
```

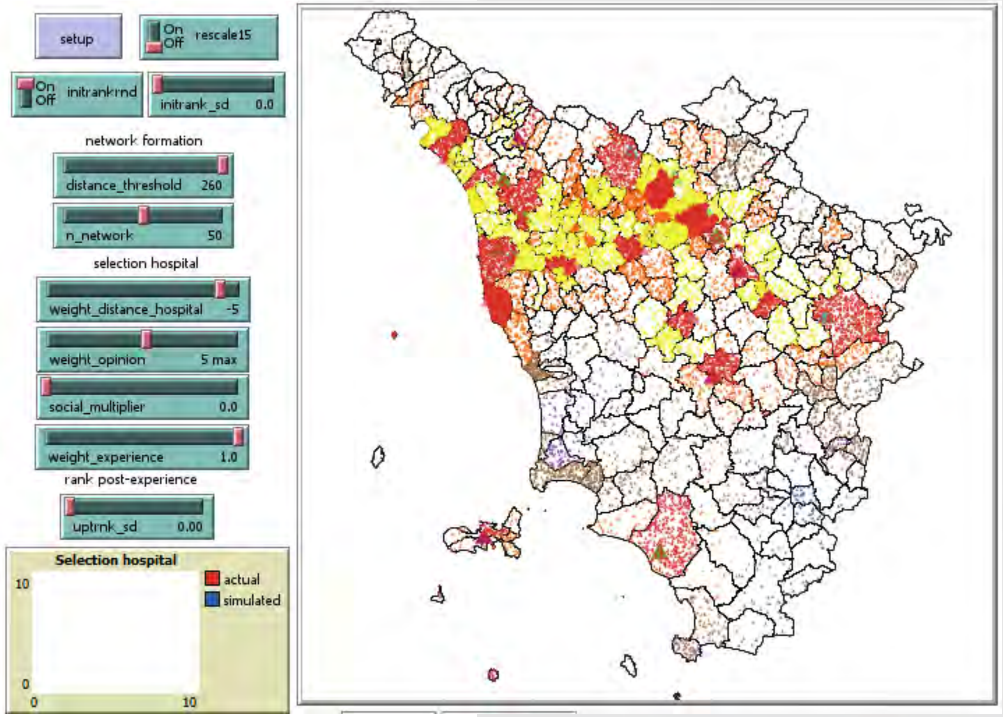
`mean, count, sort`	primitive reporters
`ask, set, forward`	primitive commands
`let h who`	local variable existing within a command block (to alleviate memory)

Let's implement in a case study...

Childbirth Mobilities: a Geo-Spatial Simulation Approach

- **Context**: While some determinants of hospital maternity selection are identified in the literature, the individual decisional processes, and social influence processes underlying the choice are unknown, neither specific data available.
- **Why ABM**: we can model the weights of preferences for hospital attributes at agents' micro-level, compare different social influence processes and compare how they replicate the data
- **Data available**: Mobility patterns in Tuscany 2023:
  - municipality residencies of women who gave birth

  (aggregated and anonymous)
  - municipality hospital where they gave birth
  - ranking of hospital (PNE performance indicator)
  - matrix of ditances
  - shapefile to map geographies to data



rescale15: Off
initrankrnd: On
initrank_sd 0.0

network formation

distance_threshold 260
n_network 50

selection hospital

weight_distance_hospital -5
weight_opinion 5 max
social_multiplier 0.0
weight_experience 1.0

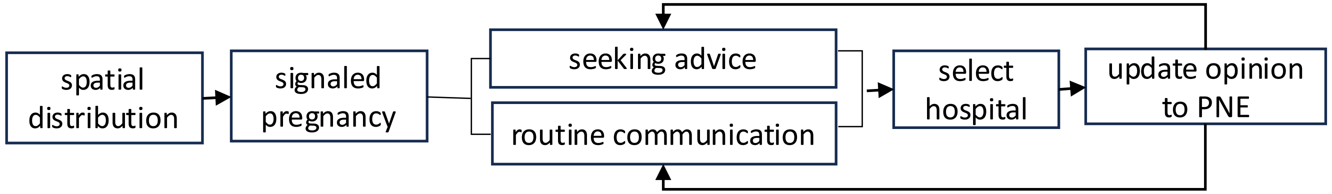
rank post-experience

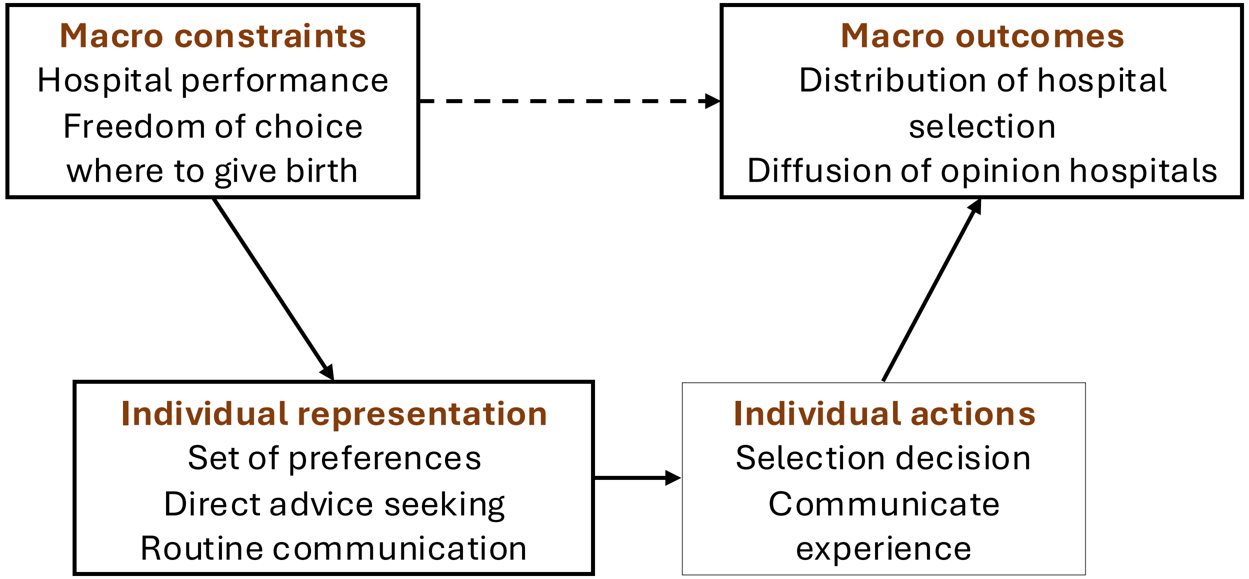
uptrnk_sd 0.00

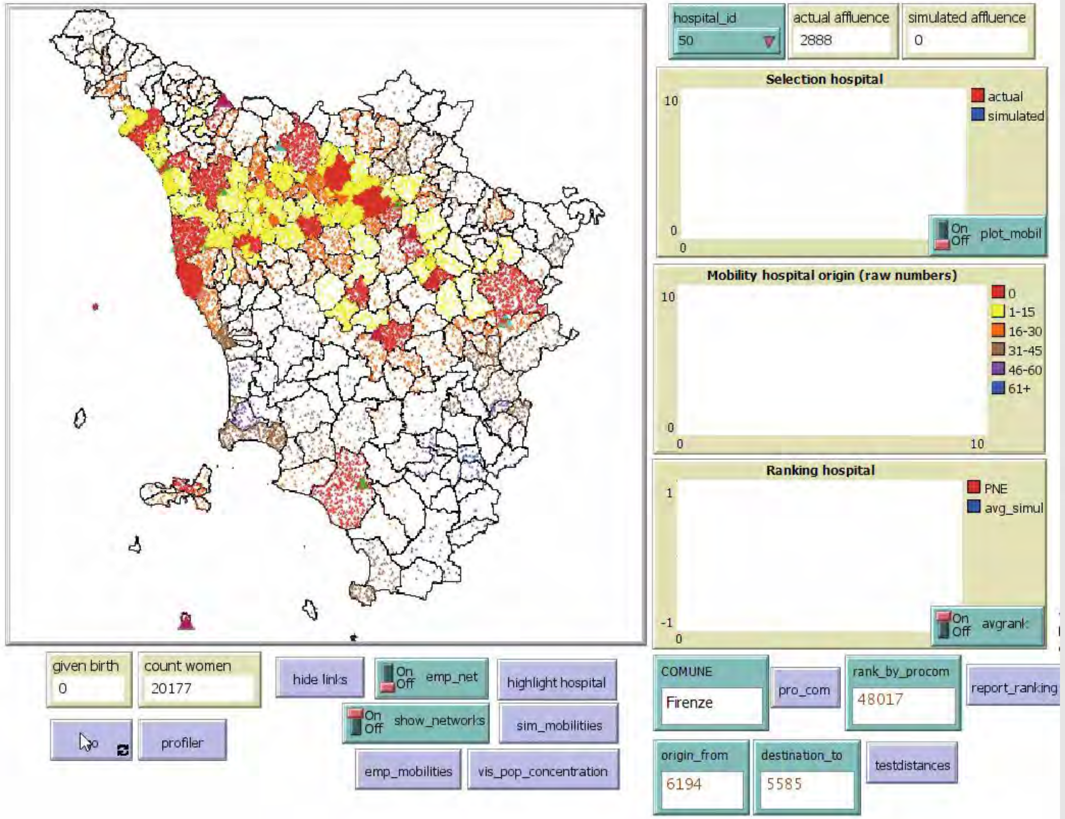
Goal

- Model a combination of individual decisional processes and social influence processes that can underline the selection of maternity hospital
- Which condition can best replicate the mobilities we observe?

Paolillo, Accordino, Pecoraro, https://github.com/RoccoPaolillo/childbirthnet/tree/MIE







hospital_id: 50

actual affluence: 2888

simulated affluence: 0

given birth: 0

count women: 20177

COMUNE: Firenze

rank_by_procom: 48017

origin_from: 6194

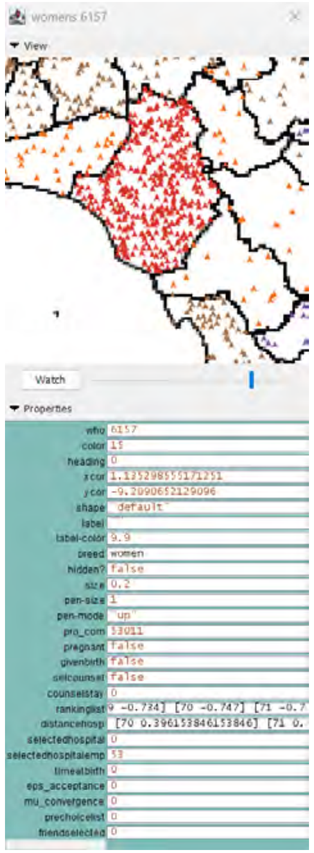
destination_to: 5585

Let’s compare with concepts presented in part one…

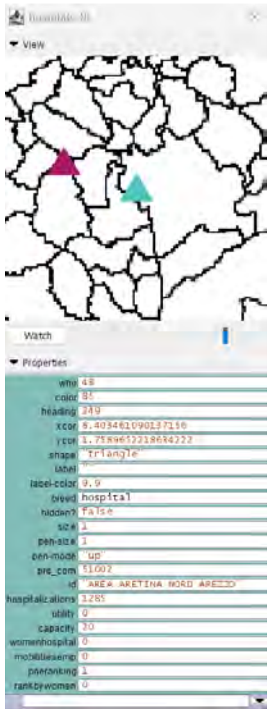
Agents – Who are the actors involved in the phenomenon?

women:

- they hold an initial random distribution of ranking opinion for each hospital in the region, when they become pregnant activate for choosing one hospital.
- they can expressely ask advice to friends in their municipality or base on common opinion of hospitals from routine communication
- after selecting one hospital, they can vehiculate the opinion of actual performance of hospital (PNE)



who	6157
color	15
heading	0
xcor	1.135298555171251
ycor	-9.2090652129096
shape	"default"
label	""
label-color	9.9
breed	women
hidden?	false
size	0.2
pen-size	1
pen-mode	"up"
pro_com	53011
pregnant	false
givenbirth	false
selcounsel	false
counselstay	0
rankinglist	9 -0.734] [70 -0.747] [71 -0.7
distancehosp	[70 0.396153846153846] [71 0.
selectedhospital	0
selectedhospitalemp	53
timeatbirth	0
eps_acceptance	0
mu_convergence	0
prechoicelist	0
friendselected	0



who	48
color	85
heading	249
xcor	6.403461090137156
ycor	1.7589612218634222
shape	"triangle"
label	""
label-color	9.9
breed	hospital
hidden?	false
size	1
pen-size	1
pen-mode	"up"
pro_com	51002
id	"AREA ARETINA NORD AREZZO"
hospitalizations	1285
utility	0
capacity	20
womenhospital	0
mobilitiesemp	0
pneranking	1
rankbywomen	0

hospital:

- PNE empirical ranking

Parameters – What are the conditions affecting the phenomenon?

