



Raising Open and User-friendly Transparency- Enabling Technologies for Public Administrations



Project number 645860
H2020-INSO-2014

D4.4a Alpha version of the SIM

(Final, version 1.0, 2016/07/26)



WISE & MUNRO



Document produced by

Organization: SGH Warsaw School of Economics

Author / email: Przemysław Szufel, Marcin Czupryna, Bogumił Kamiński, Grzegorz Koloch

Subject: D4.4a Alpha version of the SIM

Due date: 2016-07-31

Dissemination level: Public (PU)

Reviewed and approved by

Date	Name	Organization

Revision History

Version	Date	Status	Description of Changes
0.1	2016-01-29	Draft	Preparation of document structure. Specification of requirements for each section.
0.2	2016-02-29	Draft	Literature review & preliminary specification of the model.
0.3	2016-03-28	Draft	Description of methodology & further improvements of conceptual model.
0.4	2016-04-25	Draft	Description of tools and data used.
0.5	2016-05-30	Alpha	Added implementation of the model and preliminary description of results.
0.6	2016-06-27	Beta	Finishing of all sections of the document.
0.7	2016-07-17	Final	General review and corrections of the whole text. Final formatting of document for handover to work package leader.
0.8	2016-07-27	Reviewed	After internal review

TABLE OF CONTENTS

Executive summary.....	4
1 Introduction.....	5
1.1 ODGM MODEL	5
1.2 SIM MAIN FUNCTIONALITIES	6
2 Problem & literature review	9
2.1 Elicitation of preferences	10
2.2 Synthetic population modelling	10
2.3 Statistical properties of social networks	12
2.4 Opinion dynamics	13
3 Methodology	14
3.1 Agent based modelling and simulation.....	14
3.2 Algorithm for reconstruction of synthetic populationS	17
3.3 Meta-modelling.....	19
4 Tools.....	26
4.1 Simulation modules overview.....	26
4.2 UML diagrams for multi-agent simulation of ODGM	27
4.3 Library dependency	31
4.4 Steps required to repeat simulation of synthetic societies in the cloud	31
4.5 Visualisations for the SIM	33
5 Data from pilots for SIM calibration	34
5.1 Prato†	34
6 Results	39
6.1 Simulation experiments	39
6.2 Opinion dynamics in social networks.....	41
6.2.1 Sample simulation run	41
6.2.2 Parameter sweep – determinants of dynamics reconstruction	42
6.2.3 Parameter sweep – preference elicitation error determinants	47
7 Conclusion	52
8 APPENDIX A – Pilot requirement for SIM deliverable functionality.....	54
8.1 Global data on the SPOD platform.....	54
8.2 Data on room level.....	56
8.3 Reporting based on merging SPOD registration data with local census data	57
9 APPENDIX B – SIMULATION ALGORITHM DETAILS.....	58
10 Bibliography.....	63

EXECUTIVE SUMMARY

In the report we consider a scenario where Public Administration (PA) uses an online social platform to give access to citizens to open data and collect information on their preferences. Apart from the core functionalities provided in SPOD, namely: giving access to the data, visualizing the data and allowing for discussions and collaboration of citizens and PA about the data there is a need on the PA side (and possibly also third party NGOs) for open data governance. The SIM deliverable provides implementation of open data governance model developed to support SPOD deliverable.

By open data governance we mean, in the restricted way the information about citizen activity on SPOD. The notion of activity encompasses such data as: citizen logging in to SPOD, which datasets they view, intensity of discussions on SPOD, citizens' preferences revealed on SPOD, social links revealed and formed on SPOD. All those information are of interest for PA as they can help them better understand citizen needs and interests.

In order to develop the best possible Open Data Governance Model and implement it in SIM deliverable, on the basis of the outline of the assumptions and objectives formulated in the ROUTE-TO-PA proposal, we have used the following methods to define functional and non-functional requirements for SIM:

- a) detailed discussions of requirements of PA participating in ROUTE-TO-PA project (the summary of the collected requirements is given in Appendix A);
- b) review of literature to identify best practices in modelling such problems;
- c) Internal discussions with SPOD platform delivery team in order to coordinate SIM and SPOD functionalities and data models.

This procedure allowed us to formulate detailed requirements for SIM. Following the ROUTE-TO-PA delivery plan after delivery of the SIM Beta version, SIM is going to be integrated with SPOD. Therefore it is allowed that the requirements given in this document might be augmented in the next stages of implementation of ROUTE-TO-PA based on the feedback from pilot implementation. Therefore the objective of Alpha version of SIM, disseminated to Project Participants only, is to provide a baseline formulation of the requirements and allow other Project Participants, primarily pilot PA participants, to review them and augment so as to best meet the need to produce flexible and usable Open Data Governance Model. It is expected that the Beta release of SIM, planned for December 2016, will provide revised functional and non-functional requirements along with baseline implementation of SIM.

It should be noted, however, that conceptually SIM is a complex solution, especially in the part providing ability to analyze citizen's opinions expressed on SPOD. Therefore already in Alpha version of SIM we provide initial implementation of this module (mainly omitting data visualization module, which is currently under development and is planned for Beta release).

Main functionalities of SIM module can be grouped in 2 main categories:

- a) Statistical analysis of SPOD users' behaviour
- b) Generalization of the opinions expressed and collected in the SPOD platform on the whole population

1 INTRODUCTION

The report presents the Alpha version of the SIM module for the SPOD Platform. In this report we also present an Alpha version of open data governance model (ODGM). The goal of the ODGM is to help Public Administration (PA) provide information about citizen activity on SPOD, and in particular to design an efficient system for elicitation of social preferences in heterogeneous communities. The social platform for open data (SPOD) allows citizens to monitor allocation and spending of financial resources, controlling PA and hence increasing its efficiency. SPOD platform is described in DL 4.1. For the purposes of this model we assume that the PA shares the information with citizens and thus PA actions are controlled by a participatory society. The better knowledge of societal needs and preferences leads to more efficient decisions and regulatory actions taken by PA.

In order to understand the key challenge faced when citizens' opinions are analyzed it is important to notice that the opinions of sub-population that uses SPOD might be not representative for entire population. In the deliverable we develop a method for analyzing dynamics of preferences in population based on the limited online social network data.

The available data for analysis of preferences include information collected by the PA from the online platform (we assume that it is to be run and administered by the PA) and Census data regarding the population. Hence, the PA has access to basic personal data of platform users, position in the online social network and opinions revealed on the platform. The online user data can be analyzed along with the aggregated census data of the entire population.

We have implemented a multi-agent simulation model that takes into account distribution of personal attributes, social network data and opinion diffusion dynamics. We analyze how different algorithms can allow the PA to generalize preferences collected by the online platform on the entire population.

1.1 ODGM MODEL

The multi-agent open data governance model (ODGM) will allow for an optimal design for elicitation of preferences of heterogeneous societies in an open data setting. The proposed approach will enable creating an optimal design of an integrated interactive system which pursues a new open-data-driven, participation based paradigm of collaboration between public administration and local communities. ODGM employs the following three concepts that bring the proposed concept beyond the state-of-the-art in software modelling and on-line elicitations of social preferences:

1. Integration of virtual user community modelling with social software design process. Firstly, the ODGM will provide initial recommendations for elicitation of preferences in SPOD. Secondly, the ODGM will be capable of analyzing user activity data from SPOD and use it to derive optimal open data governance strategies that can be further used to optimize SPOD functionalities.
2. Open-data-driven, interactive collaboration. General guidelines for on-line elicitation of preferences of PA collaboration with communities will be created. Those guidelines will take into consideration the empirical data from SPOD.

3. Participation of underrepresented minorities in online collaboration. The ODGM will provide recommendations on how to take into account voice of underrepresented minorities in SPOD (elderly, people with low computing skills, residents of suburban areas with low population density).

The ODGM will be created as a simulation software package utilizing state-of-the art multi-agent simulation tool, as e.g. the MASON framework (see Luke et. al, 2005). The model will be created with Java programming language. The ODGM will be provided together with tools for SPOD importing and for output analysis. State-of-the-art data science tools will be utilized for simulation output analysis - Gnu R and Python packages.

The simulation, being the major part of final SIM deliverable will be wrapped by visualization module allowing PA to analyze this data along with other data directly collected with SPOD (like the number of active users or data popularity).

1.2 SIM MAIN FUNCTIONALITIES

In order to define main SIM functionalities discussions have been held with pilot participants of ROUTE-TO-PA project as well as with other project participants.

The main functionalities can be grouped into 2 categories:

1. SPOD users' statistics. It is important to collect and present data about usage of SPOD platform: number of active users, level of interest in different data sets, statistics of discussions and opinions on SPOD platform. This will allow the PA to understand and analyze how is the SPOD used, the main topics of interest for the citizens, the most important data as well as the most active and involved users/citizens.
2. SPOD results generalization. This functionality will support the PA with the information on how representative are the opinions expressed by the citizens (sample) observed on the SPOD platform. It will also provide the PA with the information how could the opinion distribution look like in the whole population (at the certain confidence level). The main socio-demographic features that determine the opinions shared by citizens are to be identified and presented.

One of the important objectives in this respect is to achieve representativeness of society opinions in the analysis of preferences revealed on SPOD platform. This creates a need for optimal design of preference elicitation and aggregation systems with heterogeneity in citizens' geographical location and demographic structure. Public resources are allocated to initiatives by democratic representatives of the citizens', who often are aided by judgments of field experts. Similar types of decisions, yet with a different degree of detail, are taken on respective levels of public decision making (local governments and the central government). Different initiatives influence economic and social well-being of citizens with a different scope and magnitude (e.g. effects can be global, local only, with local overlaps, with externalities, or with a specific mechanism of propagation through the system). The problem of elicitation of preferences, their aggregation and translation (representation) into operational budgets, is a well-known issue in the agenda of economic theory – e.g. see Gajdos et al. (2008) and for a literature review see Fischhoff & Manski (2000).

A feature of model economy, which draws a most pronounced line between the results derived in theoretical research, is the degree of heterogeneity of agents in the economy – e.g. see Kirman (1992). Economics of heterogeneous agents is a relatively new branch of economics, and it's true both for state of the art mainstream

economic paradigms - neoclassical economics, as well as more interaction-oriented approaches - agent-based economics.

As an example application consider the case where financial resources assigned by public administration to possible initiatives represent citizens' preferences as closely as possible, especially taking into account the fact, that discriminative outcomes/equilibriums, in which underrepresented initiatives are never financed, are not allowed in a longer time horizon. The main question that we address is "How to choose socially optimal regulatory actions that will significantly consider heterogeneous agents, and underrepresented social groups?" We approach this problem by modelling preference dynamics in heterogeneous communities.

Finding optimal mechanism design for sharing information on SPOD is required to use modelling tools that analyse heterogeneity of economic agents, their geographic location, virtual and real-world social networks and information flow (including data comprehension) within those networks. The tool that allows modelling and finding optimal design of such complex economic systems is agent-based simulation and modelling – e.g. see Farmer & Foley (2009) and Tesfatsion (2002).

The agent-based simulation model will be built within the MASON simulation framework (see Luke et al. 2005) and implemented in Java. The implemented model will be provided together with tools allowing for rigorous statistical analysis of simulation experiments. The statistical analysis and visualization of simulation will be implemented in Gnu R and Python. The simulation model will provide the Open Data implementation guidelines for government decision makers.

The report structure is following: after the Introduction we present and discuss elicitation of preferences problem along with the literature review in the Chapter 2. Next, we present general approach to synthetic population generation and analysis, network and statistical measures that will be used to estimate properties of social networks and to measure opinion dynamics.

We present the methodology that will be used for analysis of preference elicitation in the social networks in the Chapter 3. We start by description of agent based modelling simulation methodology – where a society is represented by a group of autonomous heterogeneous agents. Next we describe the algorithms that can be used for construction of synthetic populations that is later to be inputted into a multi-agent simulation model. The developed simulation model needs to be run across a large parameter space. Analysis of simulation results requires building meta-models across the analyzed parameter space. The methodology for those meta-models is discussed in the Section 3.3.

A multi-agent simulation model for synthetic populations is presented In the Chapter 4. The model has been calibrated with empirical data provided by one of the pilots, namely the Prato municipality.

Population data provided by two pilots (Prato and Dublin municipalities) is discussed in Chapter 5. We also present the developed tools to analyze the population empirical data in such way that they can be applied to a simulation model.

Chapter 6 contains description and results of the simulation experiments. Firstly, experiment design is described in detail. The experiments have been carried out on an HPC computing cluster. We use meta-models to better explain dependencies between social network properties and preference concordance between real population and synthetic social networks.

Chapter 7 concludes.

Chapter 8 is an appendix containing results of our talks with pilots regarding future SIM functionality that will be implemented in subsequent SIM release. Full model source code is available at the online repository: <https://bitbucket.org/pszufe/socialpreferencessimulation2/>.

Chapter 9 contains the detailed description of the simulation model. This is an important part of simulation model documentation. A detailed discussion of the role of source code in documenting agent-based simulation models can be found, for example, in Gilbert (2008), Law (2006) and Miller (2007).

2 PROBLEM & LITERATURE REVIEW

Local governments are increasingly interested in improving communication with citizens and want to understand their preferences in order to put policies in place in an informed way. Therefore, they are implementing social collaboration platforms, which allow members of local communities to discuss local governance issues. Such platforms allow for both C2C (citizen to citizen) and C2G (citizen to government) communication. Bertot et al. (2010) show that such platforms promote a culture of transparency, information openness and lead to a decrease of the corruption level. It should be emphasized, that such platforms provide a means of two-way communication, in which information flows not only from public administration to citizens, but also the other way around. Moreover, an information exchange between citizens themselves can be observed, i.e. citizens can discuss issues not explicitly directing their remarks to public administrators, but among themselves. For the purpose of the present paper we will call such platforms social platforms, online platforms or just platforms, although we keep in mind a particular context in which they are used. Any kind of information retained from user activity on the platform will be called online data.

Public administration usually has only limited information about individual citizens, but has plenty of aggregated, census-type data, that can be disaggregated using synthetic population simulation, and used to gain additional insight about the preferences of the community on various issues. Such information is essential to the policy maker. It can be used by the public administration to support decision-making processes and to advise PA in an informed way, i.e. on the basis of revealed preferences of the entire population, not only of the subjective opinions formed by the authority. Such a concept stands in line with the recently widely accepted paradigms of open governments and, in particular, of the open data trend. Such initiatives not only enable public administration to better understand the preferences of the population, which serves both ends – the community and the administration, but also to improve communication with citizens and allow community members to discuss local governance issues. Among other benefits, the issues discussed on the platform reveal priorities for the administration, e.g. by indicating matters that are most important for the citizens and therefore should be granted most concern on the part of administration and be included in PA strategies with highest priority.

Public administration, using the data collected from social platforms, would like to draw conclusions regarding the distribution of preferences on various matters and issues in the entire local community. However, the idea that all community members are keen social platform users, informed and advised using online data only is generally misleading. Since users of social platforms do not have to be representative for the whole local population with respect to many characteristics, such as age, sex, income level, location, education etc., a risk emerges of being, possibly seriously, biased, if only online data is used to inform and advise local authorities. This concerns both such issues as majority opinion as well as distribution or heterogeneity of opinions within the community, which itself can be very informative for the public administration. Such bias decreases the platform is utilized more by community members, i.e. with the number of users, but one can expect that, especially at early stages of platforms' implementations, biases can potentially be severe. Therefore, in order to generalize information obtained from online data, for example, from a survey or a voting (like/dislike/neutral), public administration has to understand the qualitative and quantitative nature of heterogeneity of the respondents (users) with respect to such aspects as their geographic or demographic structure, both among platform users and among the rest of community. Such information is obtained from census data, when the entire population is concerned, and from user profile data, as far as platform users are concerned. Such data is used in classical approaches.

Classically, when one wants to infer community opinion on a certain issue, one conducts a survey, in which the sample of respondents is carefully selected, so that it is representative for the general community, i.e., its structure resembles the structure of the whole community along as many important dimensions as possible (like location, age, sex, education, income etc.). In such situation the survey sample is formed exogenously – experimenters construct it according to their wills and means. In case of a social platform, to which users willingly subscribe, the situation is a very different one. One could also say, that the sample population (population of platform users) is formed endogenously – no one assumes that its structure will be of a certain form or structure, it emerges with respect to citizens' propensity to participate. This propensity is not uniform over the entire population, but it is reasonable to assume, that it correlates with census data along at least some dimensions, like age or income, or psychological characteristics, like extraversion, openness or opinion radicalism. Regardless the underlying reasons, opinions formed by users of the online platform, cannot be simply extrapolated to the whole community.

Apart from census-type characteristics, however, data retained on the social platform logs provides public administration with a new dimension of information, which is contained in links or connections that users involve in when using the platform. This information is a rich one and these links/connections get revealed when citizens discuss on-line certain issues/posts. It is worth mentioning that this is not necessarily a census-type link, like family membership, but can be formed between people that even don't know each other at all. On-line discussions can be direct, when two citizens interact with each other in a direct discussion on a certain issue/post which was hosted up on a forum, but they also can be indirect, when two or more citizens discuss the same topic not directly with each other, but with some other community members who discuss a given post, or even in an open way, i.e. posting their opinions publicly, so that it's available for all other platform users involved in a certain conversation. For the purpose of this paper, under a connection or a link between citizens established via an online platform, we are referring to a situation in which two citizens are involved in a discussion of a certain post/issue, regardless of the fact whether they conduct a direct discussion with each other – it will suffice that they just are involved in the discussion of the same post.

2.1 ELICITATION OF PREFERENCES

Elicitation of preferences problem is the classical problem considered in statistical and economical literature. We are interested in learning the preferences of the whole population; however, only the preferences of the small subsample are known. In an ideal case (when we are able to design the survey and draw the citizens randomly or the citizens are drawn randomly as the consequence of the selection process) we can directly generalize for the population (e.g. sample mean is the unbiased estimator of the population mean) and calculate the estimation error. When the sample is biased no direct reasoning is normally available and the sample results must be rescaled using statistical methods (the corrections are applied) to deduce about sample. Such techniques are applied to election surveys and pollings data.

In the context of social networks the situation is complicated as the bias is not only due to different attributes distributions in the sample and the population but also due to the social processes (opinion changes as the consequence of social interactions among citizens) that may have the different form for sample and population. This may lead to even more bias than due to only different attributes distributions in the worst case scenario. In such situations traditional statistical measures for bias correction would underperform the method proposed.

2.2 SYNTHETIC POPULATION MODELLING

Synthetic population generating means creating the dataset that contains the micro data comprising all the citizens (for the public administration level considered e.g. whole municipality) with all the relevant various attributes. These attributes are normally grouped into categories.

Due to privacy and other reasons such data is typically unavailable. The typical situation is that the anonymised (no data that would enable identification is being presented e.g. there is no detailed residence information, other information may be blurred by adding the random numbers) sample of citizens comprising the limited number of citizens is available together with marginal univariate distributions (histograms) with selected cross-tabulated multivariate distributions. Such data is available from different municipal, administrative levels, (sub) regional and national level. However, the data content (marginal distributions available) may differ depending on the administrative level. It must also be mentioned that the level of the data available, the attributes presented and their aggregation level depend very much on the national level, see. e.g. Huang and Williamson (2001).