```
to setup

  clear-all
  ask patches [set pcolor white]
  gis:load-coordinate-system "C:/../comuni_consultori_2019.prj"
  set tuscany gis:load-dataset "C:/../comuni_consultori_2019.shp"
  gis:set-world-envelope (gis:envelope-union-of (gis:envelope-of tuscany))
  displaymap

  set distservices csv:from-file "C:/../matrice_distanze_consultori.csv"
  set distservicesnorm csv:from-file "C:/../normalized_distance.csv«

  create-counselcenters
  create-hospitals
  create-womens

  let sorted-hospitals sort-by [[a b] -> [hospitalizations] of a >
[hospitalizations] of b] hospital

ask women [options_hospital]
  plot-hospitals

  reset-timer
  reset-ticks
end
```

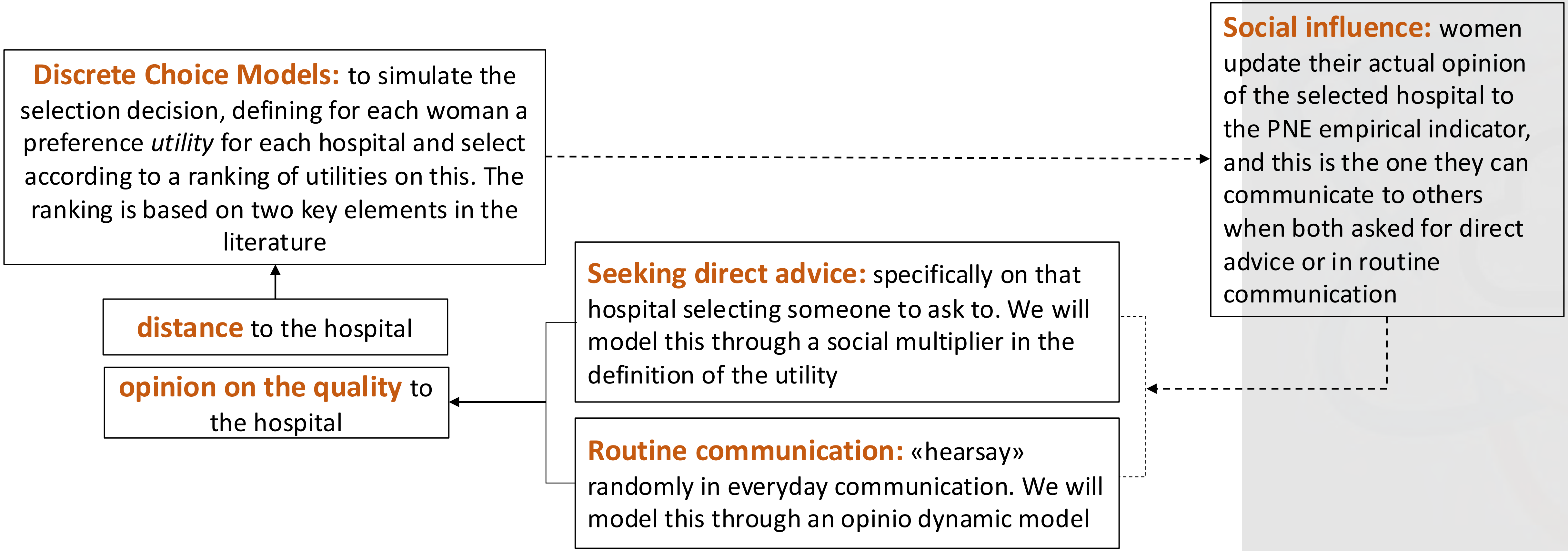
initialize the model with shapefiles

fetch data available

create agent classes

setup time schedule

Evolution - How does the order of agents' action influence each other?



Discrete Choice Models: to simulate the selection decision, defining for each woman a preference *utility* for each hospital and select according to a ranking of utilities on this. The ranking is based on two key elements in the literature

distance to the hospital

opinion on the quality to the hospital

Seeking direct advice: specifically on that hospital selecting someone to ask to. We will model this through a social multiplier in the definition of the utility

Routine communication: «hearsay» randomly in everyday communication. We will model this through an opinio dynamic model

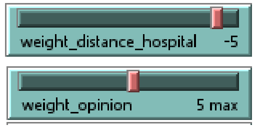
Social influence: women update their actual opinion of the selected hospital to the PNE empirical indicator, and this is the one they can communicate to others when both asked for direct advice or in routine communication

Parameters – What are the conditions affecting the phenomenon?

Discrete choice modeling: modeling the selection decision of agents, defining a utility (U) for each hospital h, based on a weight (parameter β) of how two characteristics of each hospital are relevant to the agent:
D_h: distance from the agent to the hospital
O_h: opinion on quality

The utility is used to define a probability to select that hospital *h* over the others hospital *k*

The higher is β, the more deterministic the selection is based on differences for attribute, the closer β is to 0, the more the selection is random ε



weight_distance_hospital -5

weight_opinion 5 max

We can input and manipulate the weight of each characteristic in the mind of agents

- equivalent to coefficients from regressions (clogit), not available
- test the consequences of combining different weights

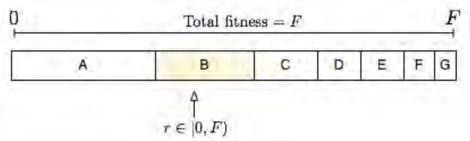
$$U_h = -\beta(D_h) + \beta(O_h) + \varepsilon$$

```
set utility ((weight_distance_hospital * 
(distancefrom*10)) +  (weight_opinion * opinionquality)  )
```

$$P_h = \frac{e^{(Uh - \max(Uk))}}{\sum e^{(U_k - \max(U_k))}}$$

```
set selectedhospital [who] of rnd:weighted-one-of hospital
[exp(utility - max [utility] of hospital)]
```

rnd:weighted-one-of *agentset reporter*



stable softmax to avoid numerical overflow

* 10 to harmonize ß of ranking and distance due to different scales

Parameters – What are the conditions affecting the phenomenon?

Social multiplier: A weight$\theta\,[0,1]$ in the definition of opinion quality O_h at the moment of decision.

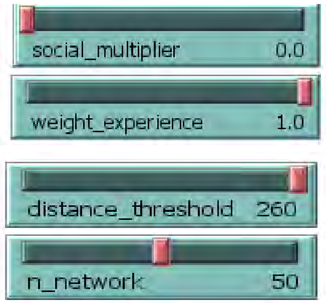
$\theta\ = 1$: the opinion quality completly aligns to that of people to whom asked for advice

$\theta\ =$0: the advice of others is not taken into consideration

We also included a weighted average to allocate different weights to friend who gave birth to that hospital (*p*), and whose opinion *o* is based on actual experience, and those who speak for hearsay (*a*)

a = 1 – w

w = 1, only those who gave birth influence



social_multiplier 0.0

weight_experience 1.0

distance_threshold 260

n_network 50

We can manipulate how influenced people will be by those they seek advice to

We can manipulate how many people advice is searched for advice and how far from hometown

$$\left(\frac{o_w + o_a + o_w + \cdots}{w + a + w + \cdots}\right)$$

```
foreach sort friends    [ z ->
  let weightfriend ifelse-value ([selectedhospital] of z = [who] of self)
  [weight_experience][(1 - weight_experience)]
  set totweightfriend lput weightfriend totweightfriend
  set ranking_othweight lput (table:get [rankinglist] of z [who] of self * weightfriend)
ranking_othweight]
```

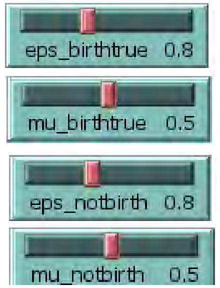
$$O_h = \ OwnOpinion_h + \theta\left(\left(\frac{o_w + o_a + o_w + \cdots}{w + a + w + \cdots}\right) - \ OwnOpinion_h\right)$$

```
set opinionquality [( opinionquality + social_multiplier *
((reduce + ranking_othweight / reduce + totweightfriend) - opinionquality ) )]
```

Parameters – What are the conditions affecting the phenomenon?

Opinion dynamics: a method to model the routine communication from women who gave birth to others in their municipality, spreading the own (updated) opinion a of that hospital. The receiver agent i accepts to listen if the distance between the own opinion of the hospital and that of the sender falls below a latitude of acceptance $|o_t^i - o_t^a| \leq \varepsilon$.
If so, the receiver aligns to the sender with convergence μ

ε = 0, not communication occurs
ε = 1, everyone is listened
μ = 0, not influence occurs
μ = 1, complete alignment occurs



eps_birthtrue 0.8
mu_birthtrue 0.5
eps_notbirth 0.8
mu_notbirth 0.5

We can manipulate how available to listen to those who gave birth and to what degree they will be influenced already in the *hearsay* routine communication

$$if\ |o_t^i - o_t^a| \leq \varepsilon$$
$$o_t^i = o_{t-1}^i + \mu(o_{t-1}^a - o_{t-1}^i)$$

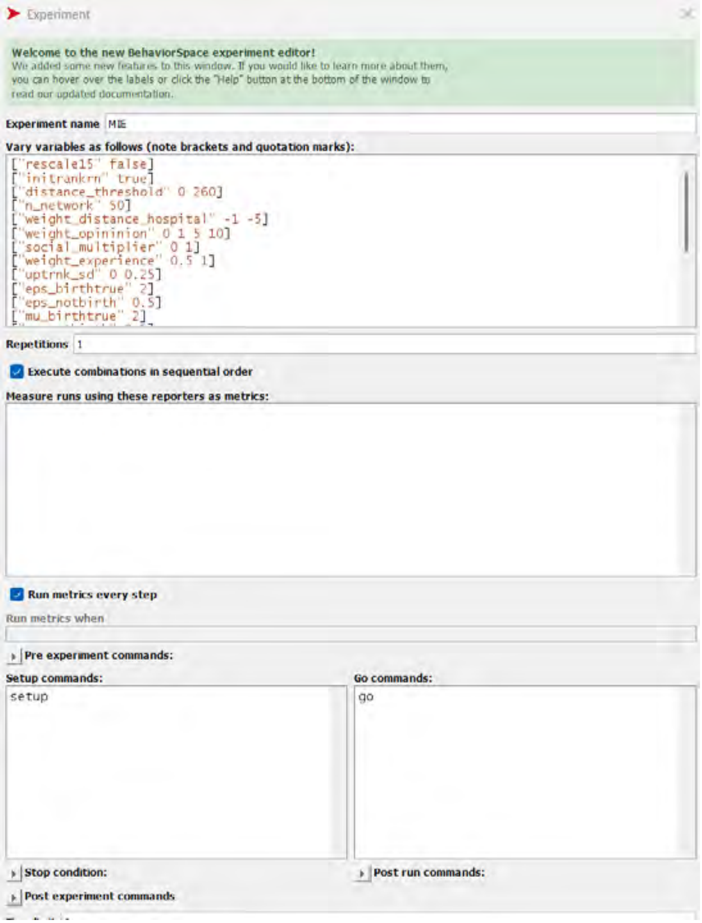
```
ask alter [
if abs(table:get rankinglist topic –
table:get [rankinglist] of myself topic) <= eps_acceptance
[table:put rankinglist topic
( table:get rankinglist topic +
(mu_convergence * (table:get [rankinglist] of myself topic –
table:get rankinglist topic)))]
```

We set to communicate every 80 time steps and to random 10% of women in municipality

Outcome – How can I identify how the conditions affect the phenomenon?

BehaviorSpace: a tool provided by NetLogo to set many experiments to run independently, setting the conditions for each parameter, define specific report measures, how many repetition wanted, and collect data in csv file

- Tools > BehaviorSpace
- Supports batch mode (headless)
- Better with a Server! For computational power, can run on laptop anyway



Experiment

Welcome to the new BehaviorSpace experiment editor!
We added some new features to this window. If you would like to learn more about them, you can hover over the labels or click the "Help" button at the bottom of the window to read our updated documentation.

Experiment name MIE

Vary variables as follows (note brackets and quotation marks):

```
["rescale15" false]
["initrankrn" true]
["distance_threshold" 0 260]
["n_network" 50]
["weight_distance_hospital" -1 -5]
["weight_opininion" 0 1 5 10]
["social_multiplier" 0 1]
["weight_experience" 0.5 1]
["uptrnk_sd" 0 0.25]
["eps_birthtrue" 2]
["eps_notbirth" 0.5]
["mu_birthtrue" 2]
```

Repetitions 1

☑ Execute combinations in sequential order

Measure runs using these reporters as metrics:

☑ Run metrics every step

Run metrics when

Pre experiment commands:

Setup commands:

```
setup
```

Go commands:

```
go
```

Stop condition:

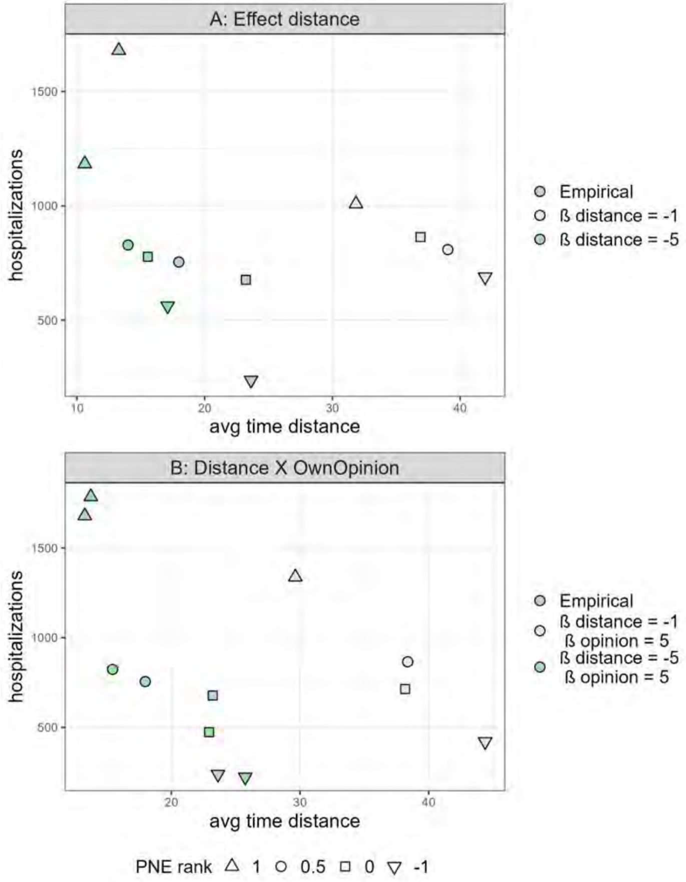
Post run commands:

Post experiment commands

['eps_birthtrue' 0 2] will run the conditions with the variable set 0 and 2

['eps_birthtrue' [0 0.1 2]] will run all the conditions with the variable set from 0 to 2 in increments 0.1 (e.g. 0 0.1 0.2...1.9 2)

Outcome – How can I identify how the conditions affect the phenomenon?

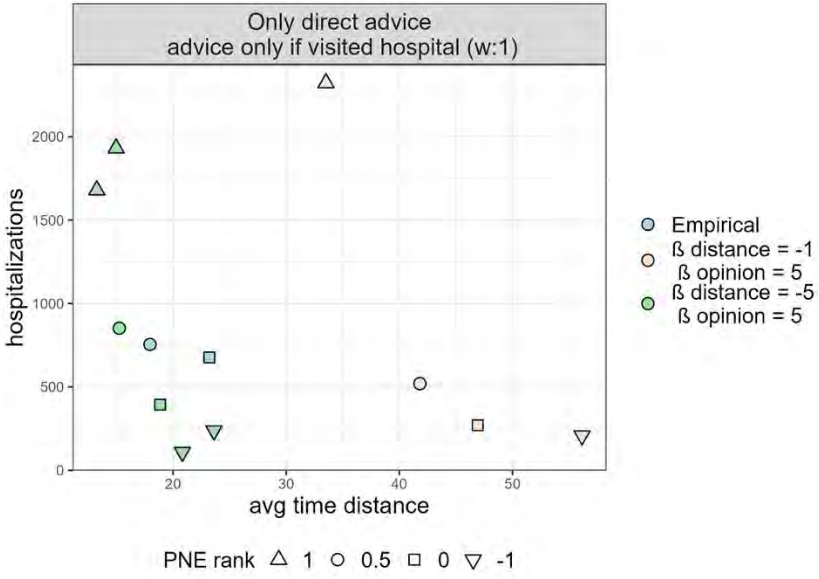


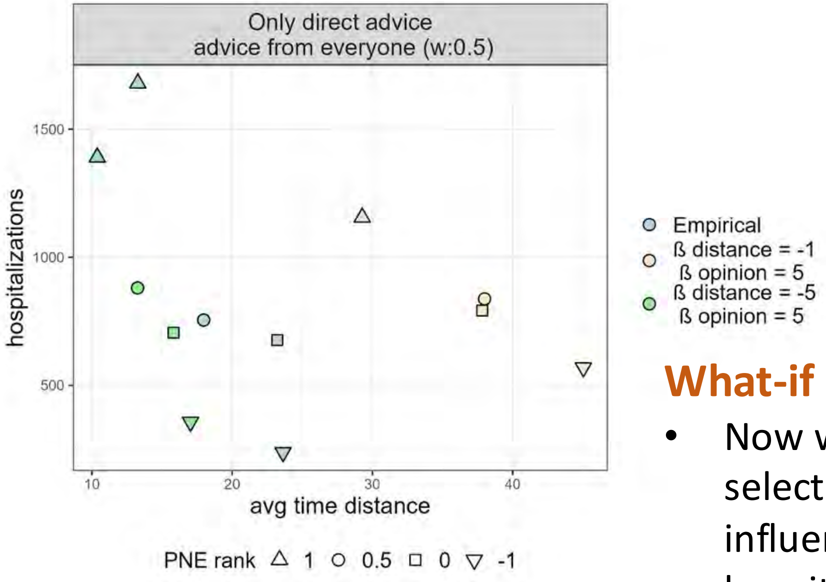
What-if scenarios

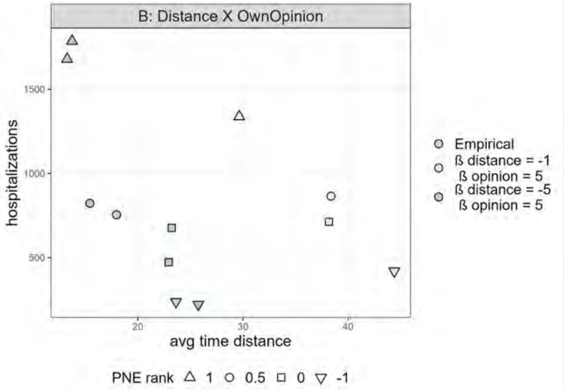
- Women communicate with everyone in routine communication (ε = 2), updating their opinion to the actual experience of those who gave birth, but they do not seek for advice (social multiplier θ = 0).
- We manipulate the weights for distance [0 -1,-5] and opinion quality updated via routine communication [0 1 5]
  - **condition A**: only distance matters (opinion weight 0)
    - with minimal weight of ditance (β = -1), women select hospitals more sparsely and difference between rankings do not emerge. Increasing weight distance (β = -5), the simulated distribution overestimates proximity of selected hospitals, and still sort by ranking not appearing
  - **condition B**: when we include also high opinion weight (β = 5), simulation results better approximate empirical data when coupled with high weight of distance (green condition β = -5, β = 5), both distance-wise and ranking-wise

What would be the effect of seeking advice then?

Outcome – How can I identify how the conditions affect the phenomenon?







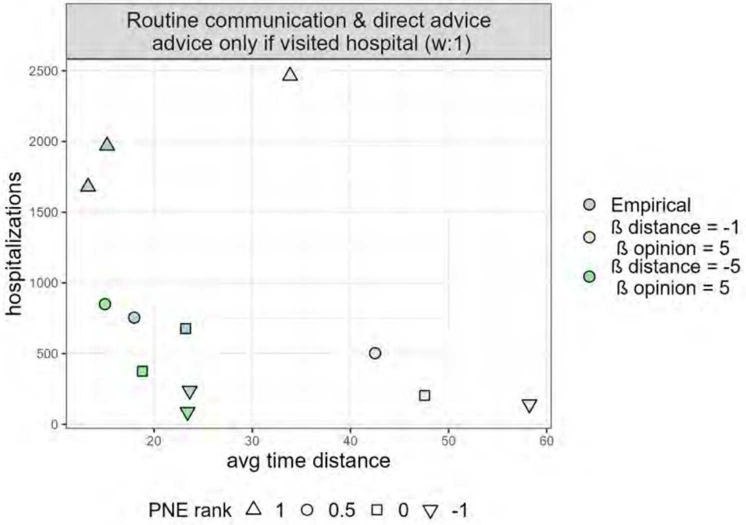
What-if scenarios

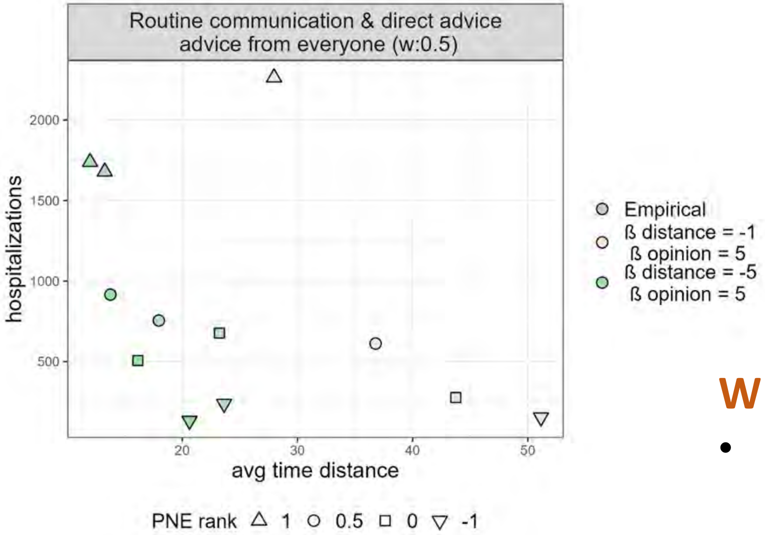
- Now women only seek for advice at the moment of selection (latitude opinion dynamic ε = 0), they can be influenced only by those who actually experience the hospital (*w* = 1) or by everyone equally (*w* = 0.5)

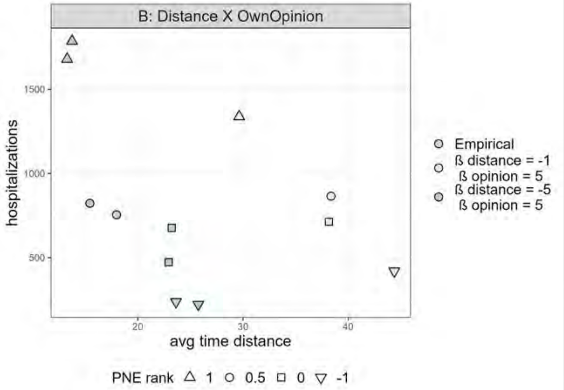
- Being influenced by those who gave birth (*w* = 1), hospitals with high PNE are overestimated, and more when the selection is more random by distance (β = -1)
- Being influenced with equal weight by everyone, those with actual experience and those with random opinion, underestimates the match with empirical data instead

What if we combine the two types of social influence?

Outcome – How can I identify how the conditions affect the phenomenon?







What-if scenarios

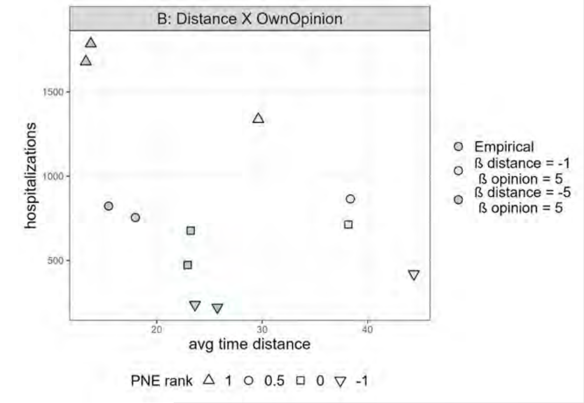
- Now women undergo both types of influence. When seeking advice, they can be influenced only by those who actually experience the hospital (*w* = 1) or by everyone equally (*w* = 0.5)

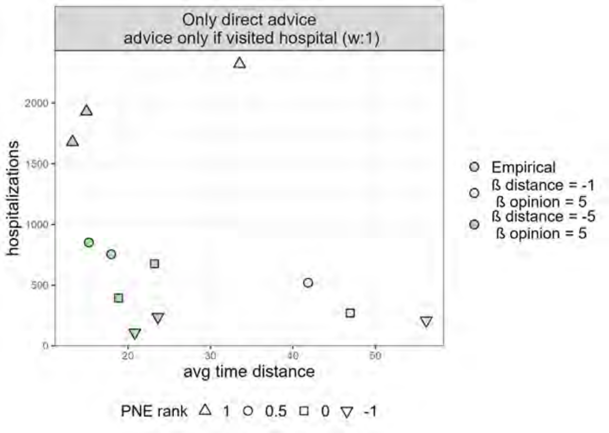
- Being influenced by those who gave birth (*w* = 1), still shows overestimation of hospitals with high, and more when the selection is more random by distance (β = -1)
- Being influenced with equal weight by everyone, those with actual experience and those who updated opinion by *hearsay* in common routine, better approximates the empirical data and reduces overestimation (sligthly)

Outcome – How can I read the evolution of the phenomenon?

Conclusions

- The best approximation to empirical mobilities is due to a combination of preference for shorter distance and high opinion quality. But high opinion quality with different weight of distance doesn’t produce the same effect. So, distance seems more relevant, and conditioning the diffusion of opinion updates
- Concerning the two modalities of social influence, seeking for advice would overestimate the effect of PNE ranking of hospitals, since the difference in opinion quality becomes more salient. The effect is higher if agents relocate randomly in space, probably because more likely to find high PNE hospitals, that are more and in more populated areas.
- Being exposed to different opinions when seeking advice and in combination with routine communication ameliorates the polarization effect of ranking coming closer to the empirical data, and routine communication seems to suffice





Outcome - How the phenomenon emerges through agents' interaction?

Limits and Next Steps

- To better understand the actual differentiation between seeking advice modality and routine communication, looking at the evolution through time and wider parameter space
- To differentiate action of women and measures considering the actual microspace within the region

BUT

- Overall, we had quite amount of data here
- Sometimes information on sociodemographic population might be missing, how could we do?