Depending on the availability of the sample of individual citizens we may distinguish synthetic population generations with and without sample, see. e.g. Lenormand and Deffuant (2013). Another characteristics of the data available is whether individual citizen data is available (a single citizen with the typical attributes as: age category, gender, place of residence, income category, employment status or marital status) or household data (household data typically comprises information on role (household head), number of children, relation between household heads and spouses attributes e.g. age difference) or both

Two main synthetic population reconstruction methods are considered in the literature: synthetic reconstruction approach using the Monte-Carlo method and the combinatorial approach. These methods are compared by Huang and Williamson (2001).

Synthetic reconstruction with Monte-Carlo is done in two steps: generating the multivariate distribution of all relevant attributes and sampling citizens from this distribution. Iterative proportional fitting (IPF) technique is mainly applied for the generation of multivariate distribution (such distribution is normally not available as only two-or three way tables are available in a standard case). Using the sample data, the cross-table (comprising the number of citizens in the sample according to two or more attributes) is presented in multidimensional matrix form. Such matrix presents the information on correlation structure, however, due to stochastic reasons the marginal distributions may differ. Using IPF matrix data is transformed in such a way that the marginal distributions will fit the known marginal distributions of the region and the correlation structure preserved to the extent possible.

The relevant formulas for one step in two dimensional case updating are presented below:

$$x^{k+1}(i, j) = \frac{x^k(i, j)}{x^k(i, .)} \times \tilde{x}(i) \quad (2.1)$$

$$x^{k+1}(i, j) = \frac{x^k(i, j)}{x^k(., j)} \times \tilde{y}(j) \quad (2.2)$$

Where $x^{k+1}(i, j)$ represents the number of citizens in (i,j) cell (the citizens with the first attribute belonging to the i-th class and the second attribute to j-th class) in the k+1 step

$x^k(i, .)$ and $x^k(., j)$ represents appropriate marginal distributions calculated based on the two-dimensional matrix values.

$\tilde{x}(i)$ and $\tilde{y}(j)$ are known marginal distributions of attributes one and two respectively.

Formulas can be easily generalized to the multidimensional case. The updating process ends when the matrix values changes cease to exceed the given threshold value. To generate multivariate distribution of attributes considered, the hierarchical process is applied. The general idea of this process is presented below, see Frick and Axhausen (2004). We start with univariate marginal distributions and step-by-step generate two-dimensional, then three-dimensional and so forth distributions, which are summarised below:

0 step:

(1,0,0), (0,1,0), (0,0,1)

1 step:

(1,0,0) and (0,1,0) \rightarrow (1,1,0)

(1,0,0) and (0,0,1) \rightarrow (1,0,1)

(0,1,0) and (0,0,1) \rightarrow (0,1,1)

2 step:

(1,1,0) and (1,0,1) and (0,1,1) \rightarrow (1,1,1)

In practice some two- and three way tables are already available for the sub-population and there is no need for generating them. Sometimes data on the upper administrative level may be applied, as the starting point of the IPF process. Having generated the multivariate distribution, individual citizens are generated using Monte-Carlo technique.

The combinatorial approach follows another approach. The available data sample is weighted in such a way that its composition fits the observed marginal distributions. The typical measures of fit are: total absolute error, standardised absolute error, phi and psi statistics and z-score.

Having both individual citizen and household data presents additional challenges. Barthelemy and Toint (2012) propose the method for generating both individual citizens and households at the same time. Their procedure consists of three consecutive steps: generating individual citizens, generating households' distributions and generating households (based on the generated individual citizens).

When no sample is available, one could use more advanced statistical techniques such as the maximum entropy generator, as proposed by Barthelemy and Toint (2012).

Guo and Bhat, (2007) present and compare the software dedicated to the problem of the synthetic population generation.

2.3 STATISTICAL PROPERTIES OF SOCIAL NETWORKS

As only a sample of citizens will be active on the SPOD platform in course of the self-selection process this selection process should be modelled. As we represent the social structure of citizens by the graph $G=(V,E)$ where V represents nodes (citizens) and E edges (social links among citizens) graph sampling methods are suitable for this purpose.

Many methods of graph sampling are described in the literature. Depending on the object sampled we can distinguish between node and edge sampling. In the literature we consider following methods of graph sampling, see Frank (1974):

- Random sampling without replacement
- Random sampling with replacement
- Bernoulli sampling
- Random walk
- Snowball sampling

Homogenous Sampling

Random sampling without replacement admits drawing nodes uniformly at random without replacement, whereas random sampling with replacement method allows for nodes replacement (individual nodes can be

chosen more than one time). Both methods have simple statistical design (no information on the nodes connections edges necessary) and the sample size is defined beforehand.

In the Bernoulli sampling method each node can be selected with the given (hetero- or homogenous) probability. The method is characterised by simple statistical design (no information on edges necessary), however, the sample size is undefined beforehand and may differ depending on the result of single drawing.

Walk Sampling, see also Lovasz (1993).

Random walk is the iterative sampling procedure that starting with initial nodes the consecutive nodes are selected among nodes linked (by edges) to the last node selected.

Snowball sampling is the iterative sampling procedure that starting with the initial sample of nodes extends the sample by so called waves (nodes sampled for the nodes that had not already been sampled that are adjacent to the nodes sampled in the previous wave).

Walk sampling allows us to sample without knowing the entire network. However, statistical properties of the sample are more complex than in case of homogenous sampling procedure. Analysis of the sample statistics and generalization requires usage of Markov chains methods and bias correction induced by unequal number of edges degrees for each node. Extension of these methods can be found in the literature e.g. Ribeiro and Towsley (2010) propose multidimensional random walk sampling methods.

Depending on the sampling methods sample statistics such as nodes distributions, dyads and triads distributions may differ, see Frank (1981) so that the different methods must be used for statistical inference. The application of logit models and logistic regressions for the estimation of networks characteristics is proposed by Wasserman and Pattison (1996).

In case of the sparse data, the Bayesian approach could also be applied, see e.g. Butts (2003) or Farine and Strandburg-Peshkin (2015).

2.4 OPINION DYNAMICS

Generally the main groups of learning models considered in the literature are: Bayesian updating and non-Bayesian updating, see Acemoglu and Ozdaglar (2011).

The Bayesian updating model is seen as a model of learning and rational opinions and beliefs updating process. However, the requirements according to the agent knowledge of priors (the prior beliefs distribution over all possible alternatives) and high requirements for computational processing (classical Bayesian probability updating formula) required of citizens, make the practical application of this kind of updating in real life almost impossible.

Therefore simpler methods have been proposed for the purpose of opinion dynamics modelling, see deGroot (1977). This class of models admits opinions updating with the weighted linear mean of owns opinion and the opinions the agent has the relations with (represented by edges in the graph representations). The limitation of such model is that in the limit agreement is reached and no permanent disagreements are possible. Many extensions and variations of the model have been proposed time varying weights, belief depending weights, Krause (2000). These models admit permanent disagreements under mild conditions, see Lorenz (2005). Such effect can also be achieved by introducing the heterogeneity among agents, implementing so called stubborn agents (the agents that do not change their opinions due to influence of the others). We may also allow for different levels of influence/persuasiveness (Zhou et al., 2015) or (Diao et al. 2014). Reaching of the consensus is studied by e.g. Shang (2014).

3 METHODOLOGY

3.1 AGENT BASED MODELLING AND SIMULATION

Socio-economic systems are classified as complex systems, which means that the system as a whole exhibits qualitatively different aggregate macro characteristics than behaviours that can be inferred from simple aggregation of micro level individual actions of individuals, households, enterprises or institutions which constitute its parts.

The emergent differences in macro- form micro- behaviour stems from effects of mutual interactions between individuals. Therefore, in order to effectively model complex socio-economic systems it is not enough to capture behaviour of individual elements and aggregate them, but it is essential to understand and represent the overall dynamics of the system, see Axtell (2007) and Tesfatsion (2002).

This principle is a founding element of agent based modelling, which is a methodology that allows the researchers to quantitatively explain complex social and economic phenomena. In this way it is possible to explain emergent behaviour that is observable in macro scale that is present due to micro scale interactions, e.g. network effects. Agent based modelling overcomes a major shortcoming of standard economic modelling that assumes that it is enough to model a single homogenous type of agent for each class of agents, so called representative agent. In this approach every individual, household, company etc. is identical and fully rational. By full rationality it is meant that it possesses full knowledge, makes optimal decisions and incurs zero costs in decision making process. This approach is clearly not valid empirically. In some situations it was found to be sufficient and provide satisfactory predictive power. However, when we want to explain effects of interactions between agents the fact that they are different and not completely rational is crucial.

The key feature of agent based model is that it contains multiple heterogeneous entities: individuals, households, families, companies etc. adapting their actions to a dynamically evolving environment. Usually agents form hierarchies, e.g. a group of individuals constitutes a household and connections, e.g. social networks. Those three elements, i.e. heterogeneity, adaptive behaviour and complex relationships between agents imply that although in theory it is possible to write down a full mathematical specification of such model, in practice it is not feasible and computer code is a widely applied and accepted method of detailed specification of such models. Additionally, not only the specification of the model is complex. Also when solving them it is usually infeasible to use standard mathematical tools used for proving of theorems but alternatively computer simulation is used. Therefore, because of the complexity of the model, its specification is not explicit (mathematical model) but implicit (computer code) and the method of analysis is not deductive (theorem proving) but inductive (statistical analysis of computer simulation output), see Kamiński (2012).

The foundation of traditional economic modelling was originally built upon what is called now Cowles Commission approach. This approach, in short, consisted in estimation of, sometimes large, systems of quantitative relationships (i.e. equations) specified between aggregate economic variables such as output, investment, unemployment, inflation, money supply, etc. These are ad hoc relationships, which abstracted from individual choice, bottom-up aggregation of economic dynamics and often abstracted from the way in which economic agents form expectations. Rapid development in computing power has permitted for the creation and usage of agent-based simulation techniques in the 1990s. Agent based models are build according to the bottom-up method. It means that agent based models are designed on the micro level, where interactions and behaviours of the individual agents are specified, and then the macro dynamics is observed as an emergent outcome of the model's simulation, see e.g. Oeffner (2009).

The fact that the agent based model is represented and analysed using computer simulation introduces some practical constraints. The most important one of them is the size of the population of agents in the model. Modelling of populations consisting of millions of agents is usually infeasible (however, it should be noted that such big models are met in practice, but they require enormous amounts of computing power) and as alternative synthetic populations of agents are constructed. They usually feature much lower numbers of agents (in magnitude of thousands). In such cases the characteristics of agents are chosen so as to accurately represent the target population. A typical approach is to collect aggregated data about distribution of attributes of entities in real life (e.g. gender, age, income, location) along with their interdependences and create a synthetic population in an agent based model that features similar distributions.

One important benefit of synthetic population approach in agent based modelling is that it allows the researcher or practitioner to analyse counterfactual scenarios. This means that we are able not only to consider and model the behaviour of actual population (as e.g., in econometric modelling). We can also consider what-if scenarios assuming alternative narrations or future events. A good example of such application is a model of New NASDAQ Stock Exchange (Darley and Outkin, 2007) where authors were predicting consequences of new regulations of trading on this market. The constructed agent based model was able to predict most of the emergent behaviour that later occurred in practice (like evolution of predatory trading). Such conclusions were not possible using traditional methods that focused on analysis of historical data.

Agent based modelling approach is very flexible in the sense that using them, one can replicate a wide range of dynamic phenomena. This flexibility can sometimes, however, be challenging. Scientific theories are based on abstraction and agent based models that try to replicate or describe the reality, what can be thought of as an overload with respect to scientific theory. According to the general opinion, models should not be as complex as reality, Leijonhufvud (2006). But this argument seems to be valid, if simplicity is needed, i.e., when the purpose of modelling is generalization (for the sake of an example). Generalization, however, does not have to constitute the only purpose of economic modelling. Apart from abstraction, one can be, and in fact often is, interested in predictions of how does a given system behaves under given circumstances or what are the consequences of changes in the structure of the system. It does not mean that every other system, even a very similar one, must, or is expected to reproduce the same behaviour as the system in question. From such a perspective, close correspondence between the model and the reality seems desirable, if not necessary.

In fact, as it was mentioned earlier, within an agent based framework, a one-to-one correspondence between the model and the real world economy could in principle be possible. One can, however, argue, whether such a correspondence makes any practical sense, since a large number of arbitrary choices would have to be made, regarding the decision making processes of all the groups of agents within the model. These could, in principle, be assumed on the basis of micro studies, experimental investigations or ad hoc assumptions, but clearly, even so, it would be extremely difficult to preserve stability of the model on an aggregated level. Moreover, once the stability has been achieved, it seems to be very fragile with respect to changes in agents' decision making functions and the structure of the model.

On a high level of abstraction, summarizing the observations of Fagiolo et al. (2007) and Oeffner (2009), the following features of agent based models can be emphasised:

1. Bottom-up perspective. Macro-level dynamics appears as a result of behaviour and explicit interactions of individuals on the micro level Tesfatsion (2002), Pyka and Fagiolo (2005).
2. Heterogeneity. Agents are heterogeneous in their behaviour, competencies, (bounded) rationality, computational skills etc.
3. Evolving complex system approach modelled by a network of direct interactions. All agents live in a network which is a complex dynamically evolving system (Kirman, 1997), aggregated properties emerge after repeated interaction between agents take place, agents' decisions are based on

present and past experience, trading of goods and services are modelled explicitly and as a result, the general equilibrium does not hold.

4. Non-linearity. Interactions between agents are highly non-linear, agent based models can contain feedback loops between micro and macro levels (small scale interactions create a macro level dynamics, which in turn influences activity on the micro level).
5. Direct interactions. Agents interact with each other directly, their decisions depend on past and present choices made by other agents (Fagiolo. 1998), subgroups of agents (local networks) can emerge and their structure can change endogenously over time, agents can decide with whom to interact according to expected payoffs (in a bounded-rational way).
6. Bounded rationality. Agents live in a world which is too complex for exact (hyper) rationality, only local or partial rationality can be imposed, agents behave as rational individuals with adaptive expectations.
7. Learning. In numerous agent based models learning algorithms are introduced Windrum and Moneta (2007), agents engage in an open-ended search within a dynamically changing environment, observed patterns constitute a relevant ingredient for learning and adaptation, initial conditions often put agents as units without knowledge about the environment in which they live.
8. Dynamics. Agent based models, due to adaptive expectations, are characterized by dynamics which is irreversible.
9. Endogenous and persistent novelty. Economic systems are non-stationary with constantly introduced novelty, which leads to emergence of new behaviour patterns, which in turn drives adaptation and learning, on top of which agents find it difficult to adapt and learn in such a turbulent and changing environment, e.g. firms introducing new products into the market in order to increase payoffs while results of research and development cannot be known ex ante (Dosi et al., 2006).
10. Selection mechanisms on the market. Goods and services produced by companies are filtered and selected by consumers, selection criteria are complex and involve numerous dimensions (e.g. product features), additional turbulence can be created firms entering or dropping out of the market (Windrum, 2005).

In a more explicit, implementation-oriented manner, a minimalistic ABM consists of the following ingredients:

1. Agents. They are specified as objects of predefined types (e.g. households, firms, banks, and the government) and implemented within the simulated economic environment as autonomous and interactive entities. Agents are characterized by micro-parameters according to which they can differ (e.g. education type, age or productivity). Micro-parameters can be fixed or variable over simulation iterations. Each agent has a set of decision micro-variables attached, which are updated according to ex ante assumed decision rules (e.g. consumption, labour demand, wage offered).
2. Interaction structure. Agents interact with each other exchanging resources they have at disposal (e.g. trading consumption goods, hiring labour supply, borrowing money holdings) and information contained in their information sets (e.g. wages, prices, labour market status). Interaction structure defines who interacts with whom and how.
3. Time. Models are simulated in discrete time steps, e.g. days in Legnick (2013), weeks in Ashraf et al. (2011), months in Giovanni (2010), and quarters in Gaffeo (2008). Different kinds of decisions can be made in different timeframes.
4. Macro variables. Result as an explicit aggregation of micro variables. Some can be defined exogenously on the macro level (e.g. a rate of interest).

The agent based model is usually so complex that it is impossible to parameterize it exactly using empirical data. Usually we have to calibrate it and test its behaviour under different values of its parameters. Therefore the workflow of working with agent based model is multi step, as shown in Figure 1 below.

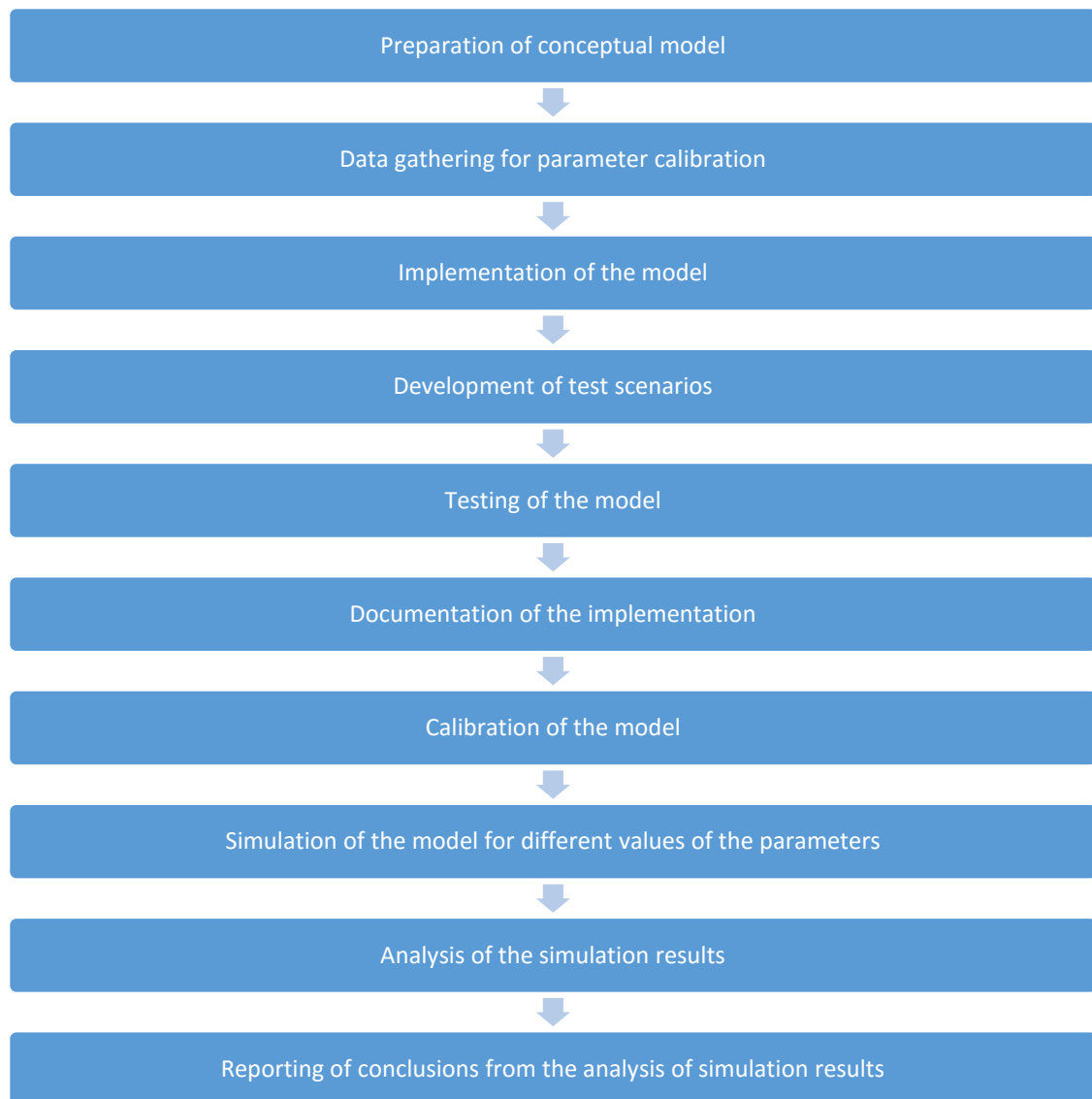


Figure 1 Steps of development and analysis of agent based model.

The above process can be iterative if the obtained results do not reflect the modelled phenomenon accurately. In the following sections we describe the components used for implementation of agent based model developed in this work package.

3.2 ALGORITHM FOR RECONSTRUCTION OF SYNTHETIC POPULATIONS

We assume that we can only observe the sample of the population (users of the SPOD platform). In particular, we observe the opinions expressed by these group of citizens (e.g., in forms of emoticons) in certain intervals of time (users are assumed not only to express their opinions but to change their opinions due to discussion with

the citizens they have links to, as e.g., friends) and the links to other citizens that represent relations between citizens as friendship. Additionally, we assume that we can observe personal attributes (e.g. sex, age, income, social status, employment, family status) of the sample citizens. We also know the marginal distributions of the same attributes. We model such situation as a network where nodes represent the citizens and the edges existing link among them.

In the following section the method which describes preferences of agents, who are not users of the online platform, are inferred using census data and data from the platform are explained in more detail. The method consists of three major steps:

- I. synthetic population generation
- II. opinion propagation dynamics simulation within the synthetic population
- III. opinion dynamics selection and results evaluation

In the first step of the procedure the synthetic population is generated. Synthetic population resembles the true population regarding the marginal distributions of the attributes considered. The primary opinions (opinions of the citizens expressed for the first time with no or almost no influence of the others) distribution and their correlation with socio-demographic data as well as the links among citizens and the correlation with socio-demographic data (in particular the degree of being homophile, similarity between citizens increases the probability of the link existence between them) are estimated based on the observed sample and then reconstructed on the synthetic population.

Wasserman and K. Faust (1994) point out that a social network has the following four distinguishing characteristics:

- (1) Actors and their actions are viewed as interdependent rather than independent, autonomous units*
- (2) Relational ties (linkages) between actors are channels for transfer or "flow" of resources (either material or nonmaterial)*
- (3) Network models focusing on individuals view the network structural environment as providing opportunities for or constraints on individual action*
- (4) Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relations among actors."*

These are network characteristics that should be considered with a special care when designing a multi-agent model of a social network.

Below we outline a 6-step procedure which produces a synthetic population of agents and implements the opinion diffusion simulation, using the network graph representation. The citizens are represented by nodes and the relations among them by edges. We also use the term agent for the citizen. We use the following notation for the citizens. Namely V^P denotes all the citizens in the synthetic population, V^S denotes only these citizens of V^P whose preferences and links are observable (SPOD users) and $V^{NS} = V^P \setminus V^S$ all these citizens that are not observable. Similar superscripts are used for the edges.

1. For a given network structure $G^S = (V^S, E^S)$, using data available for agents in V^S , i.e., $d(v)$ for $v \in V$, estimate a model M_E which predicts a probability that two given agents $v, u \in V^S$ are connected by an edge in G , i.e. a probability, that $(v, u) \in E^S$. This probability will be denoted by $p_{v,u}$.
2. Using model M_E , for all agents $v \in V^{NS} = V^P \setminus V^S$ reconstruct edges of the form (v, u) , such that $u \in V^{NS}$ and $u \neq v$.
3. For a given discussion/post $p \in P$, primary opinions $o(v, 0)$ and data $d(v)$ for all agents $v \in V^S$, estimate a model M_o , which uses $d(v)$ to predict $o(v, 0)$.

4. Using model M_o , for all agents $v \in V^{NS} = V^P \setminus V^S$ reconstruct their primary opinions $o(v, p)$.
5. Using an opinion diffusion algorithm A_o , simulate how opinions of agents $v \in V^S$ change, when interaction of all agents is taken into account, i.e. when opinions of agents $v \in V^S$ are influenced by opinions of agents $u \in V^{NS}$, for which an edge (v, u) is predicted by a model M_E . Different diffusion algorithms are allowed. They differ in how agents value her/his own opinion comparing to the opinion of the citizens she/he has a relation with and how the agents adopt the opinions of these citizens.
6. Compare the generated opinions dynamics of the agents $v \in V^S$ with the dynamics observed on the platform and choose the opinions dynamics process that generates the minimal difference. For all agents $v \in V^{NS} = V^P \setminus V^S$ the final opinions $o(v, n)$ generated by this opinions dynamics are considered.

Details of the procedure are presented in Chapter 9 of this document.

3.3 META-MODELLING

One of the significant challenges of agent based modelling is the difficulty, often impossibility, of derivation of the analytical characteristics of their properties using deductive methods. In such cases statistical induction methods are required. In order to precisely describe the above concept, we use the following formalism.

Assume that there is a given mathematical model M and its property W that is of interest for the researcher or practitioner. In general, there are the following scenarios of their relationship:

- S1) it is possible to verify property W using simulation and analytically;
- S2) it is possible to verify property W only analytically;
- S3) it is possible to verify property W only using simulation;
- S4) it is not possible to verify property W using simulation or analytically;

It is worth noting that the impossibility of analytical verification of the property can be objective (the analytical method of verification is not known) or operational (in theory the problem could be solved but the cost involved is so significant that in practice it is not achievable).

In case of agent-based models their specification is usually so complex that it is impossible to solve them analytically and one has to apply simulation methods to analyse them, therefore we are left with S3 and S4 scenarios only.

Let us emphasize the difference between those two scenarios. Consider an agent based model of open data governance (such as is prepared in this work package), that is model M in the above notation. Assume that we analyse social network between the agents in this model and we are interested in clustering coefficient¹ of the network of connections (this parameter is actually measured in our analysis). We might be interested in two properties $W1$: clustering coefficient is equal to exactly 0.5 and $W2$: clustering coefficient is greater or equal than 0.5.

It is impossible to verify property $W1$ using simulation, even if it were actually true. On the other hand, property $W2$ can be verified. For an extended discussion of this distinction, see Kaminski (2015).

In practice, the analyst is simply restricted to consideration of properties that can be verified using simulation. The basic workflow of simulation experiment with agent based model is given below

¹ A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. There are many formal definitions of, and the simplest one defines it as probability that two random neighbors of a given agent are also neighbors.

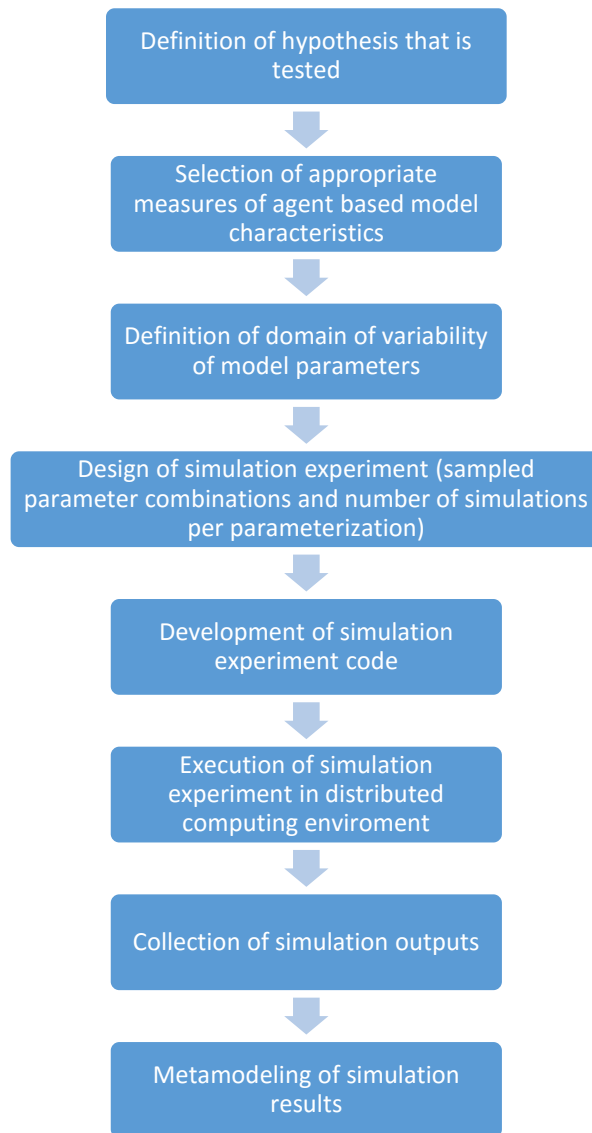


Figure 2 Process of analysis of agent based model properties.

In the above process the two critical elements are design of simulation experiment and meta-modelling and they will be explained in the following paragraphs.

When it comes to design of computer experiments the starting point is identification of model parameters and their domains. Assume that the considered model has parameters P_1, P_2, \dots, P_i with their domains equal to D_1, D_2, \dots, D_i . That means that, for example, the value of parameter P_1 has to be in the set D_1 . Therefore the whole space of the parameters is a Cartesian product $D_1 \times D_2 \times \dots \times D_i$.

Depending on the number of parameters in the product $D_1 \times D_2 \times \dots \times D_i$ four major methods of sampling points for the experiment design are used in practice:

- 1) Full Cartesian product;
- 2) Random sampling;
- 3) Latin hypercube design;
- 4) Low discrepancy sequences.

In order to avoid technical details below we illustrate all four techniques assuming that we have only two parameters P_1 and P_2 .

Full Cartesian product (see figure below) assumes that we select a subset of values of each domain and simulate all combinations of all parameters.

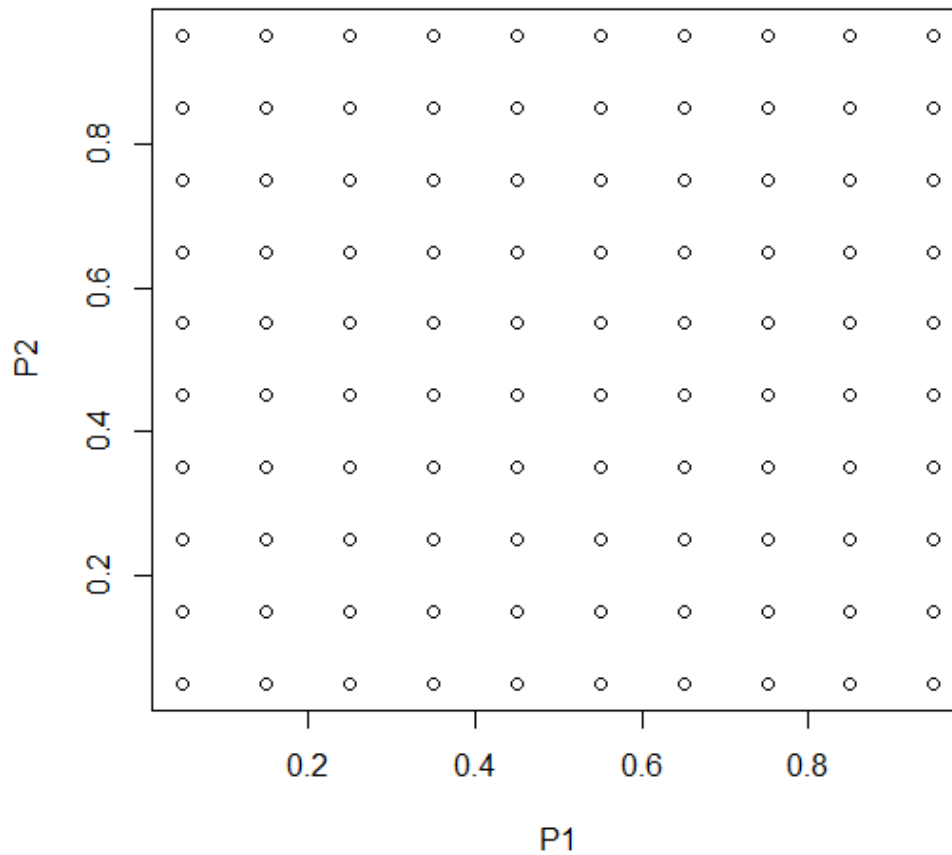


Figure 3 Full Cartesian product.

The advantage of full Cartesian design is that it evenly fills the parameter space. Unfortunately, as the number of parameters grows the number of simulation parameterization in the full Cartesian design grows exponentially. For example with 10 parameters and each measured at 10 values we need to sample 10^{10} points. Another disadvantage is that in each dimension we get only 10 distinct values and are unable to precisely capture the behaviour of the model between them. Finally, one has to know how many simulations one wants to run before the simulation experiment.

The simplest approach trying to overcome these shortcomings is random design, depicted in figure below. In this approach we sequentially randomly pick a point from design space.

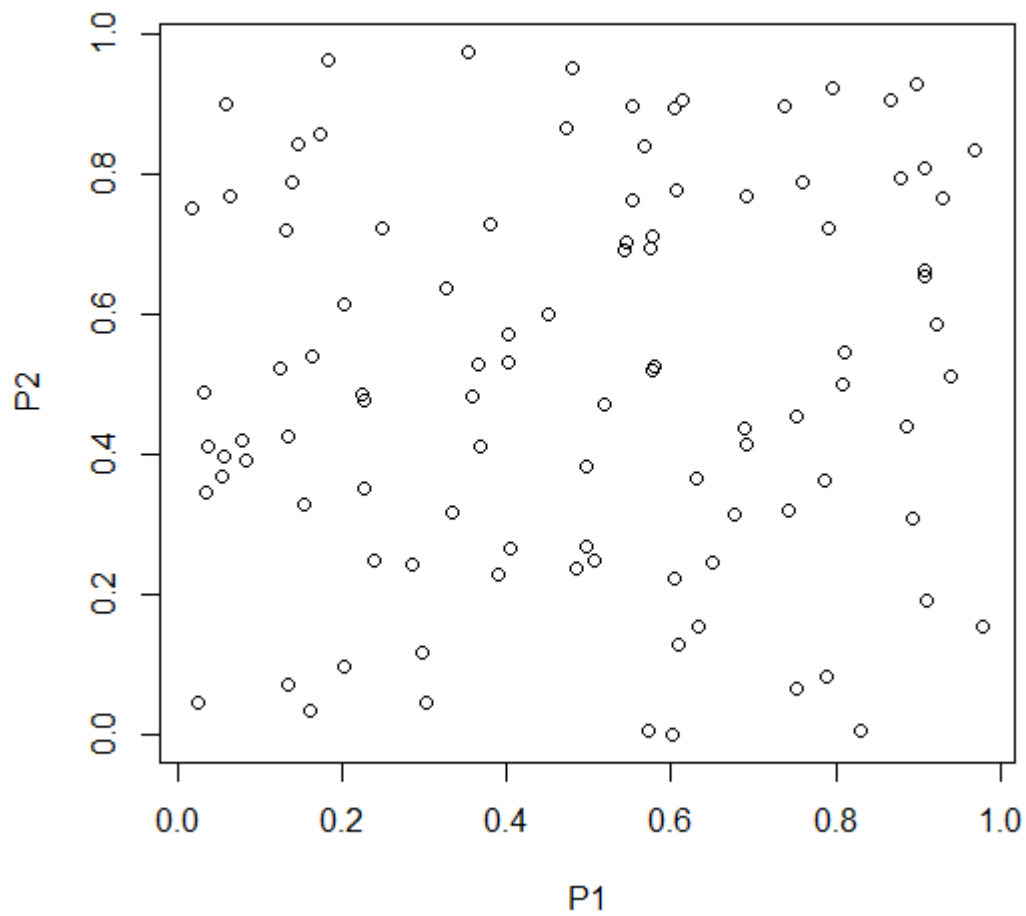


Figure 4 Random design.

The advantages of this approach are that one does not have to plan upfront how many points it wants to sample and that in each dimension all values of parameters are equally likely. However, its shortcoming is that it does not guarantee even distribution of points in design space, which can be seen in Figure above, where in some areas there are clusters of points whereas other areas are blank.

A popular method that is in between full Cartesian product and random design is Latin hypercube design, see figure below.

In this approach each dimension is selected uniformly (exactly as in Cartesian product approach) but instead of calculating their full product the dimensions are randomly matched.

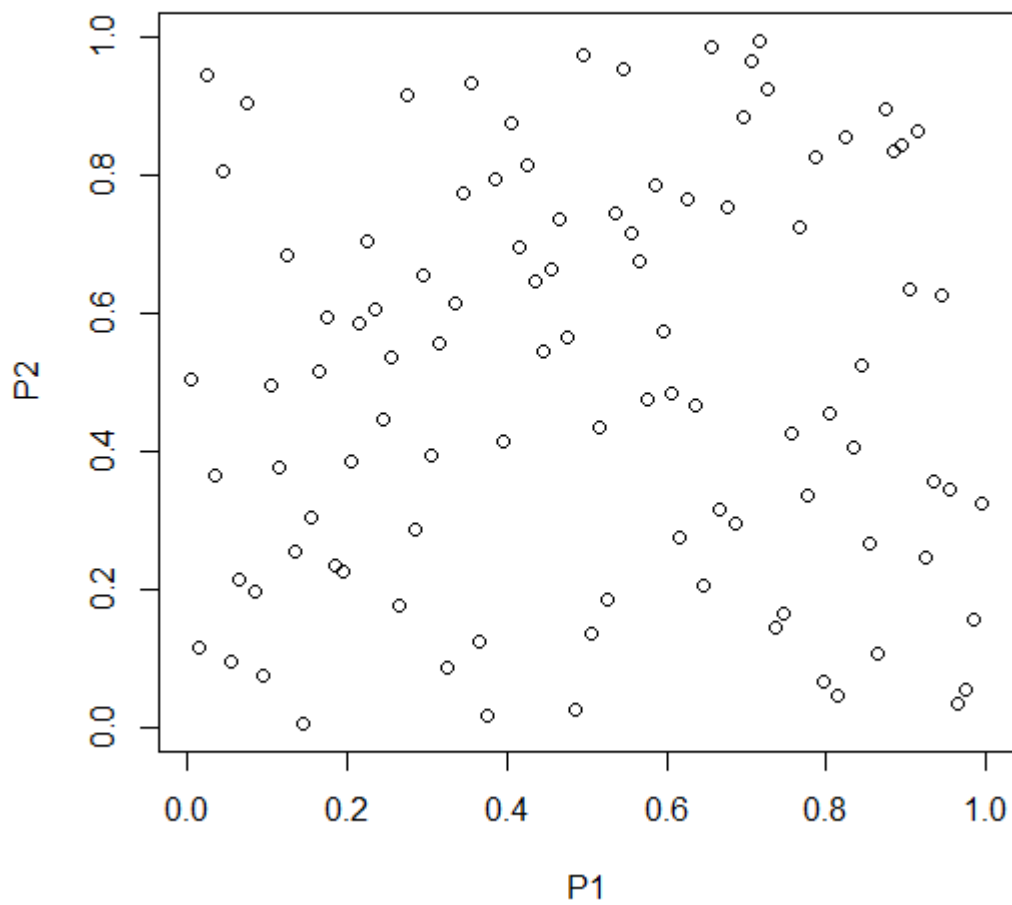


Figure 5 Latin hypercube design.