- Synthetic Populations

Let's see what synthetic populations are...

Thank you! Questions?

rocco.paolillo@cnr.it
@roccopaolillo.bsky.social

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union– NextGenerationEU under NRRP Grant agreement n. MUR IR0000008

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Agent-based Modeling and Synthetic Populations for Social Research

Rocco Paolillo

CNR-IRPPS

Rome, January, 30° 2026, Italy

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Part Three

A service for synthetic populations extraction...

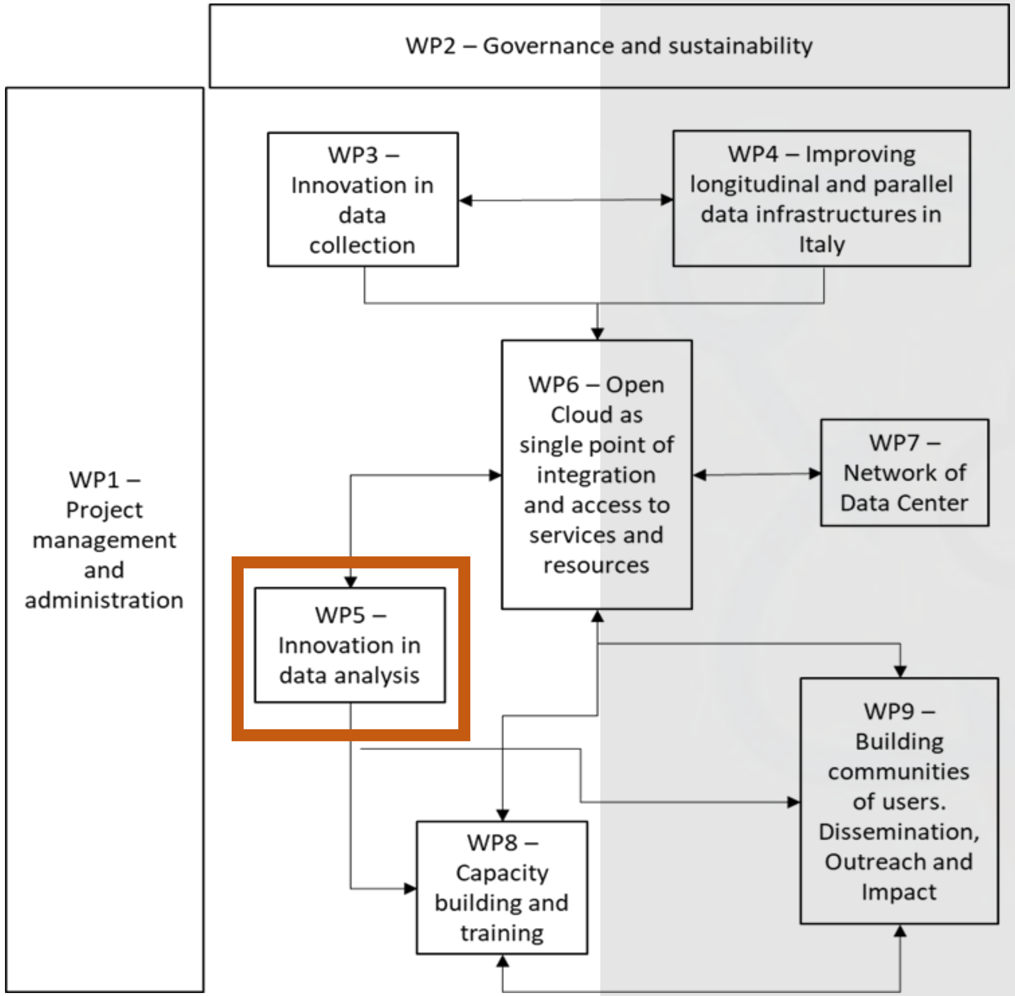
FOSSR

Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Creation of an ***Italian Open Science Cloud for the Social Sciences*** guided by ***Open Science*** principles

which shall provide **innovative tools and services** to investigate issues related to the **economic and societal change of contemporary societies** through the enhancement of **research infrastructures**

https://www.fossr.eu/



WP2 – Governance and sustainability

WP1 – Project management and administration

WP3 – Innovation in data collection

WP4 – Improving longitudinal and parallel data infrastructures in Italy

WP6 – Open Cloud as single point of integration and access to services and resources

WP7 – Network of Data Center

WP5 – Innovation in data analysis

WP8 – Capacity building and training

WP9 – Building communities of users. Dissemination, Outreach and Impact

Luciana Taddei
Mario Paolucci *Editors*

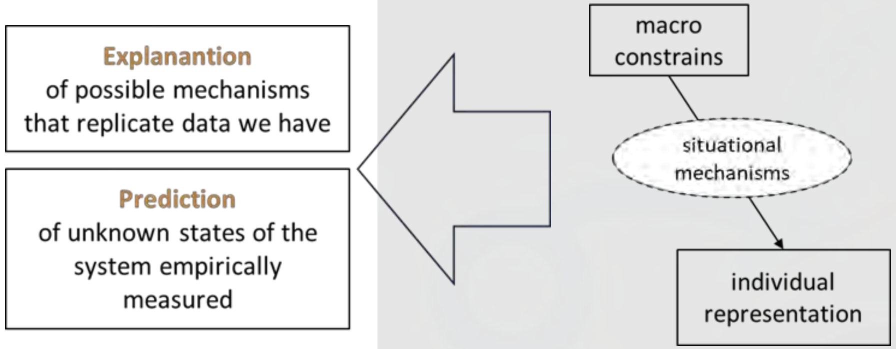
Longitudinal Data Infrastructures in Europe

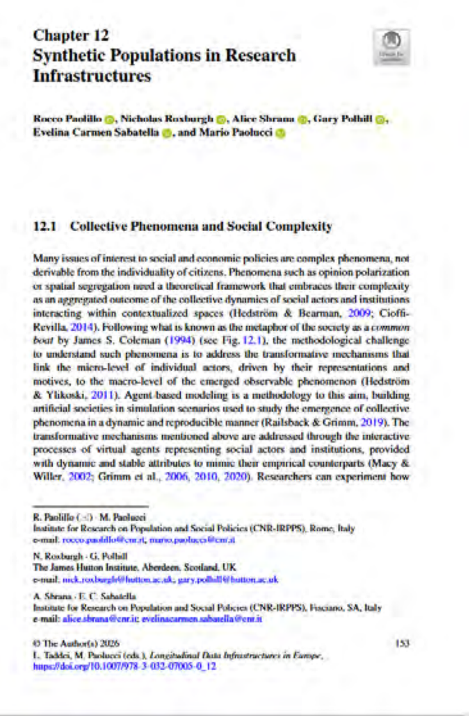
Tools for Open Science in Social Science Research

OPEN ACCESS

Springer

Ex-ante policy analysis evaluation
Post-ante policy analysis evaluation
Conterfactual policy scenarios


Explanantion
of possible mechanisms that replicate data we have
Prediction
of unknown states of the system empirically measured
macro constrains
situational mechanisms
individual representation



Chapter 12
Synthetic Populations in Research Infrastructures

Rocco Paolillo, Nicholas Roxburgh, Alice Sbrana, Gary Polhill, Evelina Carmen Sabatella, and Mario Paolucci

12.1 Collective Phenomena and Social Complexity

Many issues of interest to social and economic policies are complex phenomena, not derivable from the individuality of citizens. Phenomena such as opinion polarization or spatial segregation need a theoretical framework that embraces their complexity as an *aggregated* outcome of the collective dynamics of social actors and institutions interacting within contextualized spaces (Hedström & Bearman, 2009; Cioffi-Revilla, 2014). Following what is known as the metaphor of the society as a *common boat* by James S. Coleman (1994) (see Fig. 12.1), the methodological challenge to understand such phenomena is to address the transformative mechanisms that link the micro-level of individual actors, driven by their representations and motives, to the macro-level of the emerged observable phenomenon (Hedström & Ylikoski, 2011). Agent-based modeling is a methodology to this aim, building artificial societies in simulation scenarios used to study the emergence of collective phenomena in a dynamic and reproducible manner (Railsback & Grimm, 2019). The transformative mechanisms mentioned above are addressed through the interactive processes of virtual agents representing social actors and institutions, provided with dynamic and stable attributes to mimic their empirical counterparts (Macy & Willer, 2002; Grimm et al., 2006, 2010, 2020). Researchers can experiment how

R. Paolillo (✉) · M. Paolucci
Institute for Research on Population and Social Policies (CNR-IRPPS), Rome, Italy
e-mail: rocco.paolillo@cnr.it; mario.paolucci@cnr.it

N. Roxburgh · G. Polhill
The James Hutton Institute, Aberdeen, Scotland, UK
e-mail: nick.roxburgh@hutton.ac.uk; gary.polhill@hutton.ac.uk

A. Sbrana · E. C. Sabatella
Institute for Research on Population and Social Policies (CNR-IRPPS), Fisciano, SA, Italy
e-mail: alice.sbrana@cnr.it; evelinacarmen.sabatella@cnr.it

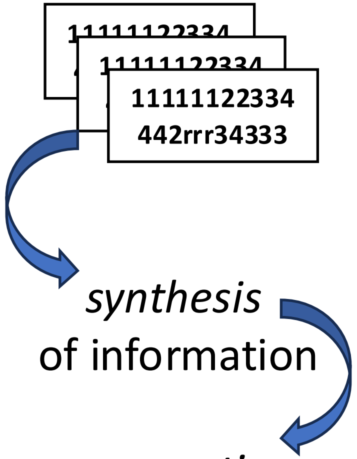
© The Author(s) 2026 153
L. Taddei, M. Paolucci (eds.), *Longitudinal Data Infrastructures in Europe*,
https://doi.org/10.1007/978-3-032-07005-0_12

- An artificial society is a stylized social system where to study the mechanisms of the phenomenon