This approach guarantees (as opposed to random design) that marginal distributions of each parameter are exactly uniform while, at the same time, the number of sampled points is elastic even for highly-dimensional data.

The drawback of Latin hypercube design is that one has to select the number of sampled points before the experiment (this is because – as it was already mentioned – marginal distributions of parameters are not random but predefined like in full Cartesian product).

In cases when one does not know the number of simulations that would be run a good method of choice is “low discrepancy sequences”, also called quasi-random sequences, see Figure below.

In this approach, intuitively, a new point in the design space is iteratively placed in the biggest “hole” in the parameter space.

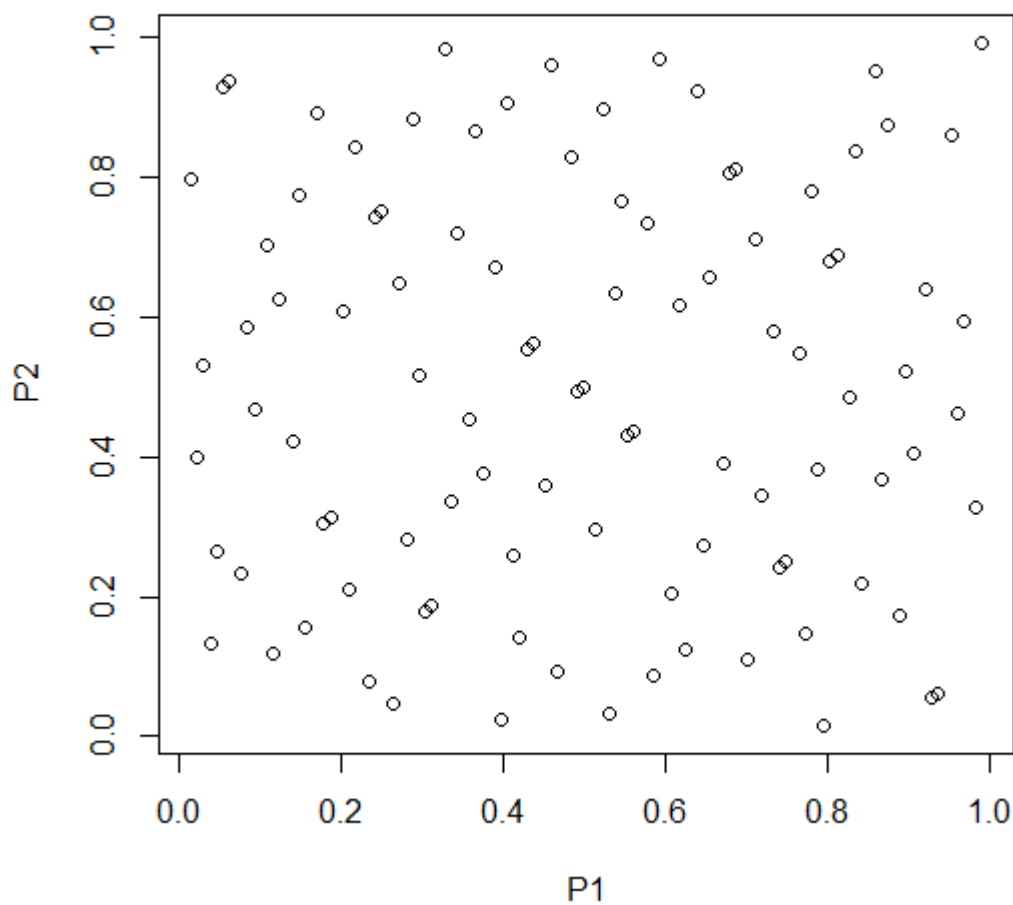


Figure 6 Low discrepancy sequence.

The most popular method for generating low discrepancy sequences are so called Sobol sequences, see Bratley and Fox (1988). They can be recommended when the number of simulations run is not known when the simulation experiment is started. In practice such a situation is quite common because usually only the computational budget is known (e.g. how many hours are given for running of the simulation) and the duration of single run of simulation is not fixed but is changing randomly from one run to other run.

In the Alpha stage of the SIM deliverable we have used reduced the dimensionality of design space and therefore full Cartesian product for experiment design was suitable. However, in beta stage we plan to extend the simulation results analysis and then more advanced methods for analysis of simulation results will be applied.

The second key step in analysis of simulation model is so called meta-modelling. As it is discussed by Kamiński (2015) complex stochastic simulations are often viewed as black boxes in the way they transform their input parameters into output characteristics of modelled systems. Therefore, it is often proposed to use their simpler approximations called simulation meta-models, cf. Barton (1992). The relevant literature identifies various reasons why meta-models can be useful for a researcher (Kleijnen, 2000; Santos, 2007). They can be summarized in three major groups: (i) understanding the shape of the relationship between the inputs and outputs of a model, (ii) prediction and (iii) optimization. These different usage scenarios imply different approaches to the selection

of a functional specification of a meta-model, simulation experiment design and parameter estimation. A review and comparison of meta-model types applied in practice is presented in Wang (2006). In the analysis of agent based model prepared in current work package the meta-modelling has two major objectives: understanding and prediction. Therefore the meta-models are expected to have two features: simple interpretation of their structure and good predictive power.

Formally if we denote by $S(x)$ a simulation model that given parameters x produces random value $S(x)$ the objective of meta-modelling in the context of work package is to find a deterministic function $f(x)$ such that it approximates expected value of simulation $E(S(x))$ as closely as possible. If we assume that the domain of parameter space is D (i.e., $x \in D$) then we want to find the function $f(x)$ that minimizes the following objective function:

$$Q(f, S, D, d) = \int_{x \in D} d(f(x), E(S(x))) dx,$$

where $d(\cdot, \cdot)$ is a measure of distance between $f(x)$ and $E(S(x))$. Most common distance measures d are absolute value $d_a(x, y) = |x - y|$ and squared distance $d_s(x, y) = d_a^2(x, y)$.

In practice it is impossible to evaluate $Q(f, S, D, d)$ and minimize it. Therefore its approximation based on sampled data is used. Assuming that we have run the simulation in points x_1, x_2, \dots, x_n respectively n_1, n_2, \dots, n_k times and collected observations $s_{i,j}$ the simplest approximation of $Q(f, S, D, d)$ is the following:

$$q(f, S, D, d) = \sum_{i=1}^n d\left(f(x_i) - \frac{\sum_{j=1}^{n_i} s_{i,j}}{n_i}\right).$$

This estimator is unbiased i.e., $E(q(f, S, D, d)) = Q(f, S, D, d)$.

The key challenge in this process is the choice of an appropriate space F of approximation functions f . As it was mentioned earlier this space should have two desirable properties: good explanatory power and ease of implementation.

In the current work package we have selected the following classes of models. They all have the desired properties noted above. Here they are ordered by increasing explanatory power and decreasing simplicity of interpretation:

- Linear regression;
- Generalized additive models;
- Random forest.

4 TOOLS

The goal of this Chapter is to present tools used for the execution of agent-based open data governance simulation model. We start with an overview of simulation modules. Next, we present UML diagrams of the multi-agent simulation model. We discuss a set of libraries required to run and analyse the simulation. Subsequently, we describe tools used to set up a full HPC simulation environment in the cloud. All source code required to run the simulations along with provisioning the computational infrastructure in the cloud is available at <https://bitbucket.org/pszufe/socialpreferencessimulation2/>. Finally, we shortly discuss tools that will be used for data visualisation by the SIM module.

4.1 SIMULATION MODULES OVERVIEW

We have developed a multi-agent simulation model that enables modelling of preferences in social networks. The implementation was done with the following Free Open Source Software: R, Java, Python, MASON, Weka, JUNG.

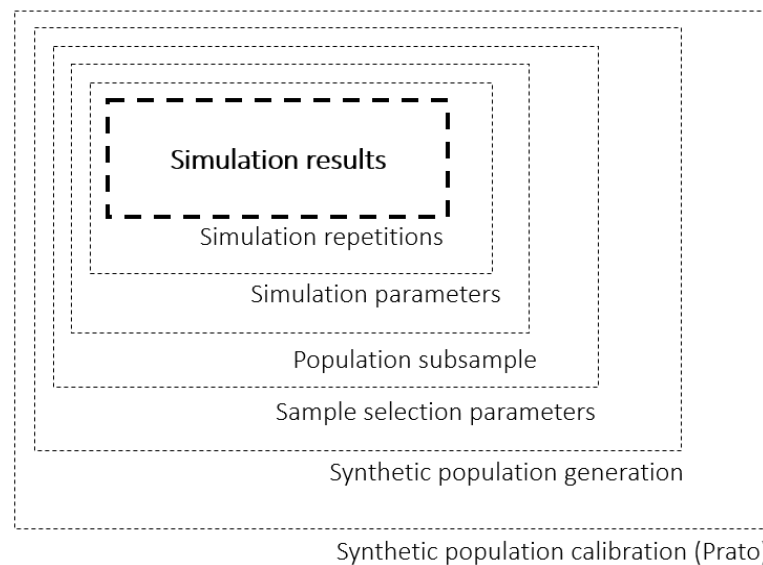


Figure 7 Parameter Layers of multi-agent model for simulation of Open Data Governance Model

Figure 7 presents layers of the simulation for open data governance model. This approach allows to compare various parametrizations (discussion of model parametrization can be found in section 6,1) and hence test robustness of various population preference dynamics.

Computational flow is presented in the Table 1. We present the required steps along with the description of the developed simulation tools.

Steps & tools for synthetic population modelling & simulation	Implementation
Simulation setup script	A set of java classes to set-up the entire database structure
The Prato census data are analysed aggregated and subsequently used to generate a synthetic population.	R implementation for synthetic population generation - PRATO_SYNTHETIC_POPULATION_HSQLDB.R See Appendix B for details
Links between synthetic population agents are created	Class EdgeConstructorPrato (see appendix B for details)
Primary preferences within the synthetic population are created	Class PreferencesGeneratorPrato (see appendix B for details)
Population dynamics simulation	Class SimulatorPrato (see appendix B for details)
Choosing a sample subpopulation through forest fire method	Class SampleSelectorPratoReduced (see appendix B for details)
Running actual simulation and parameter sweeps	Class SampleSimulatorPrato2 (see appendix B for details) Supporting bash script compatible with the MIT Starcluster tools for large simulation execution

Table 1 Steps and software tools developed for Open Data Governance Model. Details of source code can be found on project repository reachable at <https://bitbucket.org/pszufe/socialpreferencessimulation2/>

4.2 UML DIAGRAMS FOR MULTI-AGENT SIMULATION OF ODGM

We will describe our multi-agent simulation model implementation through UML diagrams (see the following figures). The literature stresses the importance of UML diagrams in documenting multi-agent simulation models – e.g. see Oechslein (2002). Full source code of simulation model is available at the repository <https://bitbucket.org/pszufe/socialpreferencessimulation2/>.

Due to high model complexity we have divided the UML diagrams into three parts:

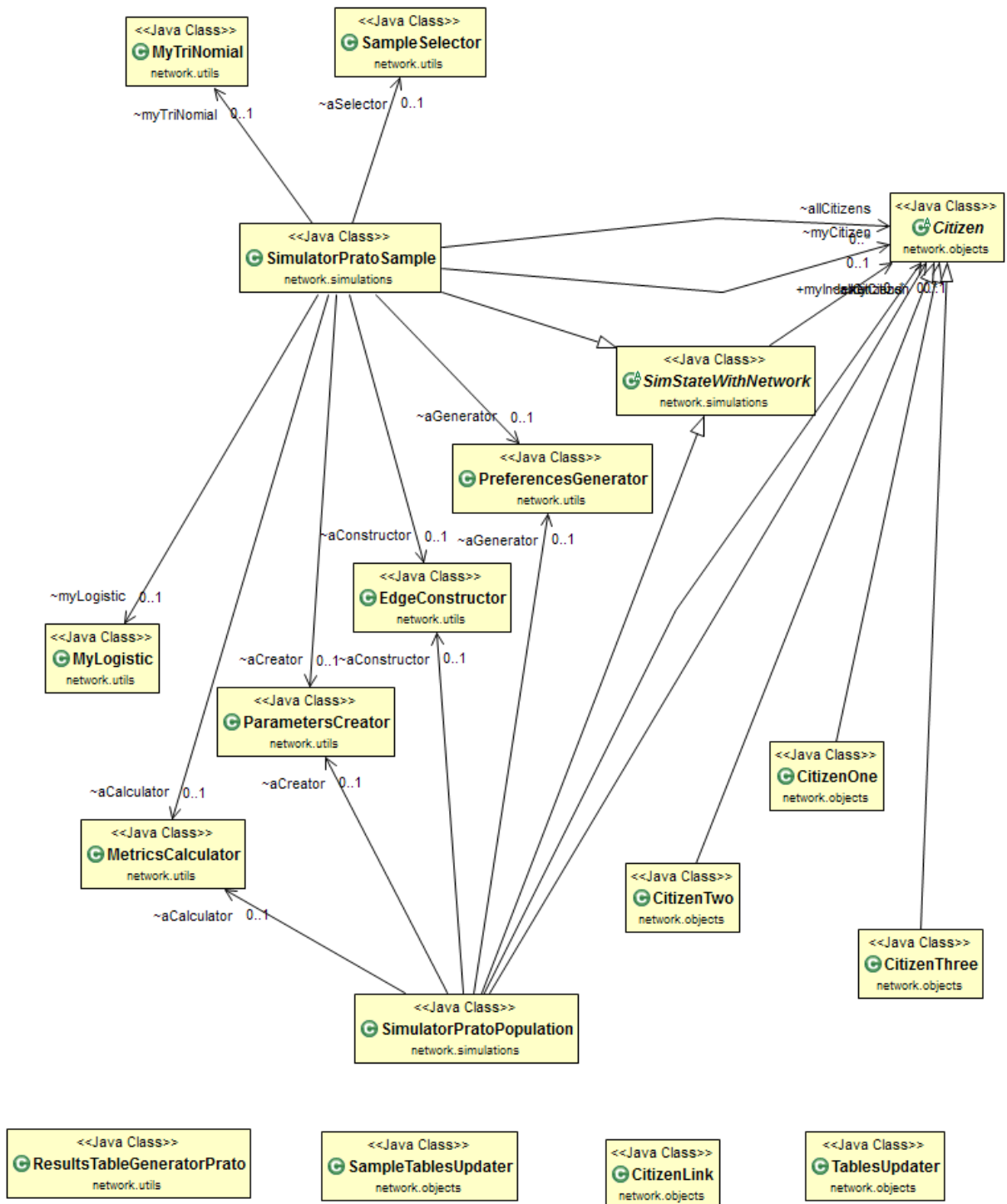


Figure 8 Class diagram of the multi-agent simulation model for open data governance modelling. Model allows to model preference dynamics in a local municipality population – general overview of the model

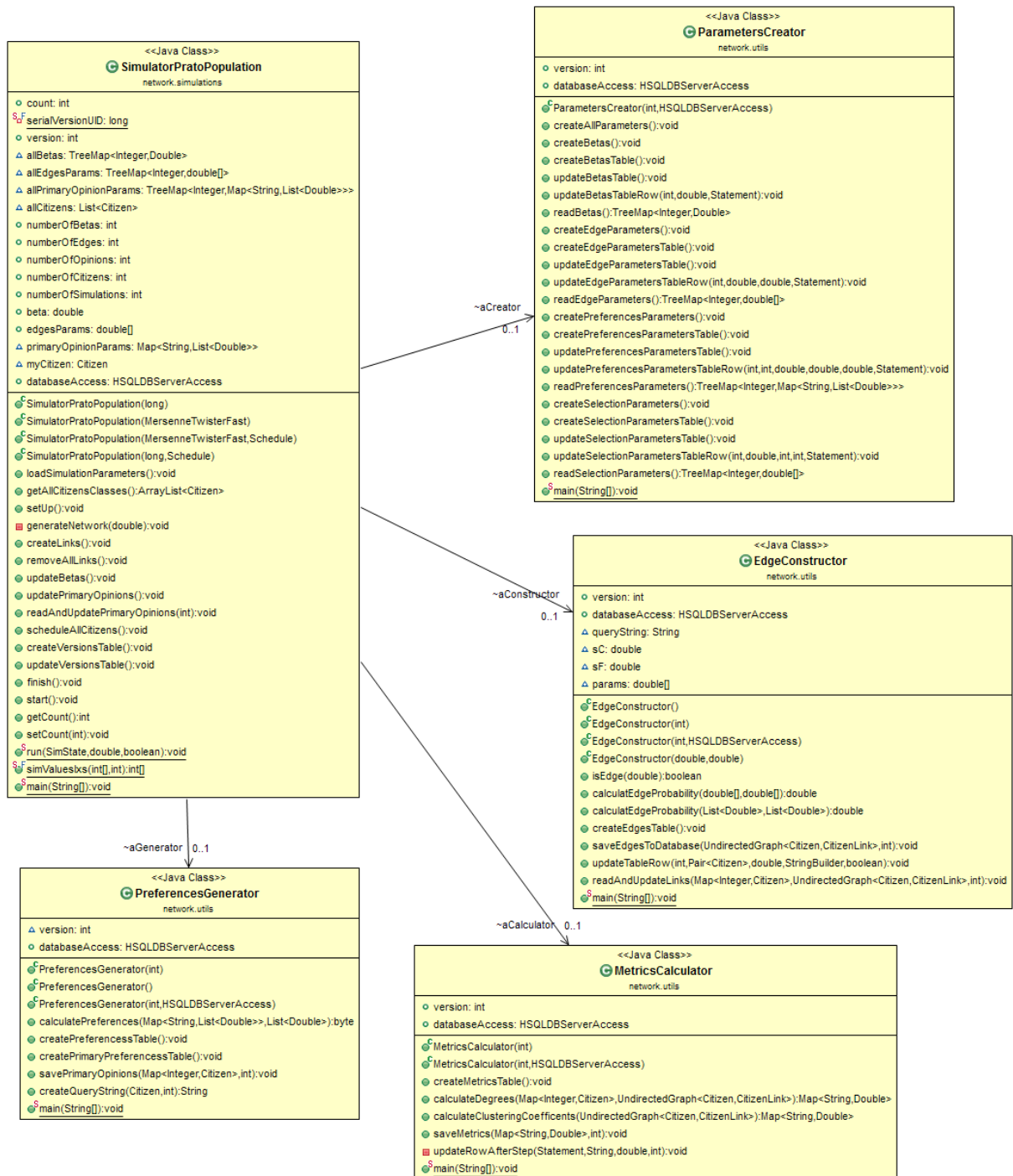


Figure 9 Class diagram of the SIM module – classes for representation and measuring network structure

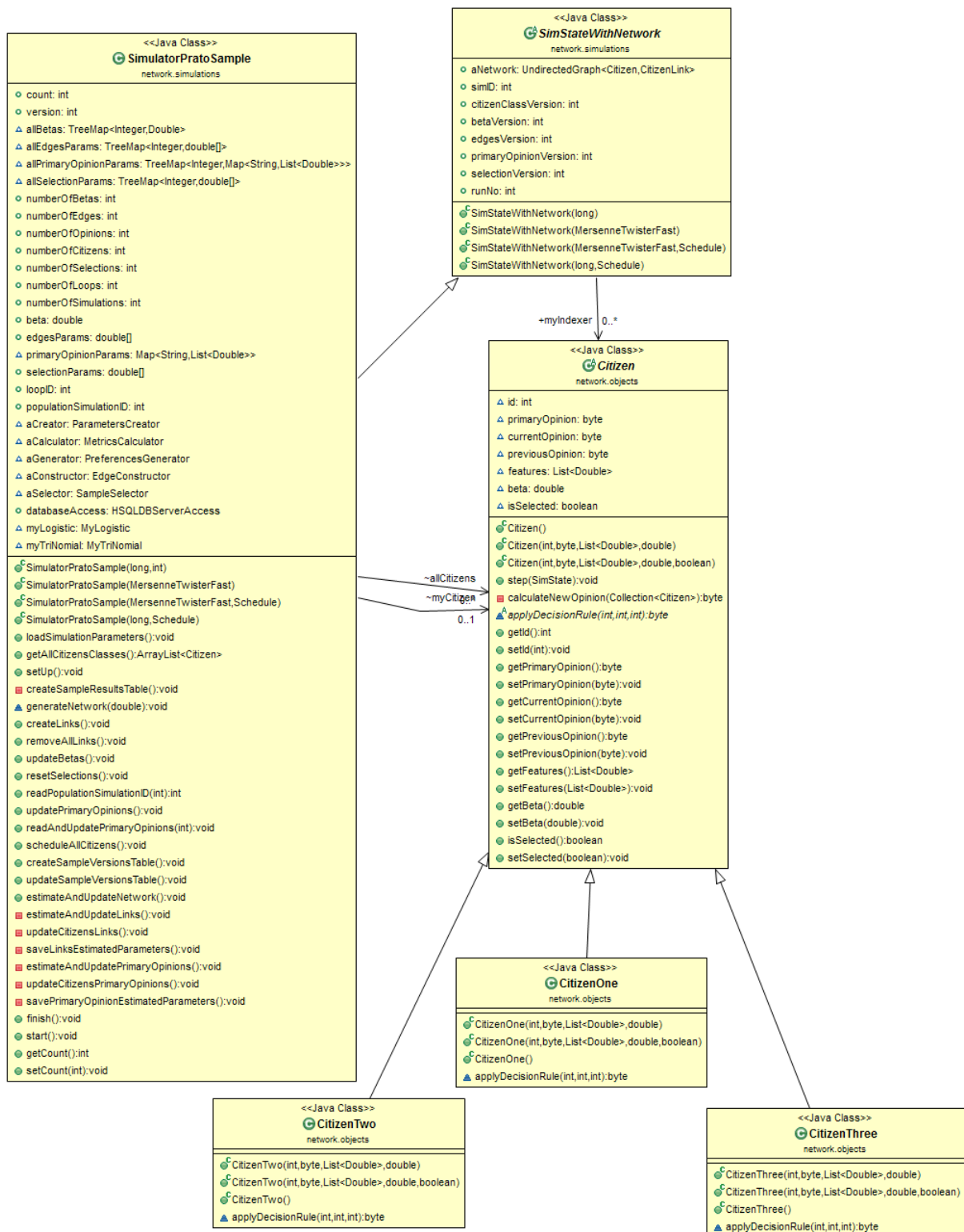


Figure 10 Class diagram – agent representation within the model. Three agent types are considered having different decision rules

4.3 LIBRARY DEPENDENCY

The simulation module depends on several Open Source libraries. We use functionalities for network processing, relational database, mathematical computations and machine learning.

The Java library includes:

- Apache commons: commons-lang3-3.4.jar, commons-math3-3.6.1.jar
- Apache COLT: colt-1.2.0.jar
- HSQLDB database: hsqldb.jar
- JUNG: jung-3d-2.0.1.jar, jung-algorithms-2.0.1.jar, jung-api-2.0.1.jar, jung-graph-impl-2.0.1.jar, jung-io-2.0.1.jar, jung-jai-2.0.1.jar, jung-jai-samples-2.0.1.jar, jung-samples-2.0.1.jar
- MASON: mason.19.jar
- Vector math: collections-generic-4.01.jar, vecmath-1.3.1.jar
- Weja machine learning engine: weka.jar

The synthetic population is generated with R scripts that depend on the following libraries: openxlsx, plyr, stringr, RMySQL and RJDBC.

4.4 STEPS REQUIRED TO REPEAT SIMULATION OF SYNTHETIC SOCIETIES IN THE CLOUD

The goal of this section is to describe how to replicate our simulation results.

The approach presented in Figure 7 and discussed later in section 6.1 leads to a very highly dimensional computational space. Hence, the open data governance simulation model needs to be run on an HPC computational cluster. The following Open source tools have been utilized to perform computations: StarCluster, Open Grid Scheduler and Linux operating system for computational and master nodes. The entire infrastructure has been run on Amazon Web Services and its structure is presented in Figure 11.

The computations are run within Amazon Web Services computational framework Figure 11.

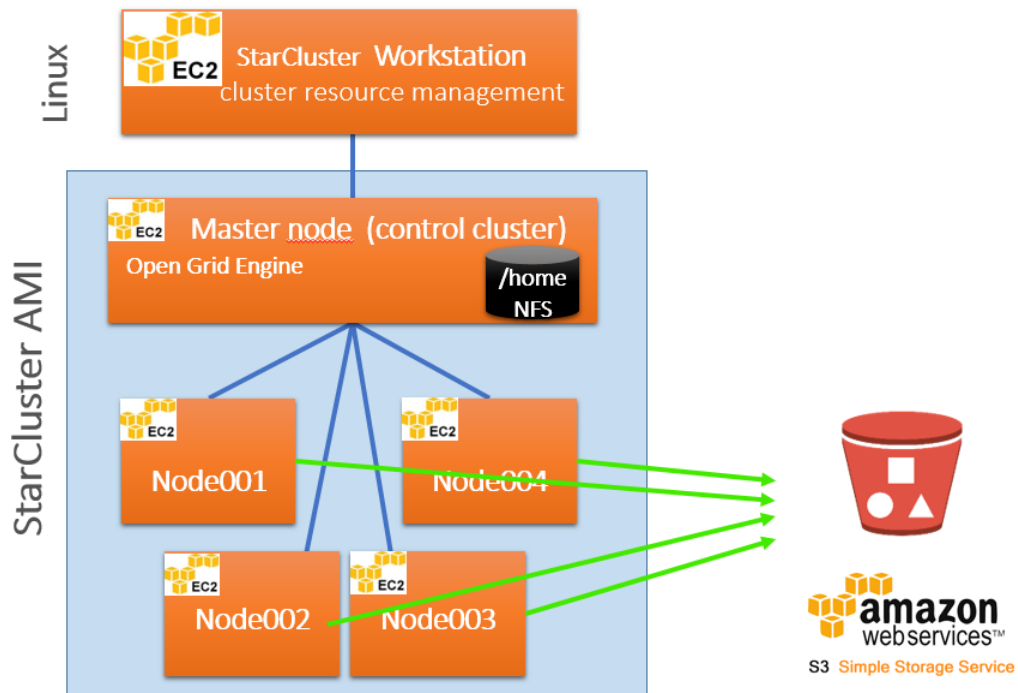


Figure 11 Architecture for execution of simulation of an Open Data Governance Model. Open source tools are used including StarCluster, Open Grid Scheduler and Linux operating system for computational and master nodes. The entire infrastructure is runs on Amazon Web Services.

1. Create an AWS S3 bucket that will be used to distribute application and collect simulation results. Create folders within the bucket "app" to store the application data and "res" to store the results.
2. Pull the simulation source code from the repository and compile it (the easiest way is to open it with Eclipse IDE). The full source code is available at: <https://bitbucket.org/pszufe/socialpreferencessimulation2>
3. Copy your simulation libraries (including dependencies) to the app folder
Create an IAM role SimulationNode along with the policy to write and list buckets (see bucket.policy file in the repository):
4. Create an IAM admin user with administrative access. Save the credentials credentials.csv file
5. Spin up a standard Ubuntu instance
6. Execute the following commands (update server address respectively)
`scp -i cluster-key.pem credentials.csv ubuntu@your_ip_address:/home/ubuntu/`
7. Download the "sc_setup.sh" from the "aws" folder in repository and execute
`bash sc_setup.sh`
8. Execute "nano .starcluster/config" and edit the configuration for your needs
9. configure & prepare the cloud-init file – config line should have the following content:
`USERDATA_SCRIPTS = ~/.starcluster/init.sh`
10. Upload the cloud_init.sh and job.sh files
11. Execute commands to start up the cluster and submit computations
`starcluster put cluster1 job.sh /home/`
`starcluster sshmaster cluster1`
`qsub -V -t 1-4000 -cwd -N job1 /home/job.sh`

The above steps allow execution of a massively parallel simulation runs. The results will be saved in an S3 bucket. Next they can be downloaded with a tool such as S3 Browser. Finally run the class “utils.Tylda” to extract and aggregate the data from simulation logs. These data can be analysed with appropriate machine learning tools (see the meta-modelling section 3.3).

4.5 VISUALISATIONS FOR THE SIM

Results of discussion with Pilots (see Appendix A) show a need to develop tools to visualize various aspects of SPOD usage dynamics along with opinion representativeness. In order to achieve this goal a development of a set of visualisation tools for SIM has started. The tools will be provided in an open source model and are currently available at the following address: <https://github.com/mwasiluk/ODC-d3>. The visualisation tools focus on presenting dependencies within the data. So far for the SIM module tools include a scatter plot, scatter plot matrix, correlation matrix and linear regression modelling (see Figure 12).

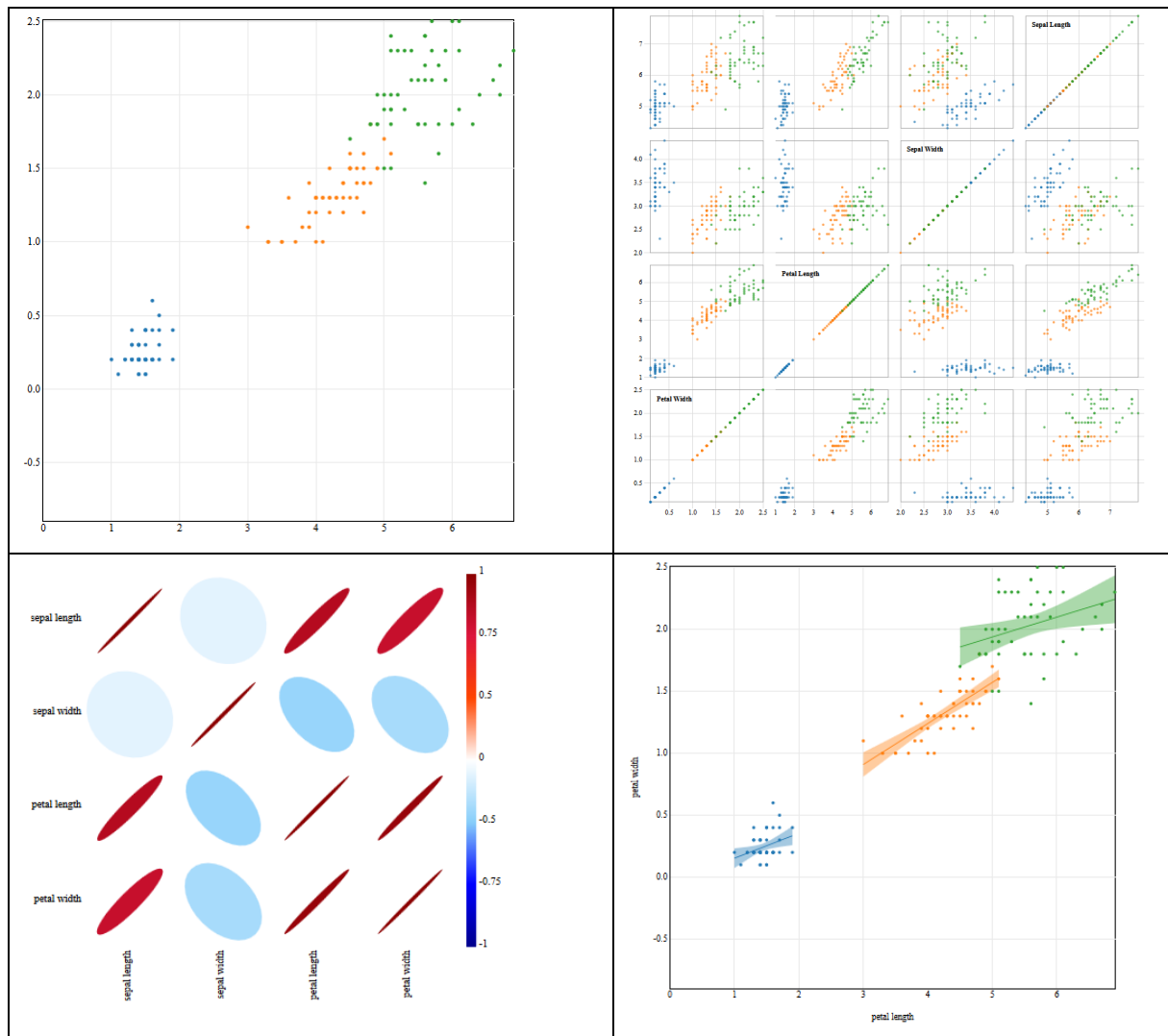


Figure 12 Sample data dependence visualisation currently under development within the SIM module. The visualisations will be used to illustrate preference structure within a virtual society.

5 DATA FROM PILOTS FOR SIM CALIBRATION

The starting point for the simulations is the synthetic population generated using the local census data. The census data available is heterogeneous with respect to attributes available, the aggregation of attributes into separate categories, level of availability (individual citizen in Prato city versus a household in Dublin city) and the cross-tables (mostly two- or three-way tables showing the joint distribution for 2 or 3 different attributes at the same time). Within this phase of the project we have used the census data from Prato municipality. Such data comprises marginal distributions for single citizens and is described below in more detail. Further work will concentrate on generating the synthetic population for other cities e.g. Dublin that are available in other forms than Prato census data.

5.1 PRATO†

Prato (Italian city in Tuscany region) data is available on an individual citizen level. We have received the census demographic data from the Prato municipality and income data based on tax information. In particular following tables were available and were the basis for the Prato synthetic population generation:

- Table population that contains the distribution of Prato population among 6 regions of residence (00, East, West, North, South and Central regions), 14 age categories (0-2,3-5,6-10,11-13,14-17,18-24,25-34,35-44,44-54,55-59,60-64,65-74,75-84,84-) and sex (woman, man)
- Table marital status that contains the distribution of PRATO population among 6 regions of residence (00, East, West, North, South and Central regions), 4 marital status categories (single, married, divorced and widowed) and gender (woman, man)
- Table employment that contains the distribution of PRATO population among different employment categories and gender
- Table income that contains the distribution of PRATO population among 7 income categories (0-10000,10000-15000,15000-26000,26000-55000,55000-75000,75000-120000,>120000)

The goal was to generate synthetic population of households' heads each characterized by the following features: region of residence, age, sex, marital status, income category. As we did not have all the cross tables we made some assumptions about correlation structure among attributes considered within synthetic population generation process. Such assumptions are not crucial as citizens data sample (the citizens that registered on the SPOD platform and provided the socio-demographic information) will be available after the SPOD platform is launched and thus the IPF procedure as described in Chapter 2 can be used for generating the correlation structure among all considered attributes.

Synthetic population generation (we have only considered households heads) followed the following steps:

- 1) Selection of citizens of minimum age of 18
- 2) Reading in, as a starting point, the table population with marginal distribution according to: region of residence, age and gender
- 3) Assuming theoretical age and marital status correlation structure (correlation structure among age and marital status is briefly described by higher than proportional share of singles among younger citizens and the higher than proportional share of widows/widowers among older citizens) we obtain the correlation structure among these two attributes for Prato by fitting the theoretical correlation structure to the real marginal distributions for age and marital status (applying Iterative Proportionate Fitting)
- 4) Households heads were randomly (with equal probability for a woman or a man) chosen from the married citizens

The distributions received for Prato are presented in the following figures.

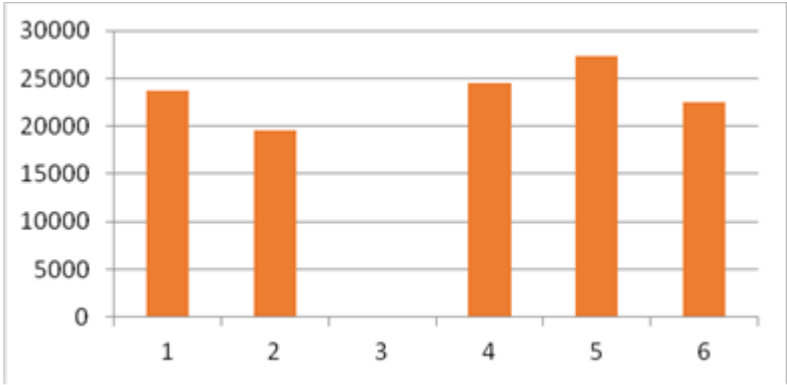


Figure 13 Prato synthetic population distribution according to district of residence.

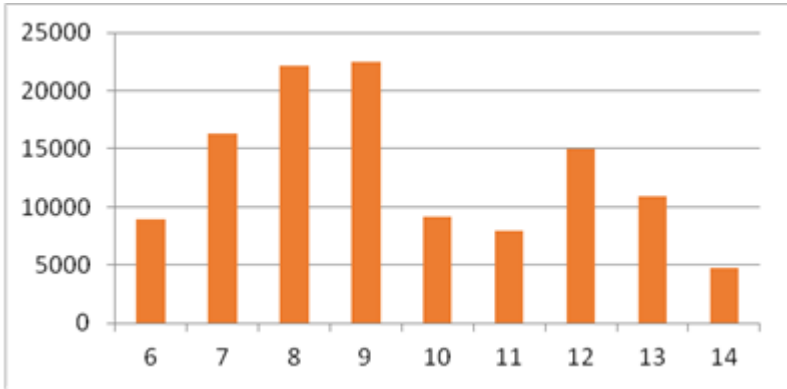


Figure 14 Prato synthetic population distribution according to age.

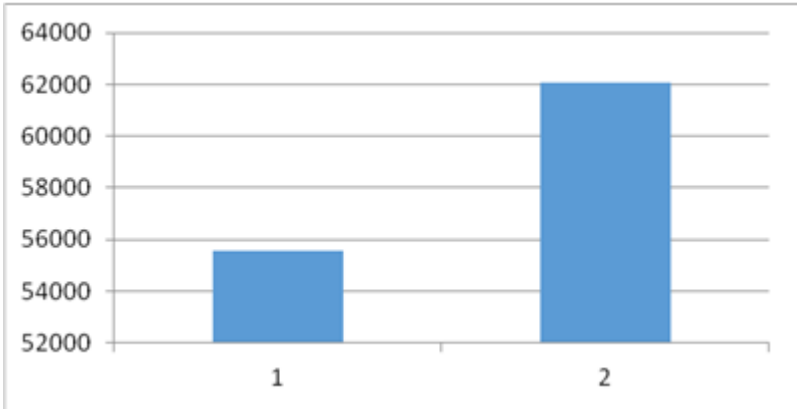


Figure 15 Prato synthetic population distribution according to gender.