    - Especially when agent-based modeling is used for policy purposes, the mechanisms observed need to be bounded to the conditions of the system they want to operate in, e.g. Digital Twin systems
    - The system needs to be a synthesis of the information available of the target society

- Challenge to identify attributes at micro-level:
  - data not collected
  - separate datasets
  - privacy issues

Synthetic populations: a series of techniques to handle available data and replicate attributes of the target population

‘While a synthetic population is implicitly an artificial population, an artificial population is not necessarily a synthetic population’

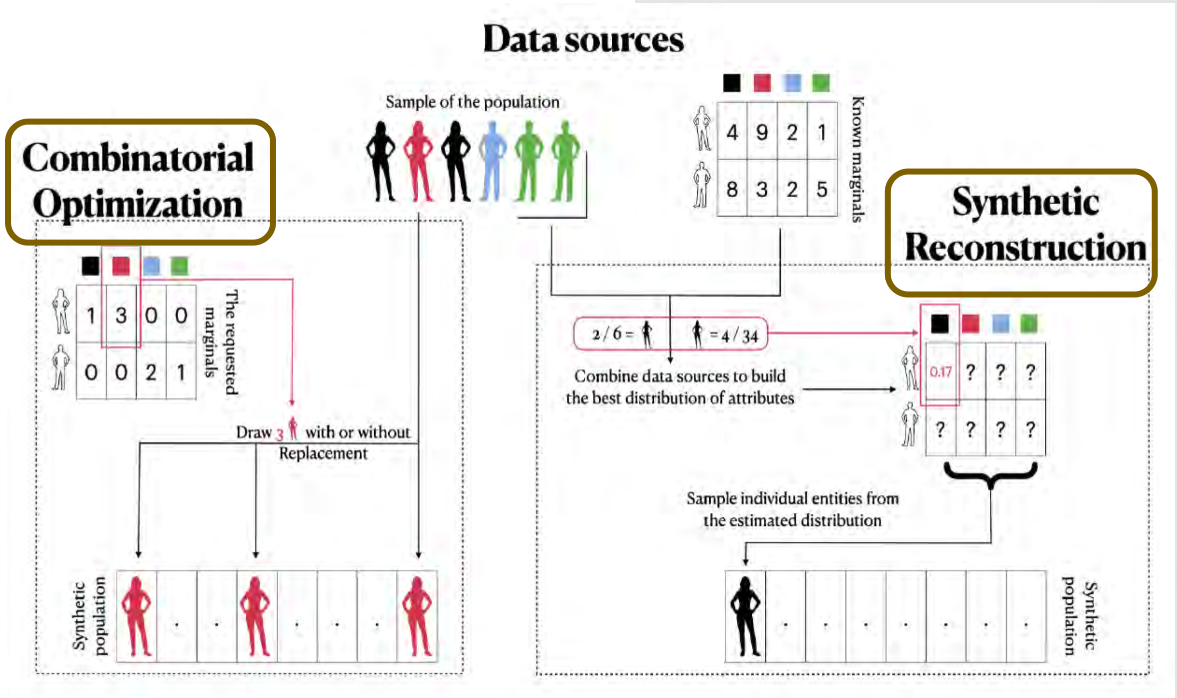


synthesis
of information

generation
of synthetic data

Bigi et al. (2022)

comparison
with available data



Chapuis et al. (2022)

Machine Learning

Every method has its peculiarities and boundaries not set in stone

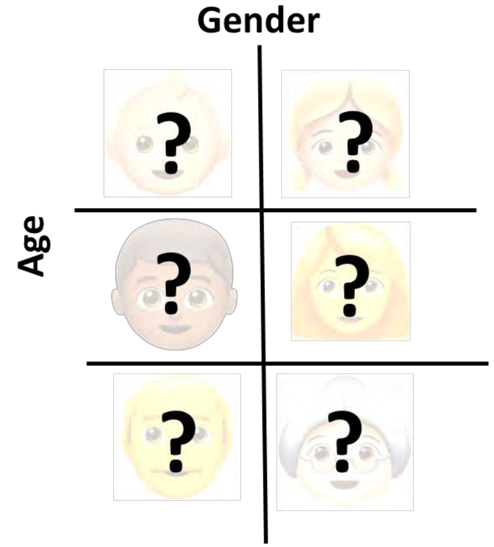
Synthetic Reconstruction

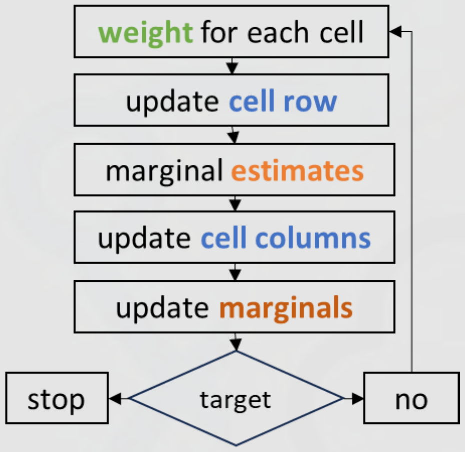
- Focus on estimate of unknown joint distributions from data available from marginal distribution
- macro to micro

Extensions

- **Multiple Iterative Proportional Fitting (MIPF)**
  taking the estimated joints as marginal to next step dimension
- **Hierarchical Iterative Proportional Fitting (HIPF)**
  nested data fitting marginals from one level (e.g. household) to the narrow (e.g. citizens) (Yamaego et al., 2021)
- **Iterative Proportional Updating (IPU)**: from micro data finds weight for cross-category multiplied by marginals and correct backwards -> combinatorial optmization

- Archetype in the **Iterative Proportional Fitting** (IPF, raking)





$$\text{weight} = \frac{\text{observed marginal}}{\text{fitted marginal}}$$

✓ mathematically transparent, robust, focus on weights

🟥 require ad hoc setting of algorithms, zero-cell problem

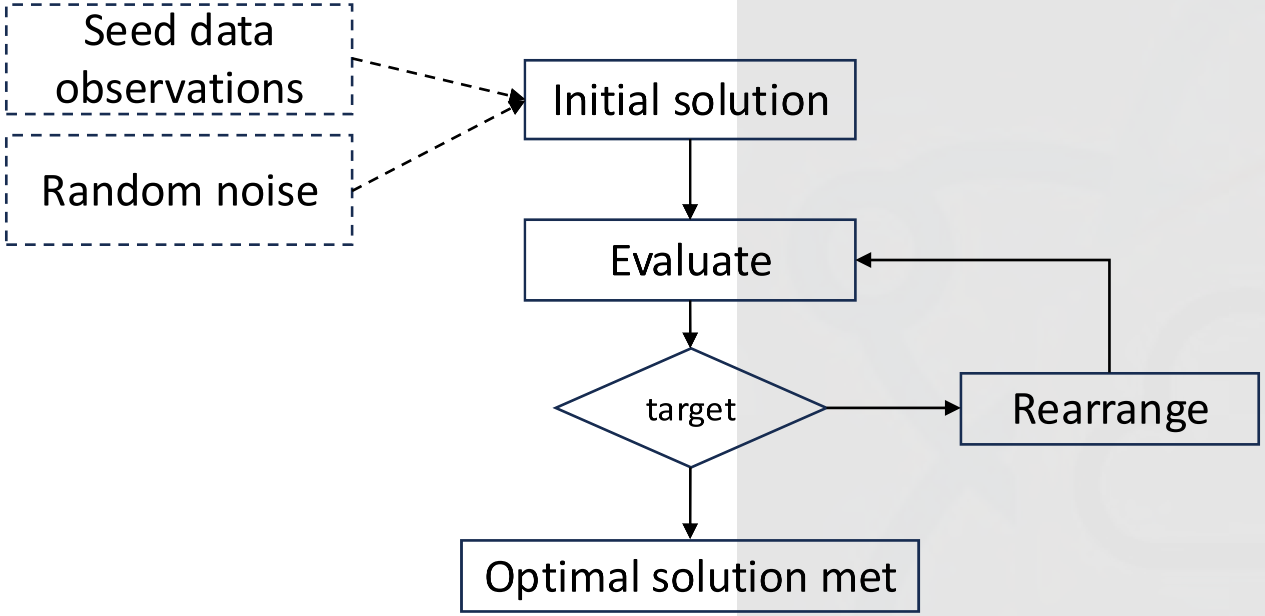
FOSSR:SPG

R::synthpop (MIPF)

Combinatorial Optimization

- Scaling of the synthetic population
- Draw random data and optimize against marginals observed

  - **Simulated Annealing (SA)**
    from the micro-data identifies some seed numbers and compares synthetic marginas to empirical marginals to correct backwards
  - **Markov Chain Monte Carlo (MCMC)**
    random number generations for all intersection and correct backwards



✓ overcome some setbacks of IPF family

✓ better with complex intersections (handled *at once*)

🟥 focus on the outcome rather than weights & inner conditions

🟥 computationally more demanding

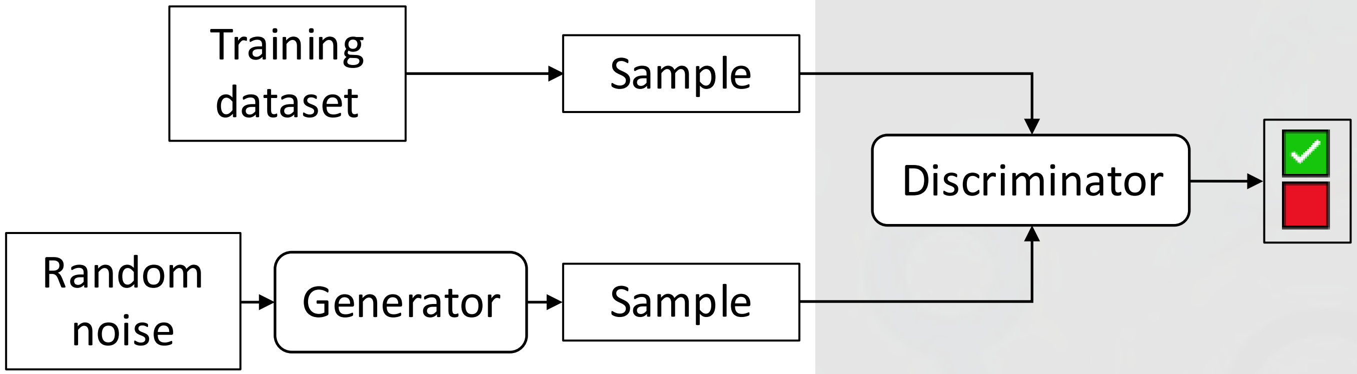
py::simmaneal

R::MCMCpack py::PyMC

Machine Learning Approach

- Most recent in time
- Statistical learning
- micro to macro

- **Generative Adversarial Networks (GAN)**
  - originated from images AI, applied to data
  - two competing (neural) networks:
    - **Generator** who produces random data
    - **Discriminator** that discriminate realistic data from not realistic
    - Goal of Generator is to get better to *deceive* the Discriminator who gets better in discriminate, meaning that synthetic data are very realistic



✅ promising because they integrate the performance of combinatorial methods with transparency, multidimensionally compared to IPF family

🟥 they require a training dataset from which the learning process occurs, where underrepresented groups are likely to be ignored in estimates

- Pre-adjust the training dataset with ad hoc weights to marginals (Falck, 2025)
- Post-adjust the synthetic outcome with weights to marginals

py::PyTorch

FOSSR
Fostering Open Science in Social Science Research

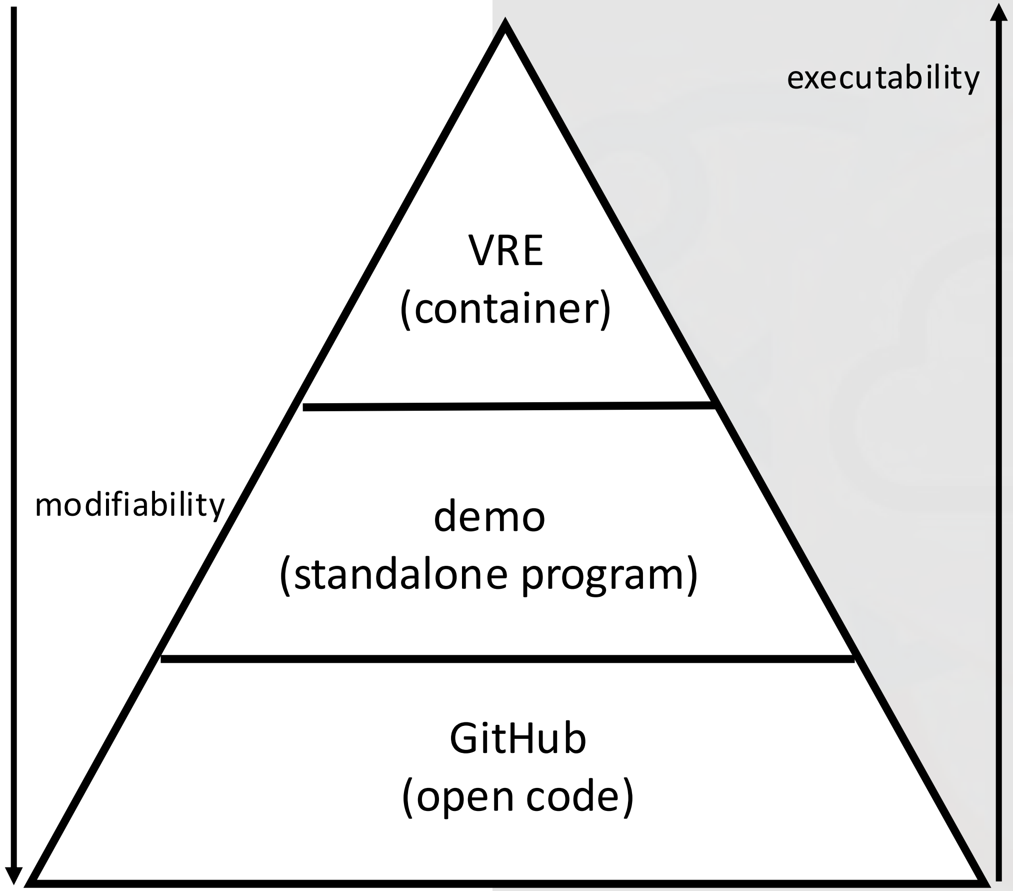
Synthetic Populations Generator (SPG)

Service to enable researchers and policy makers to extract synthetic populations at desired level of information from input dataset

- agent-based modeling
- spatial analysis
- conditional model
- ...

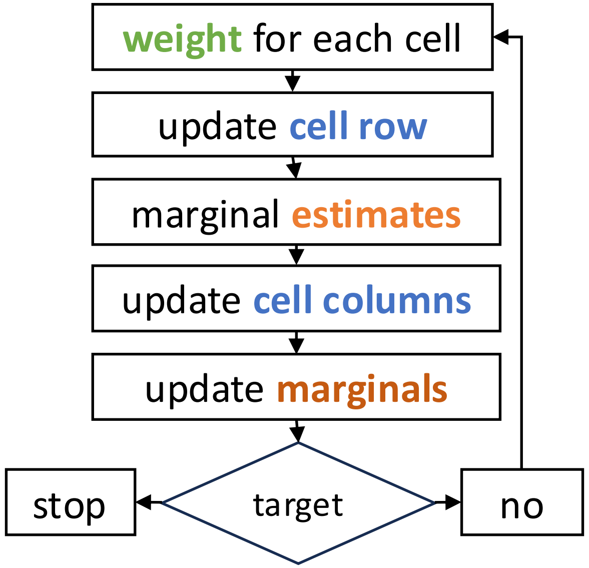
Open Science & Source Software

https://github.com/RoccoPaolillo/IPF_multidim.git > synthpopgen.py



(Jimenez et al., 2017; Hong et al., 2022)

Iterative Proportional Fitting (IPF)



weight for each cell

update cell row

marginal estimates

update cell columns

update marginals

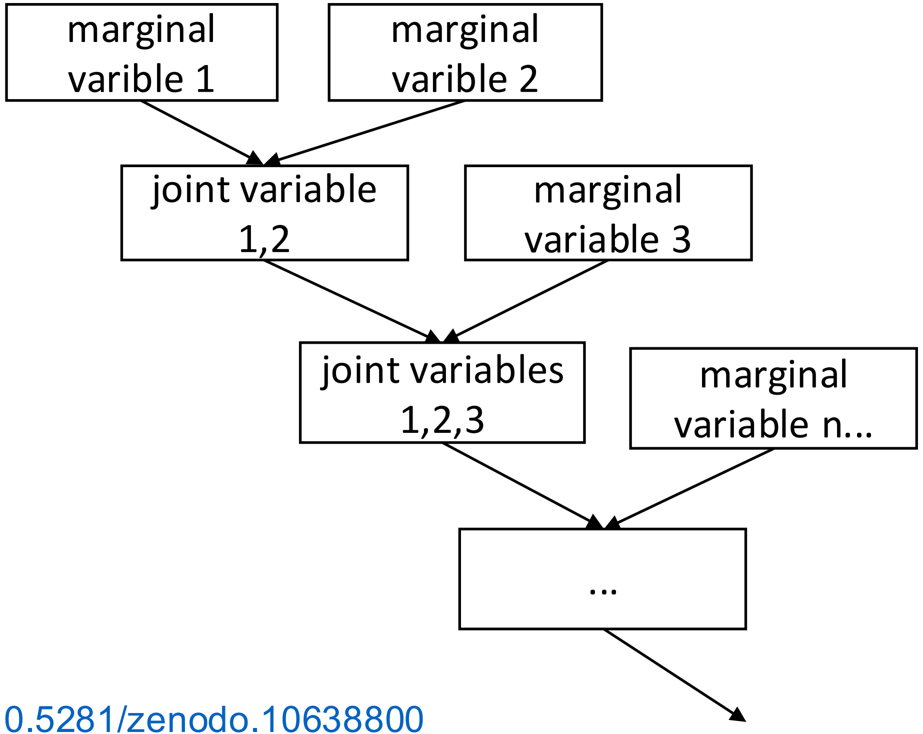
target

stop

no

$$\text{weight} = \frac{\text{observed marginal}}{\text{fitted marginal}}$$

Multiple Iterative Proportional Fitting (MIPF)



marginal varible 1

marginal varible 2

joint variable 1,2

marginal variable 3

joint variables 1,2,3

marginal variable n...

...