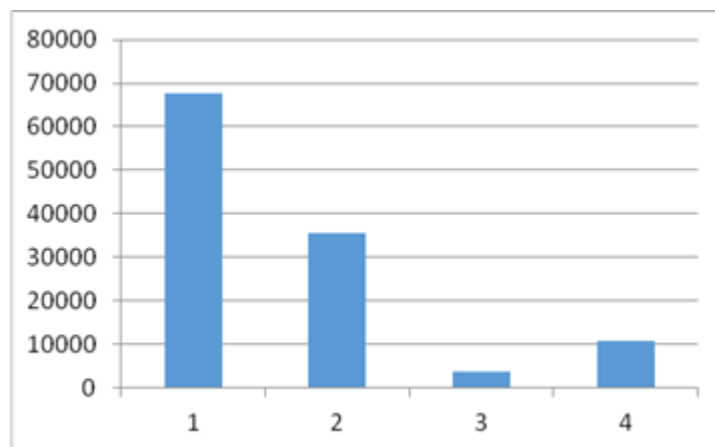


Figure 16 Prato synthetic population distribution according to marital status.

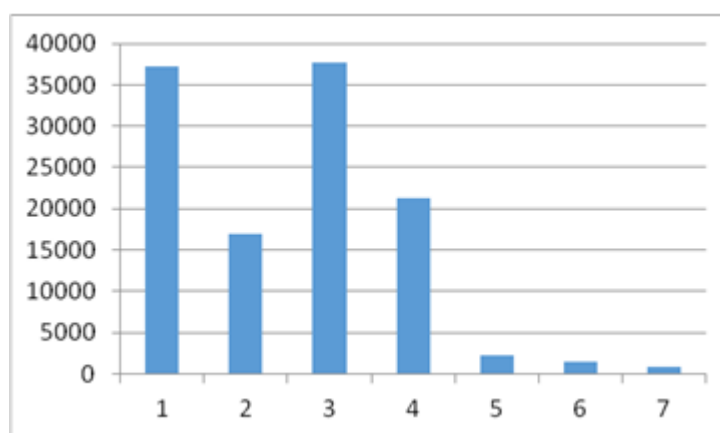


Figure 17 Prato synthetic population distribution according to income.

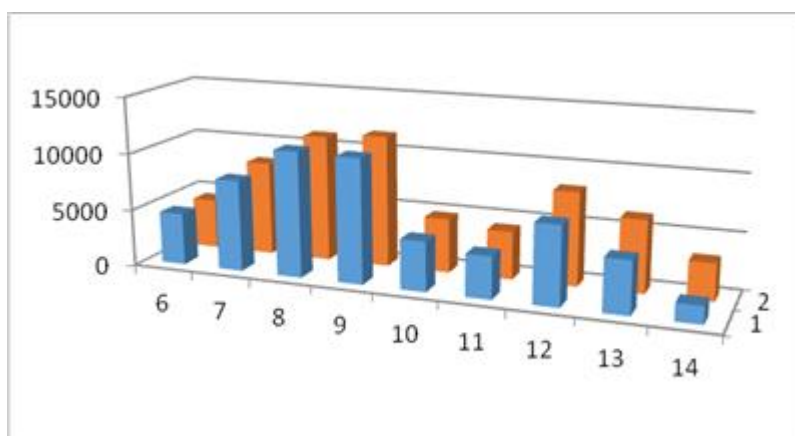


Figure 18 Prato synthetic population distribution according to age and gender.

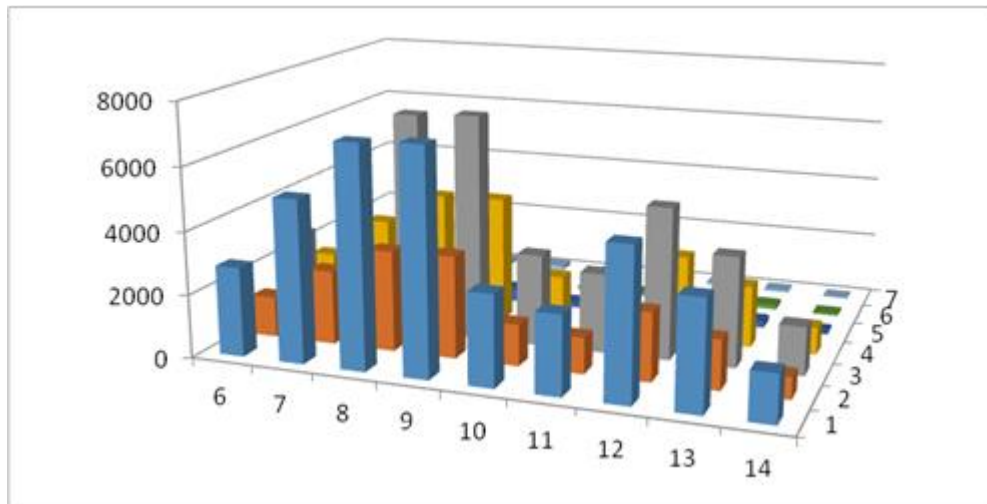


Figure 19 Prato synthetic population distribution according to age and income.

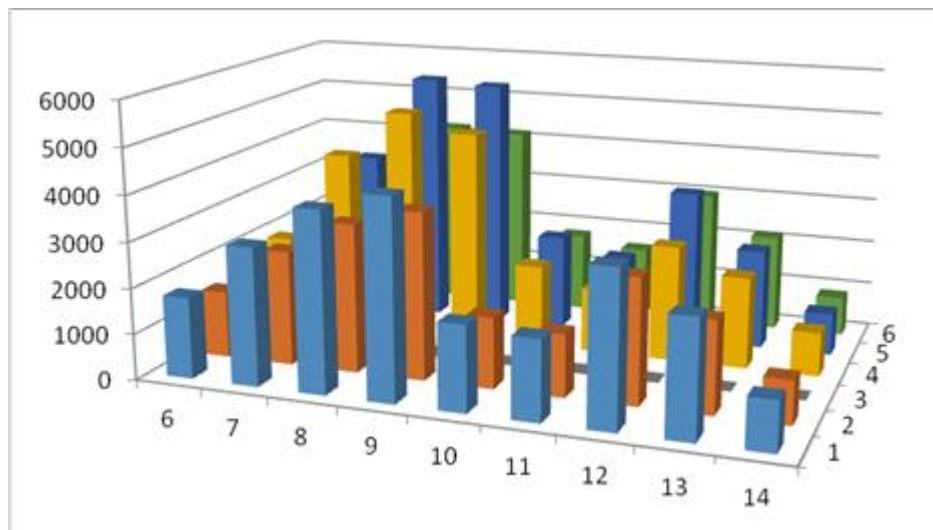


Figure 20 Prato synthetic population distribution according to age and district of residence.

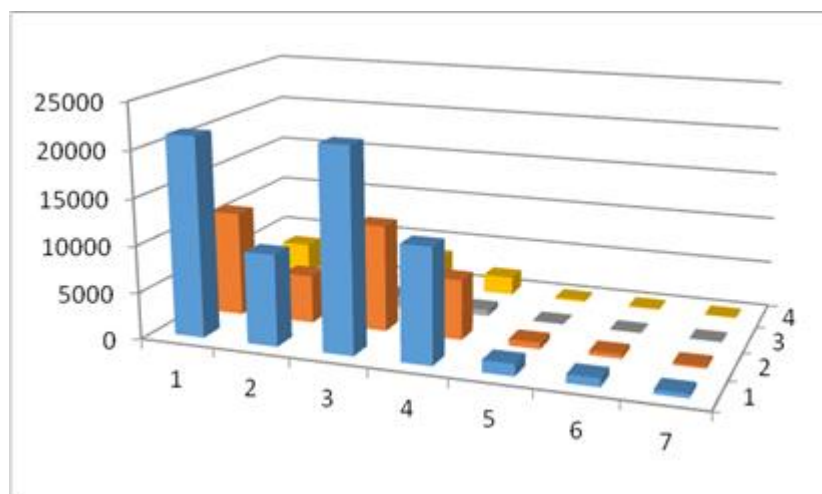


Figure 21 Prato synthetic population distribution according to income and marital status.

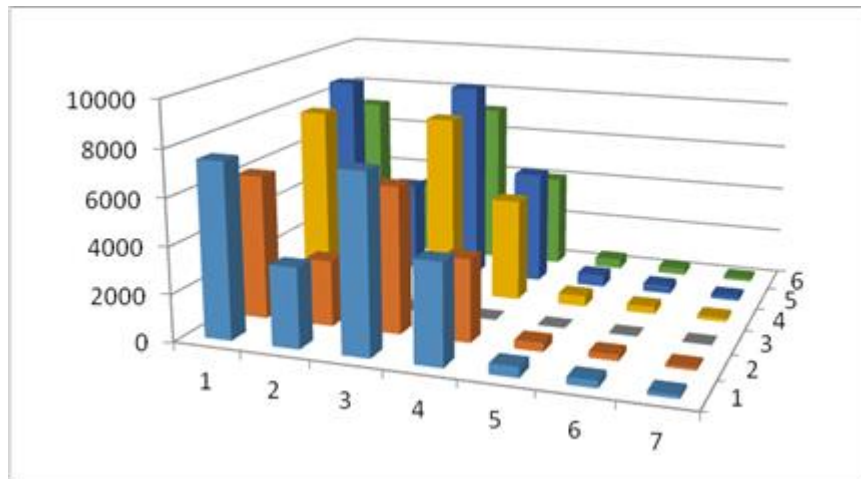


Figure 22 Prato synthetic population distribution according to income and district of residence.

6 RESULTS

This chapter consists of two sections. Firstly, scenarios for simulation experiments are presented. Secondly, results of multi-agent simulations are given and discussed. The experiments have been run on an HPC large scale computing cluster running Linux Ubuntu nodes and Oracle Java. For the analysis of the results the meta-modelling approach has been used, that was earlier discussed in the section 3.3

6.1 SIMULATION EXPERIMENTS

The simulation experiments in the alpha SIM are based on data provided by the pilot (municipality of Prato – see Chapter 5). The population of Prato town amounted to 191 thousand citizens at the end the year 2014. We have used the census data with the following socio-demographic features: city district of inhabitancy, sex, age category, occupation status, marital status and yearly income category. We have all the marginal distributions and additionally selected cross marginal distributions. Based on this a representative synthetic population of 2'480 citizens was generated to use in simulation experiments. The primary opinions are based on the observed socio-economic features and the trinomial model. In order to introduce representativeness bias in our model of preferences older citizens tend to vote for “yes” and the wealthier tend to vote “no”. We have used the trinomial model with 3 different opinion values, however, one can also allow for a continuous preferences distribution. In each step (we allow for limited number of steps in the simulation as it is more realistic assumption of user activity on the social platform) we may observe the preferences dynamics of the subpopulation (sample) and population.

For simulation experiments we consider a 5-dimensional parameter space (citizenClassVersion, edgesVersion, beta, primaryOpinionVersion, selectionVersion). The explanation of each parameter along with possible parameter values have been presented in the Table 2. A full parameter sweep is considered. The full Cartesian product of all parameters consists of 1536 points. For each parameter set we repeat simulation 30 times with different random seed values that determine the order in which opinions are updated within the social network.

Within the parameter space we consider 16 distinct scenarios (0-15) for subpopulation selection. The parameterization for those scenario is given in the Table 3. “Selection probability” is the snowball sampling probability that a particular node will be selected. Network propagation depth defines how many neighbour levels are used for sampling. The number of citizens in a sub network defines the initial selection size.

Parameter name	Parameter description /detailed model description can be found in the section 3.2/	Values considered in the parameter sweep
citizenClass	Type of decision making process for agents within the model. Three types are considered: {a, b, c}. For a detailed discussion see Section 3.2	(a) Mean neighbourhood opinion (b) Dominating opinion of neighbours (c) Polarizing opinion
edgesVersion	Type of edges propagation in the social network	0, 1
beta	β , such that $\beta \in (0,1)$ (homogenous for all the agents in the current implementation) represents the weight of the agent's own opinion to the opinions of the neighbours, the agents that the agent is connected to.	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
primaryOpinionVersion	Type of primary opinion	0, 1
selectionVersion	How a subpopulation of agents is selected?	0 - 15

Table 2 Parameter values. We perform a full Cartesian product parameter sweep. Hence the parameter space size is $3 \times 2 \times 8 \times 2 \times 16 = 1536$. For each data point 30 simulations are carried out that last over 6 periods on a population of 2840 agents.

Subpopulation selection type	Selection probability	Network propagation depth	Number of citizens in a sub network
0	0.3	1	30
1	0.3	1	40
2	0.3	1	50
3	0.3	1	60
4	0.3	2	30
5	0.3	2	40
6	0.3	2	50
7	0.3	2	60
8	0.4	1	30
9	0.4	1	40
10	0.4	1	50
11	0.4	1	60
12	0.4	2	30
13	0.4	2	40
14	0.4	2	50
15	0.4	2	60

Table 3 Parameterization for subpopulation selection procedure

6.2 OPINION DYNAMICS IN SOCIAL NETWORKS

The goal of this section is to present initial simulation results from the SIM alpha module. We start with illustrative results of a selected single simulation. Next we move to meta-modelling (see Section 3.3) across the entire parameter space.

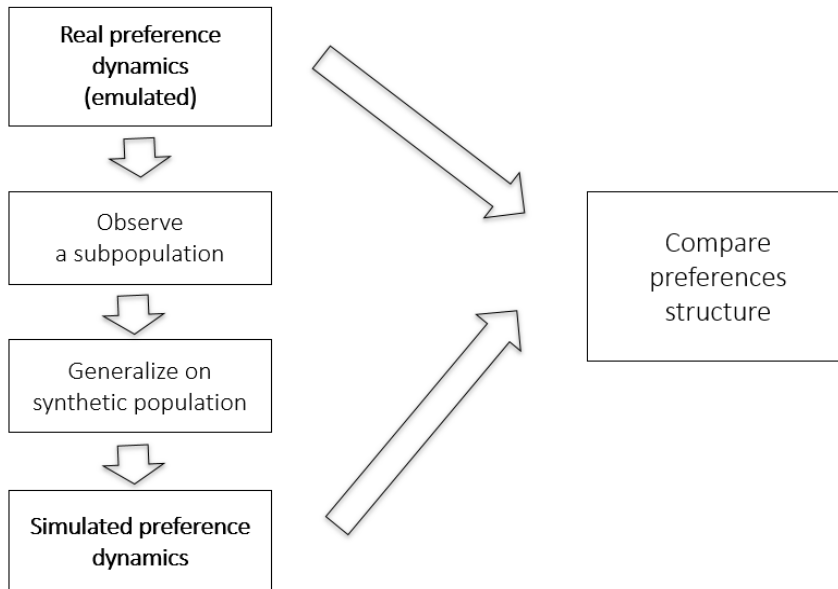


Figure 23 The goal of simulation experiments is to measure preference concordance between real preferences and simulated preferences. The concordance is a measure of validity of the taken approach for preference solicitation in the Open Data Governance Model.

6.2.1 SAMPLE SIMULATION RUN

Figure 24 and Figure 25 show that along the simulation course the preference concordance between real and synthetic population increases, while the average preference elicitation error is decreasing.

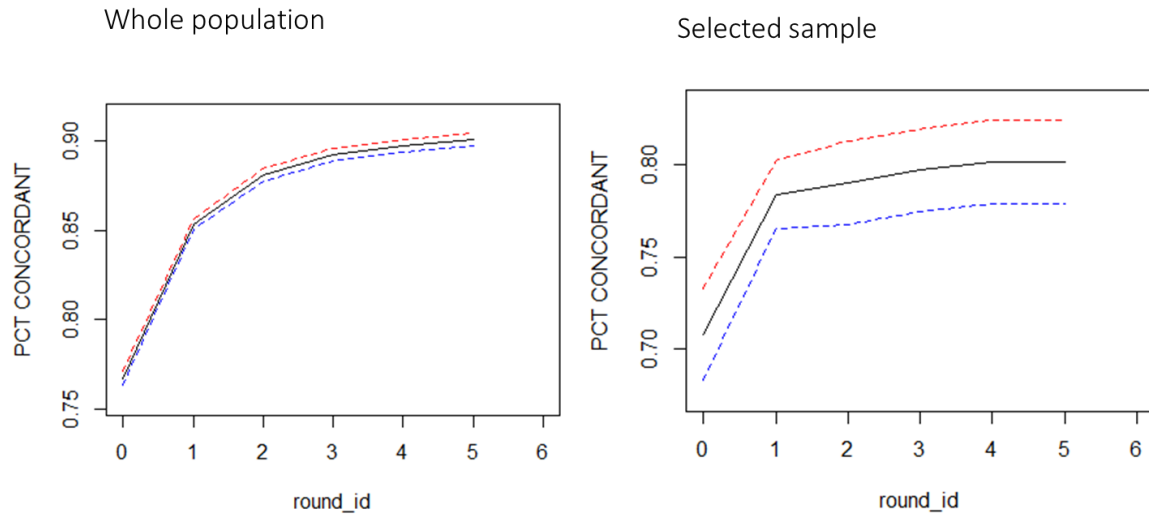


Figure 24 Example simulation results: Preference concordance increases in the whole population as well as in the selected subsample

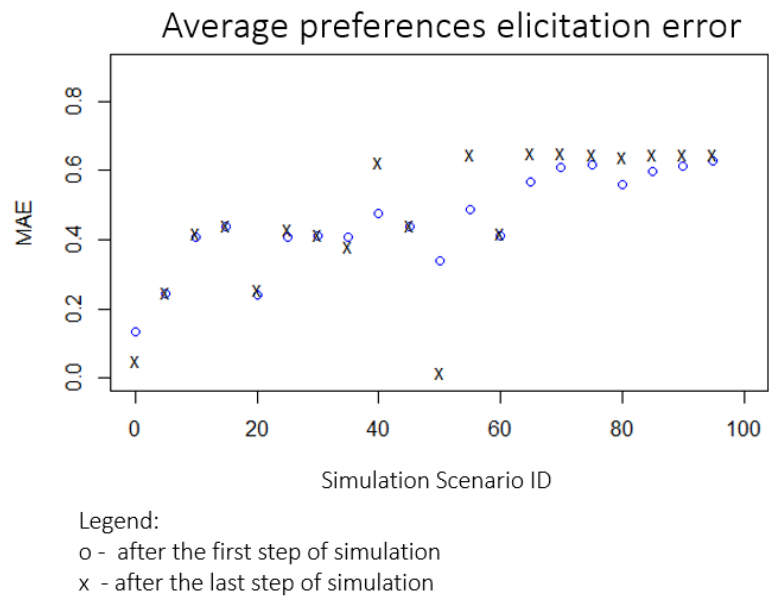


Figure 25 Example simulation results: Average preference elicitation error is decreasing through the course of simulation run

6.2.2 PARAMETER SWEEP – DETERMINANTS OF DYNAMICS RECONSTRUCTION

In the meta-analysis we will consider and compare two simulation states – at the beginning of simulation and at the end of simulation. We have 1536 parameterizations and 30 simulations for each parameterization which results in 46080 data points.

In this subchapter we analysed determinants of variable Sample_yes_PCT. Sample_yes_PCT represents part of the dynamics that was successfully reconstructed. The explanatory variables have been discussed and explained in the Section 3.2 and in the simulation experiment parameterization in Section 6.1.

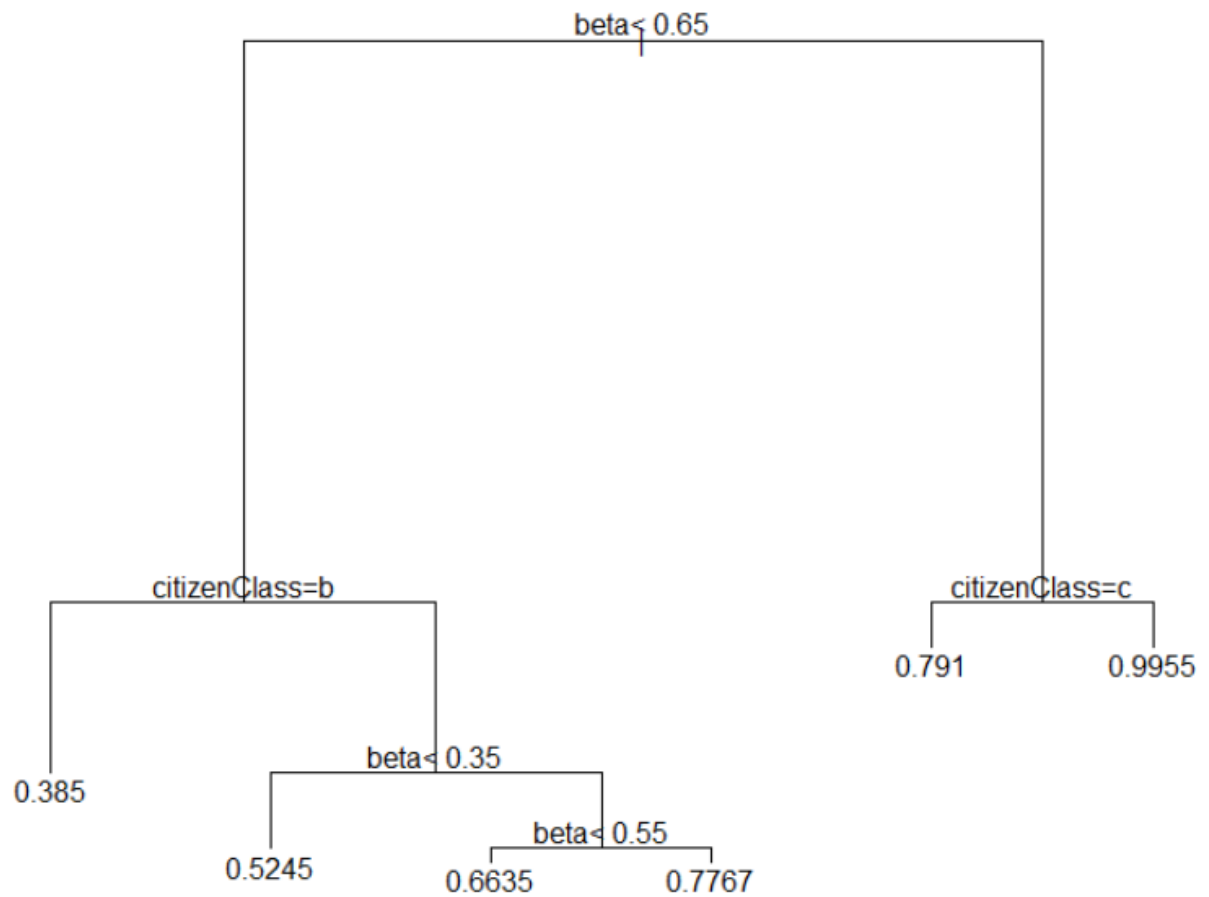


Figure 26 Decision tree across the parameter sweep and simulation runs at the beginning of simulation. The quality of population dynamics reconstruction is highest for large beta values (weight of agent's own opinion) and for citizen type not equal to "Polarizing opinion".

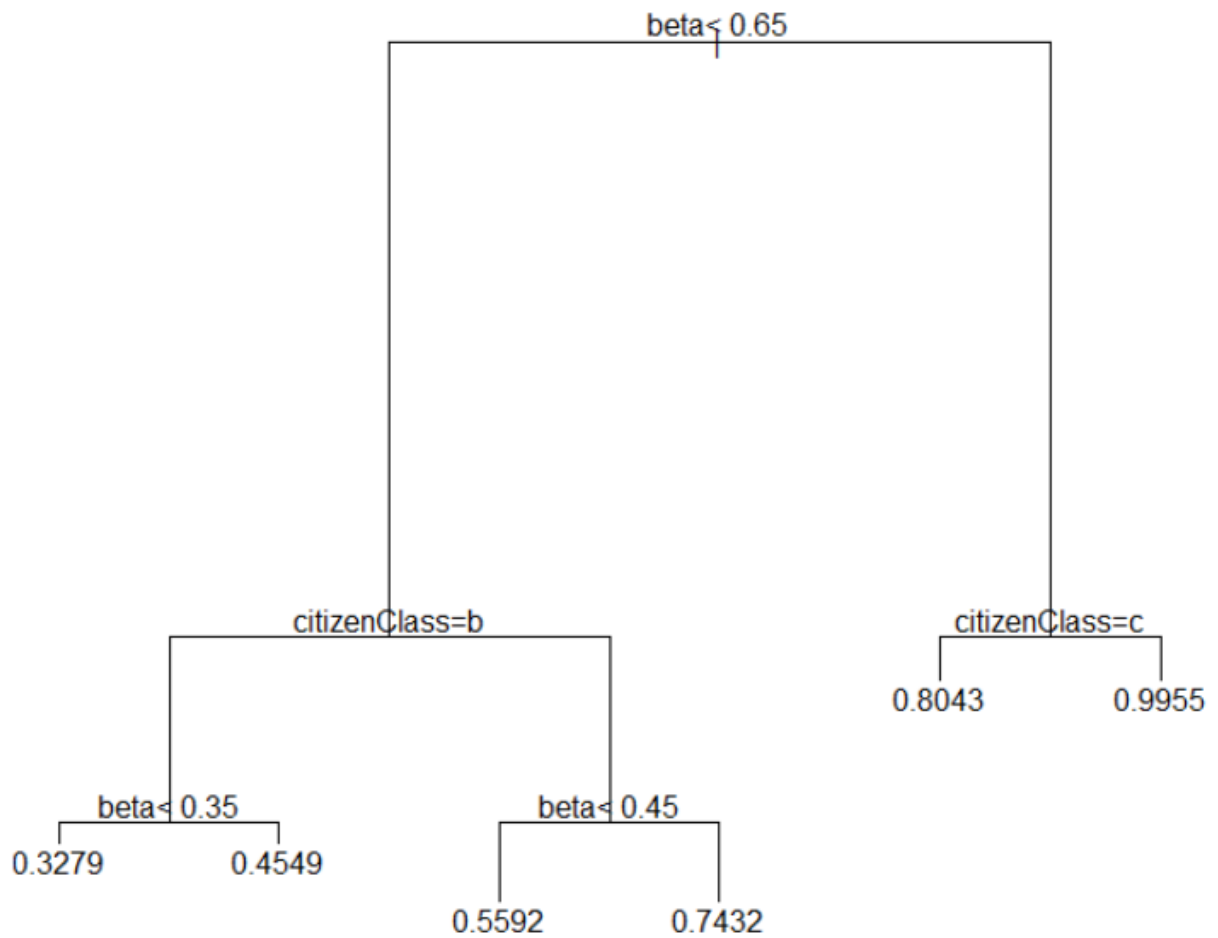


Figure 27 Decision tree across the parameter sweep and simulation runs at the end of simulation. The quality of population dynamics reconstruction is highest for large beta values (weight of agent's own opinion) and for citizen type not equal to "Polarizing opinion". Please note that for high betas the preference concordance values are higher than the ones at the beginning of simulation.

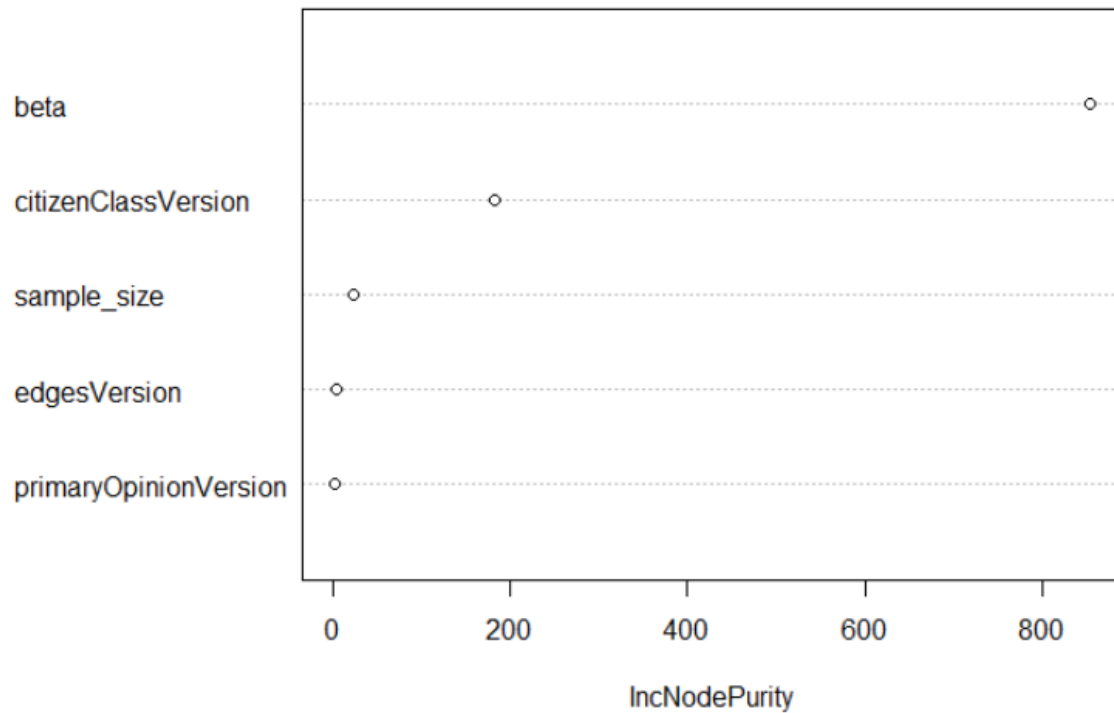


Figure 28 Random forest node purity at the beginning of simulation across the parameter sweep and simulations runs. It can be clearly seen that the preference concordance is determined by weight of agent's own opinion and type of opinion diffusion dynamics

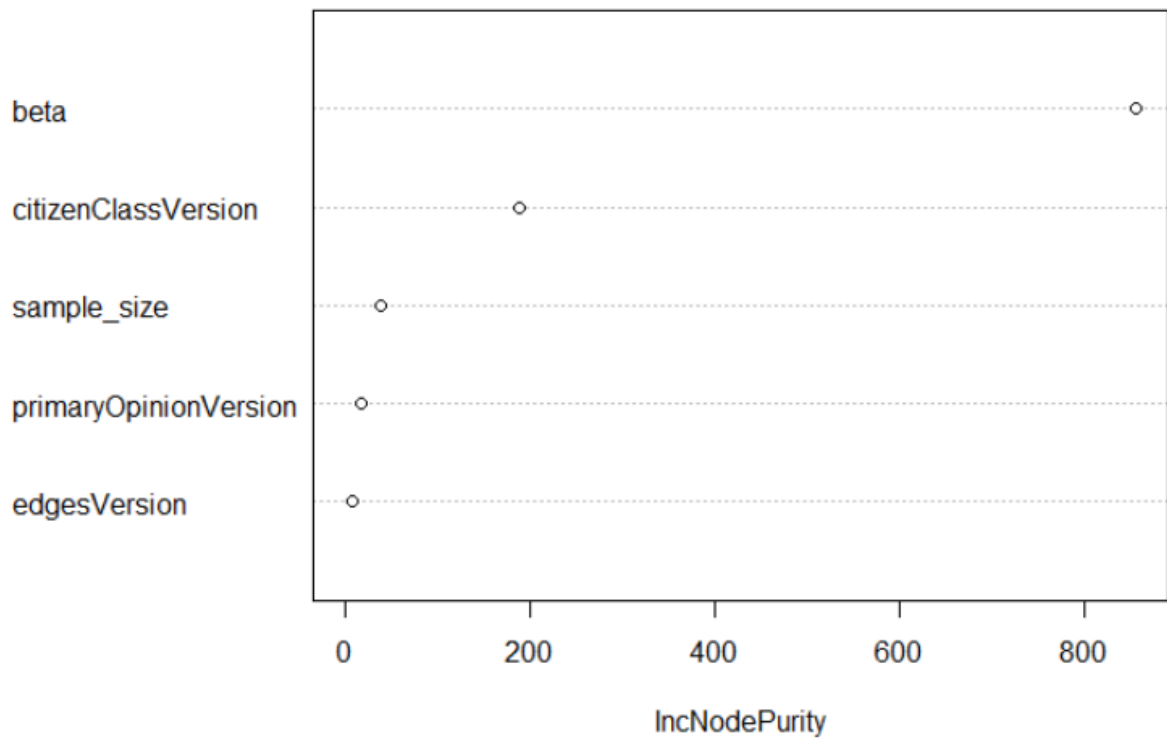


Figure 29 Random forest node purity at the end of simulation across the parameter sweep and simulations runs. Similarly to the previous graph it can be clearly seen that the preference concordance is determined by weight of the agent's own opinion and the type of opinion diffusion dynamics

Formula:
sample_yes_PCT ~ citizenClass + beta + edgesVersion + primaryOpinionVersion
+ sample_size

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.465e-01	2.198e-03	157.59	<2e-16	***
citizenClassDominating_opin.	-1.674e-01	1.478e-03	-113.25	<2e-16	***
citizenClassPolarizing_opin.	-4.712e-02	1.478e-03	-31.88	<2e-16	***
beta	7.363e-01	2.633e-03	279.61	<2e-16	***
edgesVersion1	1.187e-02	1.245e-03	9.53	<2e-16	***
primaryOpinionVersion1	6.159e-04	1.207e-03	0.51	0.61	
sample_size	8.474e-05	4.281e-06	19.79	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.667 Deviance explained = 66.7%
GCV = 0.016775 Scale est. = 0.016773 n = 46080

Figure 30 Linear model at the beginning of simulation. The main factor for opinion concordance is again the beta parameter (weight of agent's own opinion). "Dominating opinion" agent class leads to decrease in preference elicitation quality. Note that sample has a negligible effect on preference elicitation quality.

Formula:
sample_yes_PCT ~ citizenClass + beta + edgesVersion + primaryOpinionVersion
+ sample_size

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.345e-01	3.049e-03	109.704	< 2e-16	***
citizenClassDominating_opin.	-1.712e-01	2.050e-03	-83.514	< 2e-16	***
citizenClassPolarizing_opin.	-4.953e-02	2.050e-03	-24.167	< 2e-16	***
beta	7.441e-01	3.652e-03	203.753	< 2e-16	***
edgesVersion1	1.310e-02	1.727e-03	7.583	3.44e-14	***
primaryOpinionVersion1	3.356e-02	1.673e-03	20.054	< 2e-16	***
sample_size	8.463e-05	5.938e-06	14.253	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.518 Deviance explained = 51.8%
GCV = 0.032266 Scale est. = 0.032261 n = 46080

Figure 31 Linear model at the end of simulation. Beta more strongly determines preference elicitation quality. The gap between agent types is increasing. Note that sample has a negligible effect on preference elicitation quality.

6.2.3 PARAMETER SWEEP – PREFERENCE ELICITATION ERROR DETERMINANTS

In the meta-analysis we will consider and compare two simulation states – at the beginning of simulation and at the end of simulation. We have 1536 parameterizations and 30 simulations for each parameterization what results in 46080 data points.

In this subchapter we analysed determinants of the variable `pop_extreme_PCT`. The variable `pop_extreme_PCT` represents the percentage of the population for which the preference elicitation process failed – their opinions are misrepresented. In particular agents having “yes” opinion in real world have a “no” opinion in the simulated reconstructed synthetic population. The explanatory variables have been discussed and explained in the Section 3.2 and in the simulation experiment parameterization in Section 6.1.

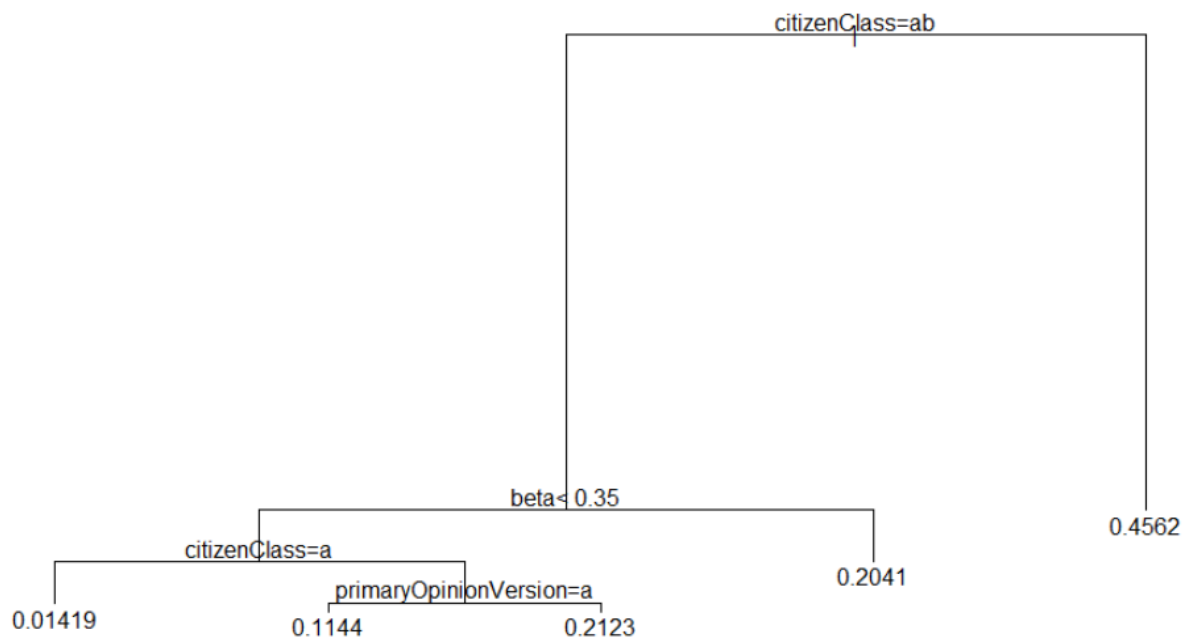


Figure 32 Decision tree across the parameter sweep and simulation runs at the beginning of simulation. The citizen class having “polarized opinion” dynamics is the type most difficult for exact elicitation of preferences.