Selected for higher transparency, robustness, light computing, core mechanism common to other method (somehow), but provided more service-oriented direction is guaranteed

First version released (ISTAT gender X age for validation) 10.5281/zenodo.10638800

Aims of Synthetic Populations Generator (SPG)

- Include **multidimensionality**
- Increase **generalizability** of variable handling
- Enable **automation** input-execution-output
- Customize **filtering** selection

- leverage estimate of joint and conditional probability over in-cell weight iteration

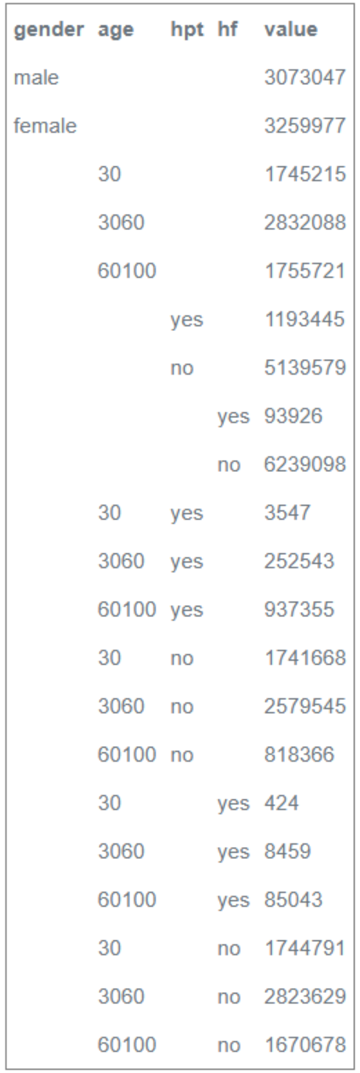
Tested with **opensalute Lazio** health data:

- gender
- age
- hyptertension (hpt)
- heart failure (hf)

known joints:

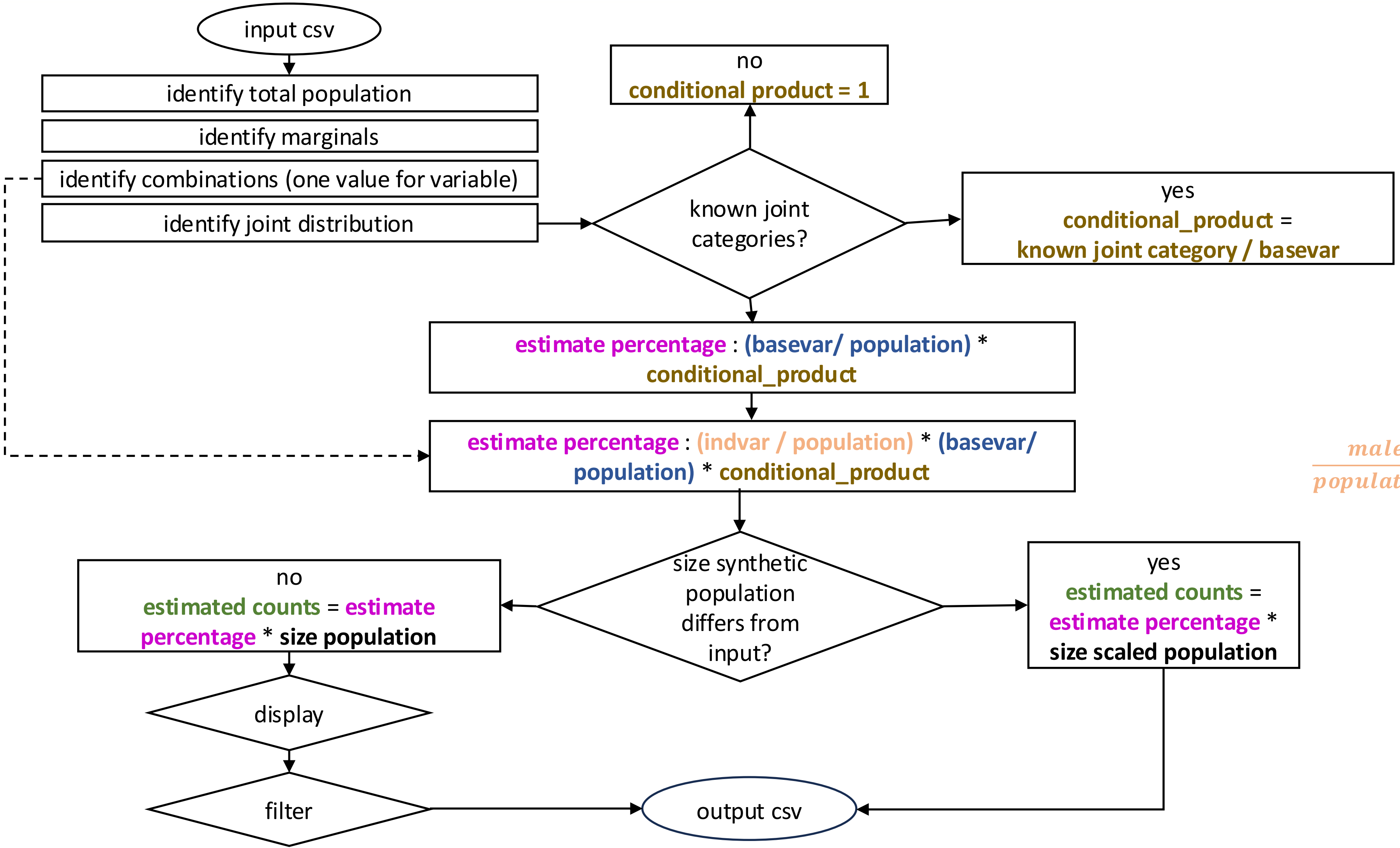
- hypertension over age
- heart failure over age

goal of service: identify joint distribution for all combinations



gender	age	hpt	hf	value
male				3073047
female				3259977
	30			1745215
	3060			2832088
	60100			1755721
		yes		1193445
		no		5139579
			yes	93926
			no	6239098
	30	yes		3547
	3060	yes		252543
	60100	yes		937355
	30	no		1741668
	3060	no		2579545
	60100	no		818366
	30		yes	424
	3060		yes	8459
	60100		yes	85043
	30		no	1744791
	3060		no	2823629
	60100		no	1670678

The algorithm



input csv

identify total population

identify marginals

identify combinations (one value for variable)

identify joint distribution

known joint categories?

no
conditional product = 1

yes
conditional_product = known joint category / basevar

estimate percentage : (basevar/ population) * conditional_product

estimate percentage : (indvar / population) * (basevar/ population) * conditional_product

size synthetic population differs from input?

no
estimated counts = estimate percentage * size population

yes
estimated counts = estimate percentage * size scaled population

display

filter

output csv

Estimate percentageM30HPTHF
gender male, age 30, hptyes, hf yes

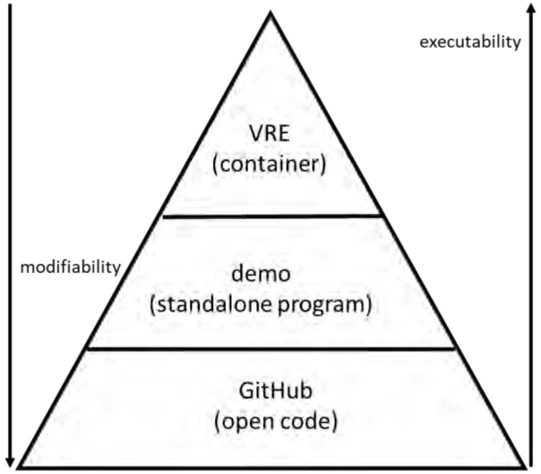
$$\frac{male}{population} * \left(\frac{age30}{population} * \frac{hptyes, age30}{age30} * \frac{hfno, age30}{age30}\right)$$

Estimate countM30HPTHF
Estimate percentageM30HPTHF *
size synthetic population

Open code

https://github.com/RoccoPaolillo/IPF_multidim.git > synthpopgen.py

cmd line



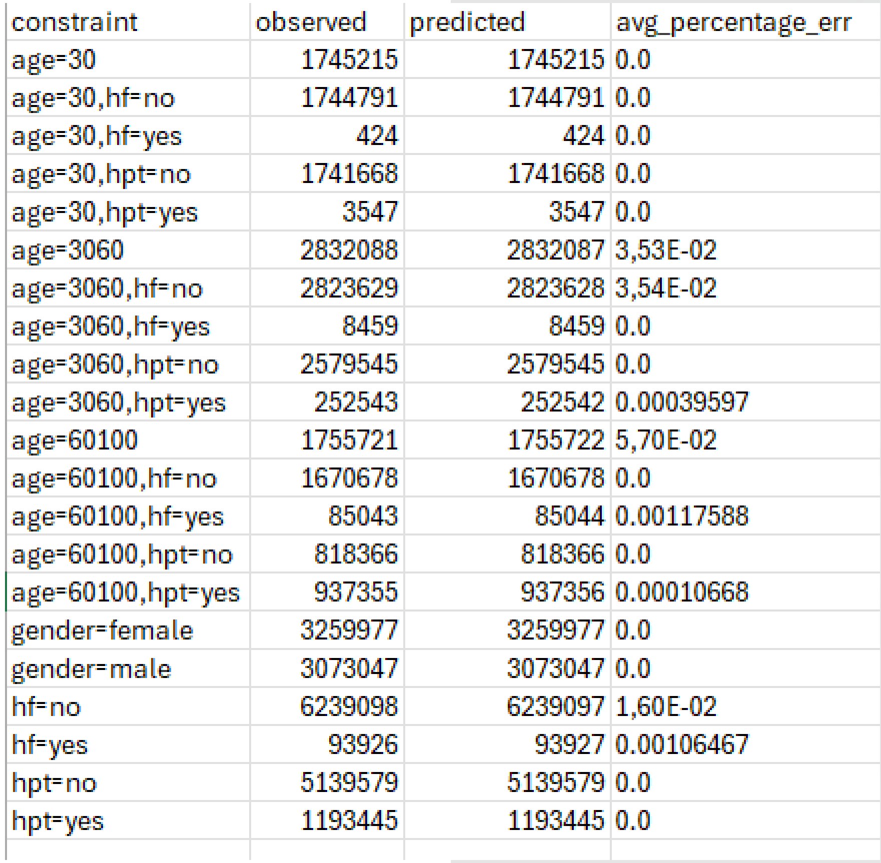
```
python synthpopgen.py -i input_file_tuples.csv \
  -f (filter)
    'all'
    'gender:female,age:3060' \
-d (display)
    'split'
    'aggregate'\
- v (validate)*
--synth-total 20303*
-o results.csv
```

→ pro: high modifiability for users(rewrite, retest...)

→ vs: knowledge coding, dependencies

measure of validation

Average percentage error between input marginals (and joints) data and stimated marginals, **RMSE 0.6**

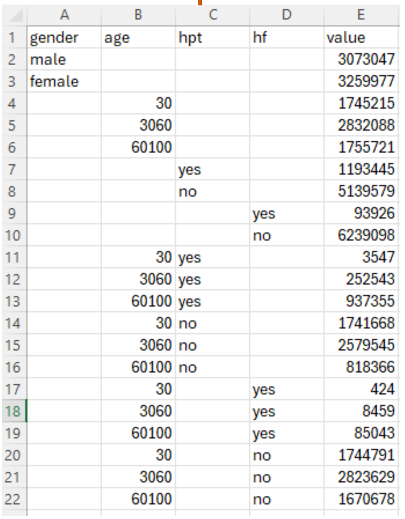


constraint	observed	predicted	avg_percentage_err
age=30	1745215	1745215	0.0
age=30,hf=no	1744791	1744791	0.0
age=30,hf=yes	424	424	0.0
age=30,hpt=no	1741668	1741668	0.0
age=30,hpt=yes	3547	3547	0.0
age=3060	2832088	2832087	3,53E-02
age=3060,hf=no	2823629	2823628	3,54E-02
age=3060,hf=yes	8459	8459	0.0
age=3060,hpt=no	2579545	2579545	0.0
age=3060,hpt=yes	252543	252542	0.00039597
age=60100	1755721	1755722	5,70E-02
age=60100,hf=no	1670678	1670678	0.0
age=60100,hf=yes	85043	85044	0.00117588
age=60100,hpt=no	818366	818366	0.0
age=60100,hpt=yes	937355	937356	0.00010668
gender=female	3259977	3259977	0.0
gender=male	3073047	3073047	0.0
hf=no	6239098	6239097	1,60E-02
hf=yes	93926	93927	0.00106467
hpt=no	5139579	5139579	0.0
hpt=yes	1193445	1193445	0.0

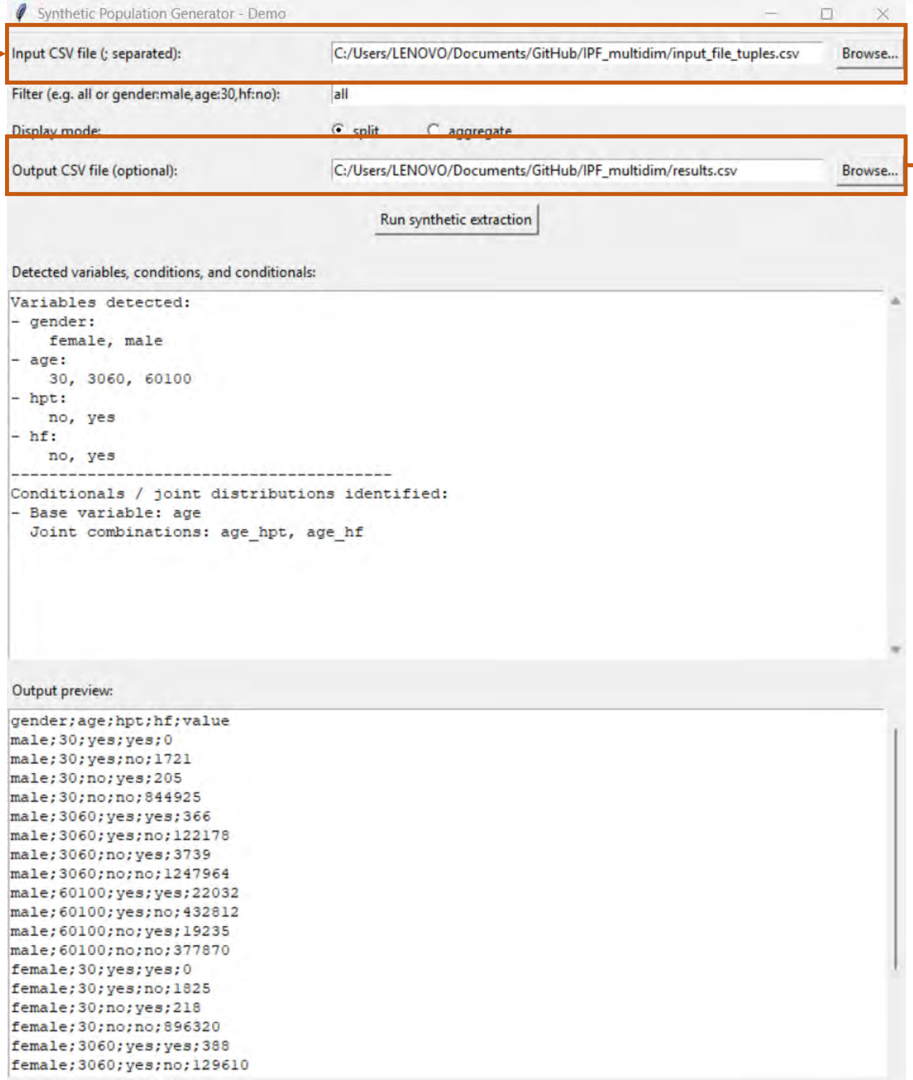
* only if whole population combinations are stimated (-f all)

Standalone program

.exe local file
py:: tkinter



	A	B	C	D	E
1	gender	age	hpt	hf	value
2	male				3073047
3	female				3259977
4		30			1745215
5		3060			2832088
6		60100			1755721
7			yes		1193445
8			no		5139579
9				yes	93926
10				no	6239098
11		30	yes		3547
12		3060	yes		252543
13		60100	yes		937355
14		30	no		1741668
15		3060	no		2579545
16		60100	no		818366
17		30		yes	424
18		3060		yes	8459
19		60100		yes	85043
20		30		no	1744791
21		3060		no	2823629
22		60100		no	1670678



Synthetic Population Generator - Demo

Input CSV file (; separated): C:/Users/LENOVO/Documents/GitHub/IPF_multidim/input_file_tuples.csv

Filter (e.g. all or gender:male,age:30,hf:no): all

Display mode: ● split ○ aggregate

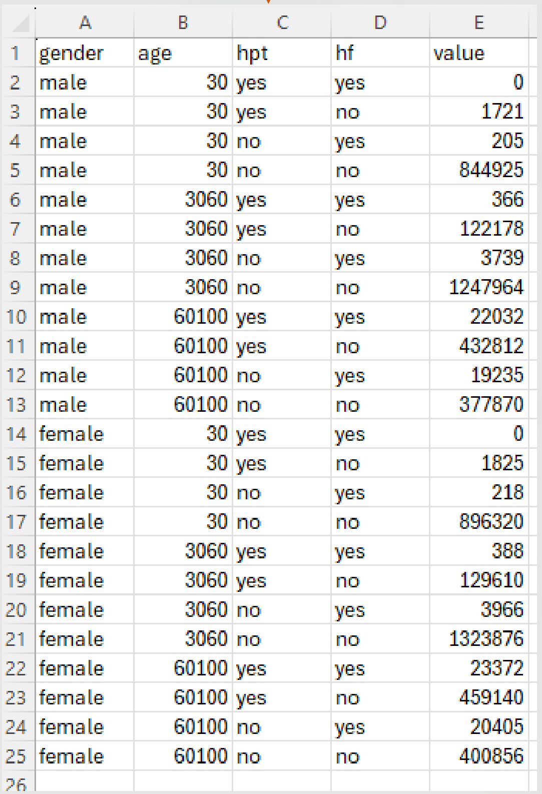
Output CSV file (optional): C:/Users/LENOVO/Documents/GitHub/IPF_multidim/results.csv

Detected variables, conditions, and conditionals:

```
Variables detected:
- gender:
    female, male
- age:
    30, 3060, 60100
- hpt:
    no, yes
- hf:
    no, yes
----------------------------------------
Conditionals / joint distributions identified:
- Base variable: age
  Joint combinations: age_hpt, age_hf
```

Output preview:

```
gender;age;hpt;hf;value
male;30;yes;yes;0
male;30;yes;no;1721
male;30;no;yes;205
male;30;no;no;844925
male;3060;yes;yes;366
male;3060;yes;no;122178
male;3060;no;yes;3739
male;3060;no;no;1247964
male;60100;yes;yes;22032
male;60100;yes;no;432812
male;60100;no;yes;19235
male;60100;no;no;377870
female;30;yes;yes;0
female;30;yes;no;1825
female;30;no;yes;218
female;30;no;no;896320
female;3060;yes;yes;388
female;3060;yes;no;129610
```



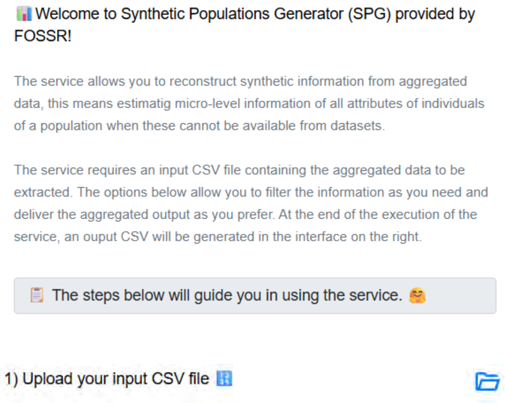
	A	B	C	D	E
1	gender	age	hpt	hf	value
2	male	30	yes	yes	0
3	male	30	yes	no	1721
4	male	30	no	yes	205
5	male	30	no	no	844925
6	male	3060	yes	yes	366
7	male	3060	yes	no	122178
8	male	3060	no	yes	3739
9	male	3060	no	no	1247964
10	male	60100	yes	yes	22032
11	male	60100	yes	no	432812
12	male	60100	no	yes	19235
13	male	60100	no	no	377870
14	female	30	yes	yes	0
15	female	30	yes	no	1825
16	female	30	no	yes	218
17	female	30	no	no	896320
18	female	3060	yes	yes	388
19	female	3060	yes	no	129610
20	female	3060	no	yes	3966
21	female	3060	no	no	1323876
22	female	60100	yes	yes	23372
23	female	60100	yes	no	459140
24	female	60100	no	yes	20405
25	female	60100	no	no	400856
26					

→ pro: not coding needed, no dependencies
→ vs: local CPU, no modifiability

FOSSR VRE Container

enhance collaboration and tools of researchers through digital platform

https://fossr.d4science.org



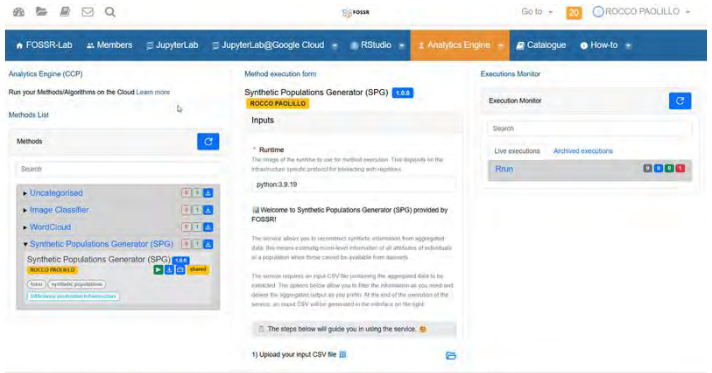
📊 Welcome to Synthetic Populations Generator (SPG) provided by FOSSR!

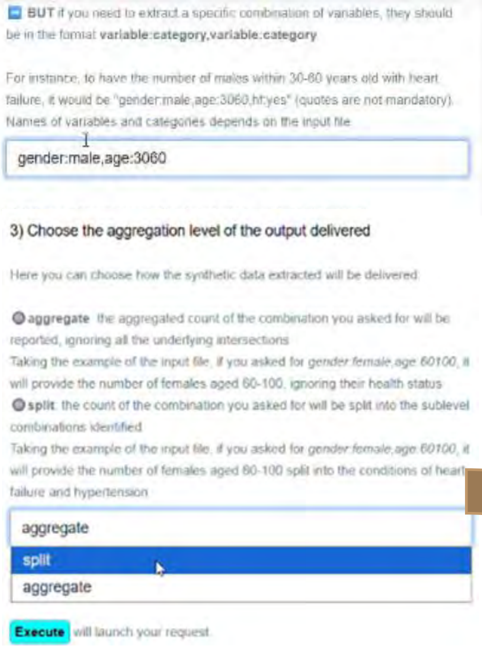
The service allows you to reconstruct synthetic information from aggregated data, this means estimatig micro-level information of all attributes of individuals of a population when these cannot be available from datasets.

The service requires an input CSV file containing the aggregated data to be extracted. The options below allow you to filter the information as you need and deliver the aggregated output as you prefer. At the end of the execution of the service, an ouput CSV will be generated in the interface on the right.

📋 The steps below will guide you in using the service. 🤗

1) Upload your input CSV file





BUT if you need to extract a specific combination of variables, they should be in the format **variable:category,variable:category**

For instance, to have the number of males within 30-60 years old with heart failure, it would be "gender:male,age:3060,hf:yes" (quotes are not mandatory). Names of variables and categories depends on the input file

gender:male,age:3060

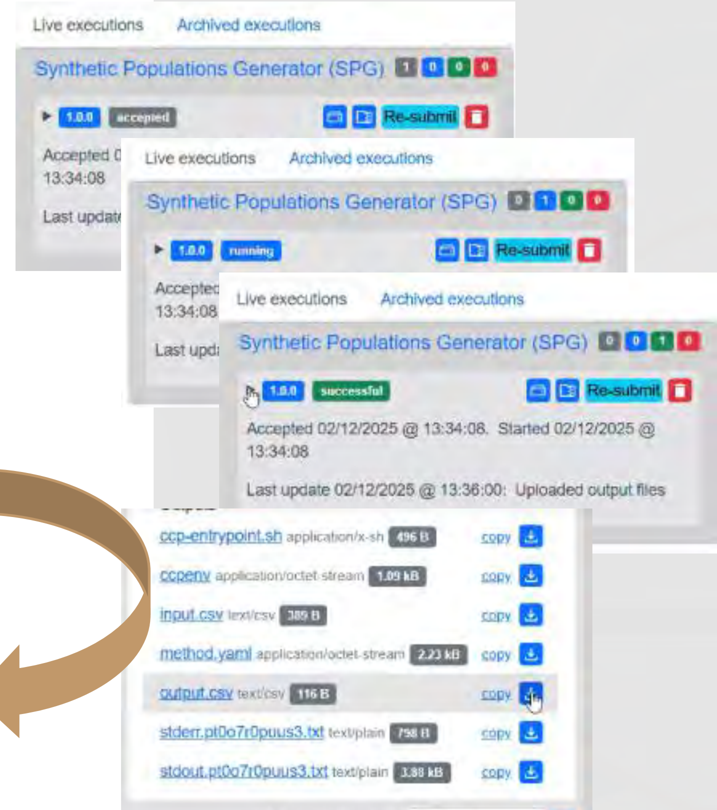
3) Choose the aggregation level of the output delivered

Here you can choose how the synthetic data extracted will be delivered

○ **aggregate**: the aggregated count of the combination you asked for will be reported, ignoring all the underlying intersections
Taking the example of the input file, if you asked for *gender:female,age:60100*, it will provide the number of females aged 60-100, ignoring their health status
○ **split**: the count of the combination you asked for will be split into the sublevel combinations identified
Taking the example of the input file, if you asked for *gender:female,age:60100*, it will provide the number of females aged 60-100 split into the conditions of heart failure and hypertension

aggregate
split
aggregate

Execute will launch your request



Live executions | Archived executions

Synthetic Populations Generator (SPG) 1 0 0 0
1.0.0 accepted Re-submit
Accepted 0
13:34:08
Last update

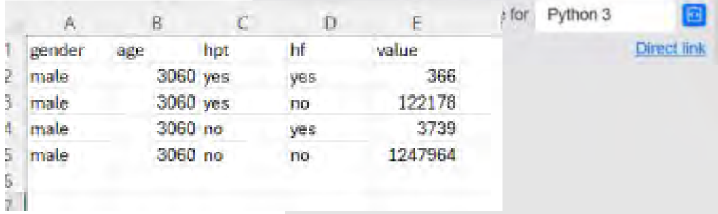
Live executions | Archived executions

Synthetic Populations Generator (SPG) 0 1 0 0
1.0.0 running Re-submit
Accepted
13:34:08
Last upd

Live executions | Archived executions

Synthetic Populations Generator (SPG) 0 0 1 0
1.0.0 successful Re-submit
Accepted 02/12/2025 @ 13:34:08. Started 02/12/2025 @ 13:34:08
Last update 02/12/2025 @ 13:36:00: Uploaded output files

ccp-entrypoint.sh	application/x-sh	496 B	copy
ccpenv	application/octet-stream	1.09 kB	copy
input.csv	text/csv	389 B	copy
method.yaml	application/octet-stream	2.23 kB	copy
output.csv	text/csv	116 B	copy
stderr.pt0o7r0puus3.txt	text/plain	798 B	copy
stdout.pt0o7r0puus3.txt	text/plain	3.88 kB	copy



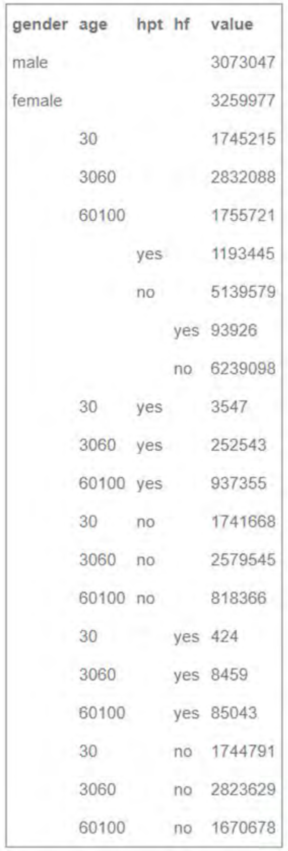
for Python 3
Direct link

	A	B	C	D	E
1	gender	age	hpt	hf	value
2	male	3060	yes	yes	366
3	male	3060	yes	no	122178
4	male	3060	no	yes	3739
5	male	3060	no	no	1247964
6					
7					

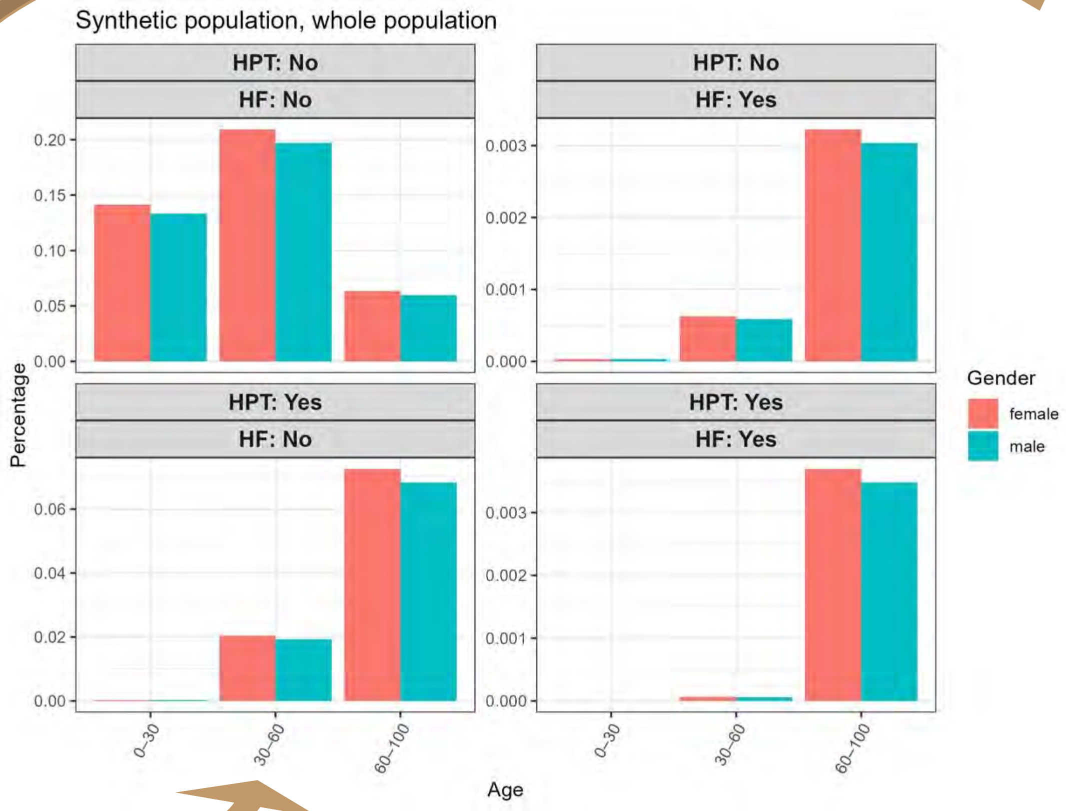
→ **pro**: uses D4Science servers, web-app
→ **vs**: no modifiable, internet-dependent

VRE > CCP (Cloud Computing Platform)> Synthetic Populations Generator (SPG)

Output & applications

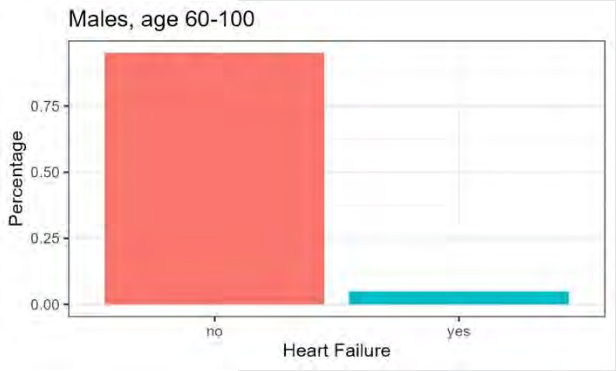


gender	age	hpt	hf	value
male				3073047
female				3259977
	30			1745215
	3060			2832088
	60100			1755721
		yes		1193445
		no		5139579
			yes	93926
			no	6239098
	30	yes		3547
	3060	yes		252543
	60100	yes		937355
	30	no		1741668
	3060	no		2579545
	60100	no		818366
	30		yes	424
	3060		yes	8459
	60100		yes	85043
	30		no	1744791
	3060		no	2823629
	60100		no	1670678



gender:male,age:60100, hpt:yes -d aggregate: 454844

gender:male,age:60100, hpt:yes -d split

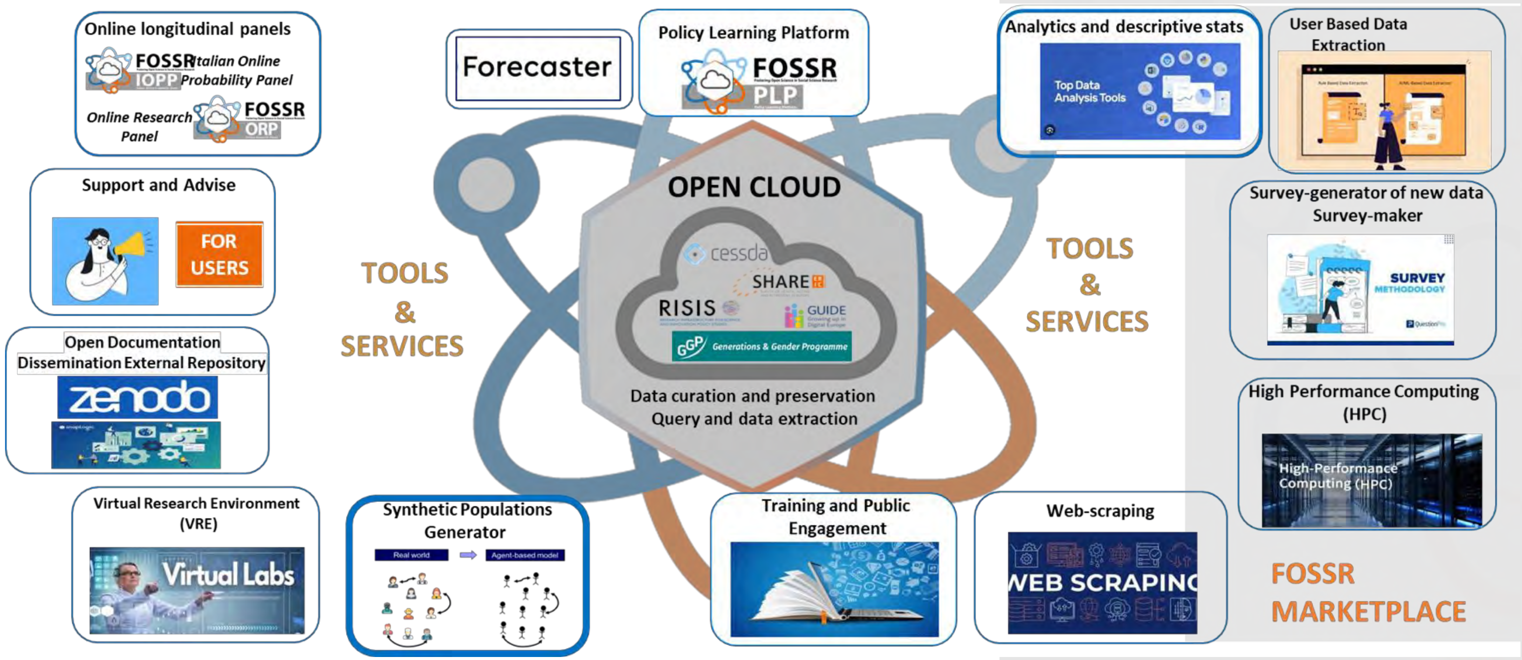


- individual-level attributes as per request
- preserving original data and restrictions
- identify intersections and heterogeneity in the population
- actor-based models initialization (agent-based models, social network analysis,...)

Current and future steps

- Assumes marginals from the same total population
- looking at machine learning as micro-to-macro approach
- Automated iteration over spatial units (census tract)
- Automated multi-source input
- Enhance UX experience and assitance
  - **LLM**

Deployment into the **FOSSR market place**



Online longitudinal panels
FOSSR IOPP Italian Online Probability Panel
Online Research Panel FOSSR ORP

Support and Advise
FOR USERS

Open Documentation
Dissemination External Repository
zenodo

Virtual Research Environment (VRE)
Virtual Labs

Forecaster

Policy Learning Platform
FOSSR PLP

Analytics and descriptive stats
Top Data Analysis Tools

User Based Data Extraction

Survey-generator of new data
Survey-maker
SURVEY METHODOLOGY

High Performance Computing (HPC)
High-Performance Computing (HPC)

TOOLS & SERVICES

TOOLS & SERVICES

OPEN CLOUD
cessda
SHARE
RISIS
GUIDE
GGP Generations & Gender Programme
Data curation and preservation
Query and data extraction

Synthetic Populations Generator
Real world
Agent-based model

Training and Public Engagement

Web-scraping
WEB SCRAPING

FOSSR MARKETPLACE

https://github.com/RoccoPaolillo/IPF_multidim.git

References

Paolillo, R., Roxburgh, N., Sbrana, A., Polhill, G., Sabatella, E. C., & Paolucci, M. (2025). Synthetic Populations in Research Infrastructures. In Longitudinal Data Infrastructures in Europe: Tools for Open Science in Social Science Research (pp. 153-164). Cham: Springer Nature Switzerland.

Chapuis, K., Taillandier, P., & Drogoul, A. (2022). Generation of synthetic populations in social simulations: a review of methods and practices. Journal of Artificial Societies and Social Simulation, 25(2).

Bigi, F., Rashidi, T. H., & Viti, F. (2024). Synthetic population: A reliable framework for analysis for agent-based modeling inmobility.Transportation Research Record, 2678(11),1–15.

Yameogo, B. F., Vandanjon, P. O., Gastineau, P., & Hankach, P. (2021). Generating a two-layered synthetic population for French municipalities: Results and evaluation of four synthetic reconstruction methods. Journal of Artificial Societies and Social Simulation, 24,27.

Falck, V. (2025). Generating spatial synthetic populations using Wasserstein generative adver-sarial network: A case study withEU-SILC data for Helsinki and Thessaloniki. Preprint.arXiv:2501.16080.

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., ... & Crouch, S. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research*, *6*, ELIXIR-876.

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A. L., Martinez, C., Psomopoulos, F. E., ... & Yehudi, Y. (2022). FAIR principles for research software (FAIR4RS principles). *Zenodo*.

THANK YOU!

fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

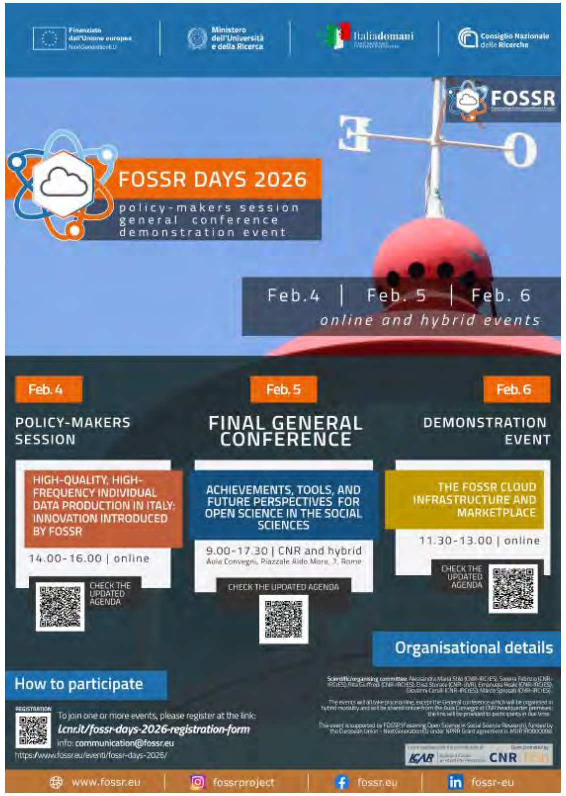
fossr-eu

@fossr

zenodo zenodo.org/communities/fossr

Thank you! Questions?

FOSSR DAYS 2026, 4-5-6 February



Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | Consiglio Nazionale delle Ricerche

FOSSR

FOSSR DAYS 2026

policy-makers session
general conference
demonstration event

Feb.4 | Feb. 5 | Feb. 6
online and hybrid events

Feb. 4
POLICY-MAKERS SESSION
HIGH-QUALITY, HIGH-FREQUENCY INDIVIDUAL DATA PRODUCTION IN ITALY: INNOVATION INTRODUCED BY FOSSR
14.00-16.00 | online
CHECK THE UPDATED AGENDA

Feb. 5
FINAL GENERAL CONFERENCE
ACHIEVEMENTS, TOOLS, AND FUTURE PERSPECTIVES FOR OPEN SCIENCE IN THE SOCIAL SCIENCES
9.00-17.30 | CNR and hybrid
Aula Convegni, Piazzale Aldo Moro, 7, Rome
CHECK THE UPDATED AGENDA

Feb. 6
DEMONSTRATION EVENT
THE FOSSR CLOUD INFRASTRUCTURE AND MARKETPLACE
11.30-13.00 | online
CHECK THE UPDATED AGENDA

How to participate

REGISTRATION
To join one or more events, please register at the link:
l.cnr.it/fossr-days-2026-registration-form
info: communication@fossr.eu
https://www.fossr.eu/eventi/fossr-days-2026/

Organisational details

Scientific/organising committee: [illegible]

The events will all take place online, except the General conference which will be organised in hybrid modality and will be shared online from the Aula Convegni at CNR headquarter premises; the link will be provided to participants in due time.

This event is supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union - NextGenerationEU under NPRR Grant agreement n. MUR IR0000008

ICAR CNR

www.fossr.eu | fossrproject | fossr.eu | fossr-eu

l.cnr.it/fossr-days-2026-registration-form

rocco.paolillo@cnr.it
@roccopaolillo.bsky.social

This work was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NRRP Grant agreement n. MUR IR0000008