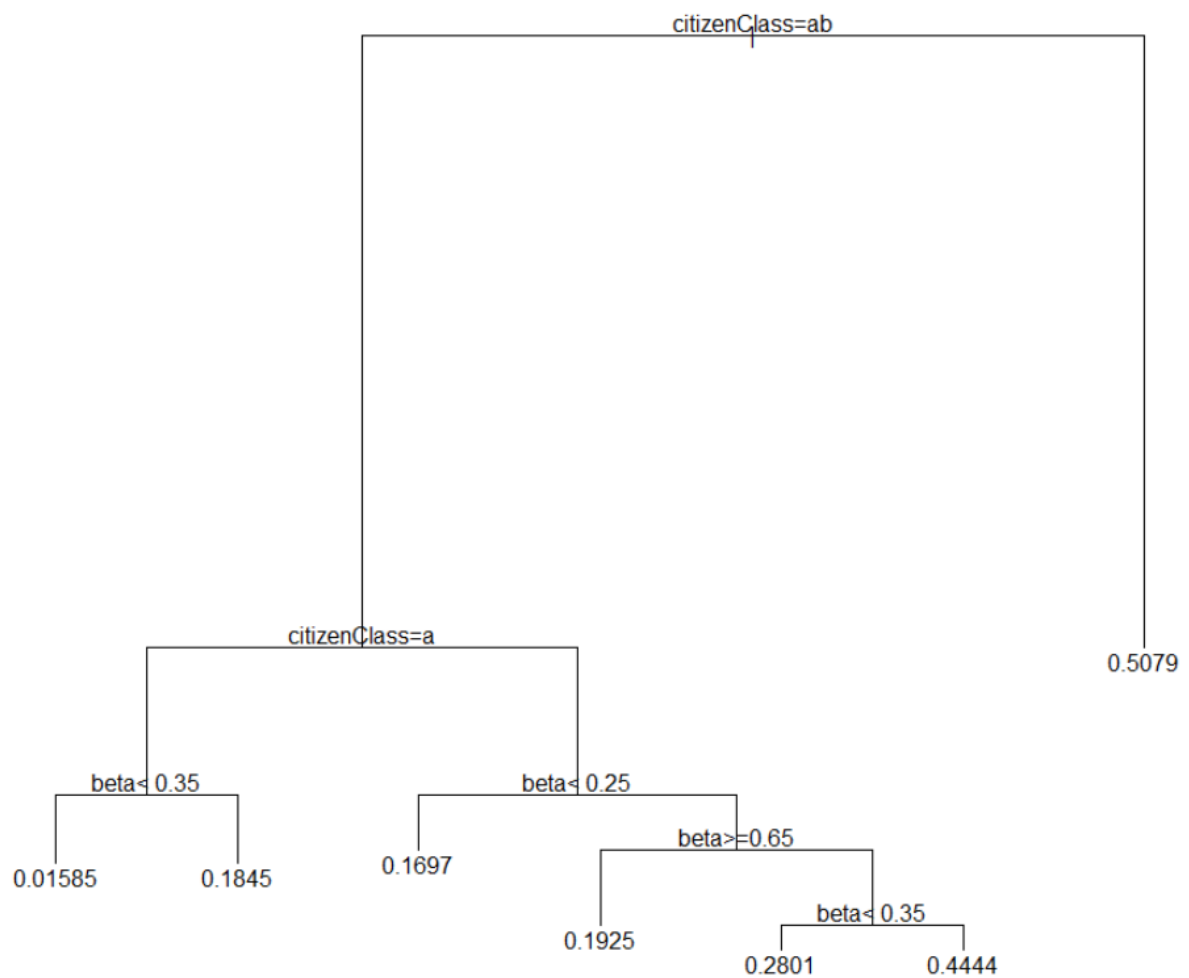


Figure 33 Decision tree across the parameter sweep and simulation runs at the end of simulation. The citizen class having “polarized opinion” dynamics type is the most difficult for exact elicitation of preferences and this difficulty has increased after simulations have been computed.

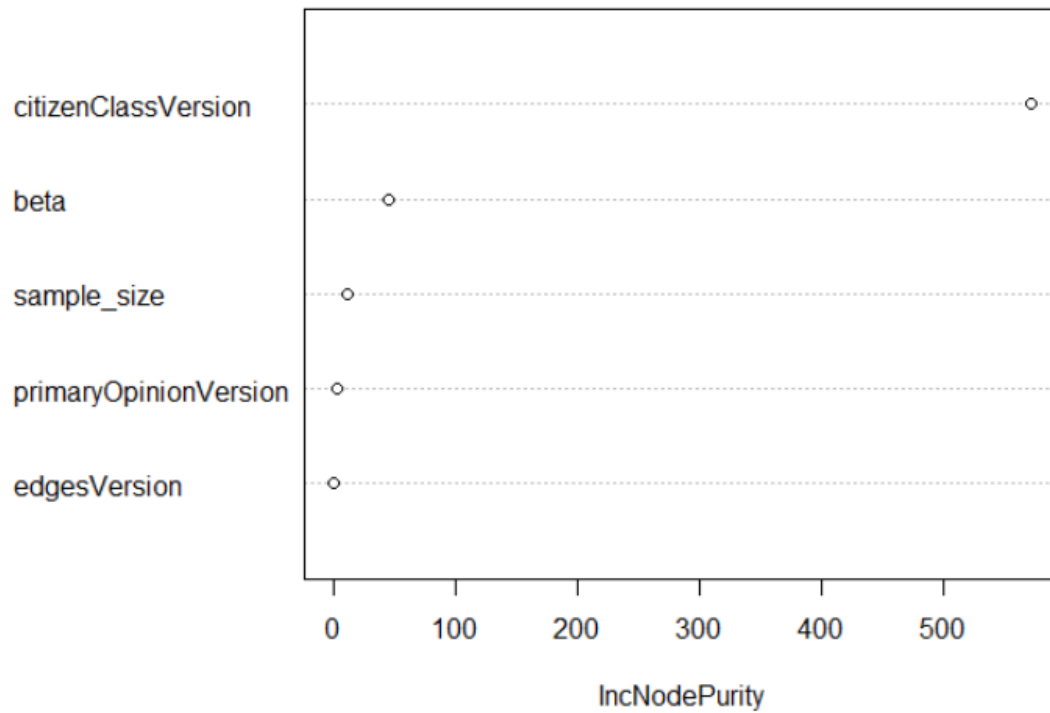


Figure 34 Random forest node purity at the beginning of simulation across the parameter sweep and simulations runs. It can be clearly seen that the preference elicitation errors are determined by the type of opinion diffusion dynamics more strongly than by the weight of the agent's own opinion.

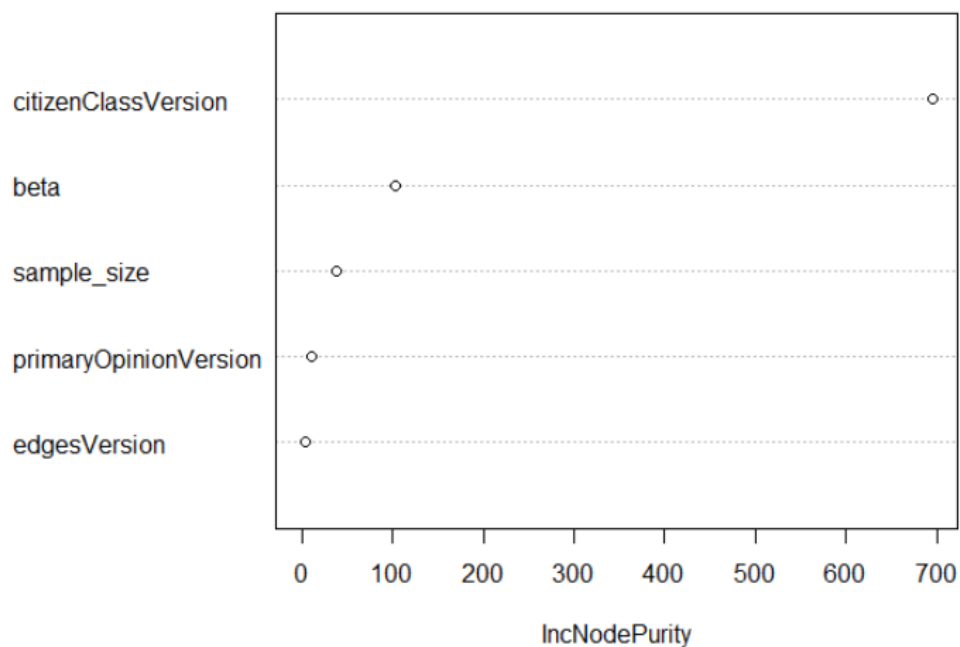


Figure 35 Random forest node purity at the end of simulation across the parameter sweep and simulations runs. It can be clearly seen that the preference elicitation errors are determined by the type of opinion diffusion dynamics more strongly than the weight of the agent's own opinion. The importance of opinion diffusion type and beta has increased.

```

Formula:
pop_extreme_PCT ~ citizenClass + beta + edgesVersion + primaryOpinionVersion
n + sample_size

Parametric coefficients:

              Estimate Std. Error t value
(Intercept)   6.796e-02  1.310e-03  51.873
citizenClassDominating_opin. 8.428e-02  8.807e-04  95.687
citizenClassPolarizing_opin. 3.374e-01  8.807e-04 383.115
beta           1.335e-01  1.569e-03  85.050
edgesVersion1 -3.197e-03  7.422e-04  -4.308
primaryOpinionVersion1 8.727e-03  7.191e-04  12.136
sample_size   -4.642e-05  2.551e-06 -18.192

Pr(>|t|)
(Intercept)   < 2e-16 ***
citizenClassDominating_opin. < 2e-16 ***
citizenClassPolarizing_opin. < 2e-16 ***
beta           < 2e-16 ***
edgesVersion1  1.65e-05 ***
primaryOpinionVersion1 < 2e-16 ***
sample_size    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.784   Deviance explained = 78.4%
GCV = 0.0059582   Scale est. = 0.0059573   n = 46080

```

Figure 36 Linear model at the beginning of simulation. The main factor for opinion elicitation error is again the opinion diffusion dynamics type. The beta parameter (weight of the agent's own opinion) is of a lesser importance. Note that sample size has a negligible effect on preference elicitation quality.

```

Formula:
pop_extreme_PCT ~ citizenClass + beta + edgesVersion + primaryOpinionVersion
+ sample_size

Parametric coefficients:

              Estimate Std. Error t value
(Intercept)  1.031e-01  3.002e-03  34.347
citizenClassDominating_opin. 1.710e-01  2.018e-03  84.763
citizenClassPolarizing_opin. 3.866e-01  2.018e-03 191.611
beta          7.930e-02  3.595e-03  22.058
edgesVersion1 -6.371e-03  1.700e-03  -3.747
primaryOpinionVersion1 -5.007e-03  1.647e-03  -3.039
sample_size   -4.563e-05  5.845e-06  -7.806

Pr(>|t|)
(Intercept)  < 2e-16 ***
citizenClassDominating_opin. < 2e-16 ***
citizenClassPolarizing_opin. < 2e-16 ***
beta          < 2e-16 ***
edgesVersion1 0.000179 ***
primaryOpinionVersion1 0.002375 **
sample_size   6.05e-15 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.448   Deviance explained = 44.8%
GCV = 0.031271   Scale est. = 0.031266   n = 46080

```

Figure 37 Linear model at the end of simulation. The main factor for opinion elicitation error is again the opinion diffusion dynamics type. The beta parameter (weight of the agent's own opinion) is of a lesser importance. Note that the importance of opinion dynamics type in determining preference elicitation has increased.

The simulation results show the representativeness bias increases while the opinion becomes more homogenous. Weight of user's own opinion (beta parameter) is the most important factor in the elicitation of average preferences for the entire population. However, when we consider preference elicitation errors the most crucial parameter is the type of information diffusion dynamics.

7 CONCLUSION

We presented an alpha version of open data governance model (ODGM). The goal of ODGM is to help Public Administration (PA) provide information about citizen activity on the SPOD, and in particular to design an efficient system for elicitation of social preferences in heterogonous communities. The social platform for open data (SPOD) allows citizens to monitor allocations and spending of financial resources, controlling PA and hence increasing its efficiency. We assume that the PA shares the information with citizens and thus PA actions are shaped by a participatory society. We have considered a scenario where Public Administration (PA) uses an online social platform to give access for citizens to open data and collects information on their preferences. Apart from the core functionality, provided in SPOD, of giving access to the data, visualizing the data and allowing for discussions and collaboration of citizens and PA about the data there is a need on the PA side (and possibly also third party NGOs) for first understanding the data usage

The SIM deliverable outlines the implementation of an open data governance model developed to support SPOD deliverable. The SIM module provides the information about citizen activity on SPOD. The notion of activity encompasses such data as: citizen logging in to SPOD, which datasets they view, intensity of discussions on SPOD, citizens' preferences revealed on SPOD, social links revealed and formed on SPOD. The report contains results of discussions with pilot participants of ROUTE-TO-PA project. Those results will be used in the preparation of graphical interface in the final deliverable.

Another important result is construction of an alpha versions of the tools for optimal design of preference elicitation and aggregation systems with heterogeneity in citizen geographical location and demographic structure. The tools are delivered as a multi-agent simulation model. In our modelling approach we have assumed that an economic system consists of interacting heterogeneous agents and hence we have considered a socioeconomic system with the agents representing members of the local community. The agent-based simulation model, presented in the report has been implemented in Java using MASON for controlling the simulation.

The goal of the report is to present a method for preference elicitation that is robust to selection bias. In particular, we propose a method for analysing dynamics of entire population preferences based on the observed preferences of the limited sub-population. In theoretical models we assume that the opinion diffusion process takes place in the entire population. However, a PA can only observe a sample sub-population. In this approach unobserved population members influence the observed information diffusion. Moreover, we assume that opinion diffusion has the same dynamics on the subsample and in entire population.

The simulation tool has been calibrated it with empirical data from the Prato population. In further analysis we will include data from another pilot as well empirical data from the SPOD platform in the analysis. The developed tool enables the finding and analysis of optimal mechanism design for sharing information on SPOD required to use modelling tools that allow the analysis of the heterogeneity of economic agents, their geographic location, virtual and real-world social networks and information flow (including data comprehension) within those networks.

In this report we have performed 46,080 simulation runs for an artificial population of 2,840 heterogeneous agents. Within those runs 1,536 virtual society parametrizations have been considered. The population structural properties have been calibrated with empirical data from one of pilots - the Prato municipality. The simulation results show determinants for successful generalisation of preferences from a subpopulation onto the entire society. The simulation results show that representativeness bias in a population increases when the opinion becomes more homogenous. Moreover, simulation results network connectivity and importance of agent's own opinion are major determinants of quality of the preference elicitation process.

8 APPENDIX A – PILOT REQUIREMENT FOR SIM DELIVERABLE FUNCTIONALITY

The goal of the meetings with Prato and Dublin was to discuss reports that will be delivered by SIM module on the SPOD platform.

Prato list of participants:

	Organisation	Participants
1	SGH Warsaw School of Economics (Poland)	Przemyslaw Szufel
2	SGH Warsaw School of Economics (Poland)	Marcin Czupryna
3	Comune di Prato	Paolo Boscolo
4	Comune di Prato	Elena Palmisano

Dublin list of participants:

	Organisation	Participants
1	SGH Warsaw School of Economics (Poland)	Przemyslaw Szufel
2	SGH Warsaw School of Economics (Poland)	Marcin Czupryna
3	Dublin City	Pauline Riordan (excused)
4	Dublin City	Nicola Graham (excused)
5	Dublin City	Brendan Fahy

During the online meeting it has been agreed that the following report types will be available for the public administration

8.1 GLOBAL DATA ON THE SPOD PLATFORM

Data in the reports will grouped by rooms within the platform.

Report type	Report priority 1 – TOP 2 – HIGH 3 – MEDIUM
a. Number of people viewing/commenting over time	1
b. Statistics on emoticons	1

c.	Number of comments	1
d.	Number of unique users	1
e.	Distributions of user data (data included in the registration process – e.g. gender, age etc)	1
f.	Users most active (by number of visits, by number and type of emoticons)	1
g.	Most popular dataset visualisation /rooms over time	1

8.2 DATA ON ROOM LEVEL

Data presented grouped by discussion within a room.

Report type	Report priority 1 – TOP 2 – HIGH 3 – MEDIUM
a. Number of people viewing & commenting over time	2
b. Statistics on emoticons – emotions over dataset visualisation	2
c. Number of comments	2
d. Number of unique users	2
e. Distributions of user data (data included in the registration process – e.g. gender, age etc)	2
f. Users most active around a given dataset visualisation or in given room (by number of visits, by number and type of emoticons)	2
g. Activity over data set as a percentage of overall activity	2

8.3 REPORTING BASED ON MERGING SPOD REGISTRATION DATA WITH LOCAL CENSUS DATA

Please note that introducing reports below requires that personal attributes will be present both in census data and the given by the users in registration process.

Report type	Report priority 1 – TOP 2 – HIGH 3 – MEDIUM
a. Emoticons room & discussion – emotions for entire population corrected by data given in the registration vs census data, each blog item – extrapolation, discussion started in each agora. Each item can be marked with emoticon – each like/dislike generalized on the population	3
b. Opinions – voting results generalized on the entire population	3

9 APPENDIX B – SIMULATION ALGORITHM DETAILS

This appendix contains the detailed description of multi-agent simulation for elicitation of preferences presented in Chapter 3. The logic in agent-based simulation models can be fully exposed fully through the sourcecode. The authors in agent-based modelling literature agree that the source code of a multi-agent model is an important part of it's documentation. A detailed discussion of the role of source code in documenting agent-based simulation models can be found for example in Gilbert (2008), Law (2006) and Miller (2007). In this report we take the same approach and fully represent simulation model source code. A detailed discussion of the developed model along with the UML diagrams for the developed classes can be found in the Chapter 4 of the report.

In the remainder of this section we explain details of steps 1 – 6 from the outlined procedure in Chapter 3.

Steps 1 and 2.

Reconstruction of edges is carried out between each agent $v \in V^{NS}$, i.e. between each agent who is not a platform user, and all the other agents $u \in V^P$, such that $u \neq v$, i.e. both those agents, who are platform users, and those, who are not. Therefore, edges that are known *a priori*, i.e. edges in E , remain as they empirically are and are not reconstructed.

Model M_E is estimated on the basis of data available for agents $v \in V^S$, and then it predicts probabilities $p_{v,u}$, that an edge (v, u) exists between an agent $v \in V^{NS}$ and an agent $u \in V^P$ such that $u \neq v$. Probability of such an edge between v and u is inferred using data on $d(v)$ and $d(u)$. Several approaches can be chosen. We applied a logistic regression model:

$$y_{v,u} = \frac{\exp(\theta x_{v,u})}{1 + \exp(\theta x_{v,u})}$$

where $y_{v,u} = P((v, u) \in E^*)$ denotes a probability, that a link between agent v and u is formed, which, in the process of estimation, is approximated by:

$$y_{v,u} = \begin{cases} 1, & \text{if } (v, u) \in E \\ 0, & \text{if } (v, u) \notin E \end{cases}$$

i.e., by a binary variable which indicates which agents in V are connected by an edge and:

$$x_{v,u} = f(d(v), d(u))$$

where $f: R^n \times R^n \rightarrow R^{m_E}$ is a function, which constructs an m_E -dimensional vector of explanatory variables for the logistic regression.

More specifically, we assume that, $m_E = n$ and that:

$$f(x^1, x^2) = |x^1 - x^2|$$

The chosen function enables us to take into consideration the degree the similarity influences the probability of the link existence between citizens.

Once a logistic regression model has been estimated, predicted probabilities $p_{v,u}$ can be derived. Then for all agents $v \in V^{NS} = V^P \setminus V^S$ the edges of the form (v,u) , such that $u \in V^{NS}$ and $u \neq v$ are reconstructed according to the estimated probabilities. The process is simulated many times.

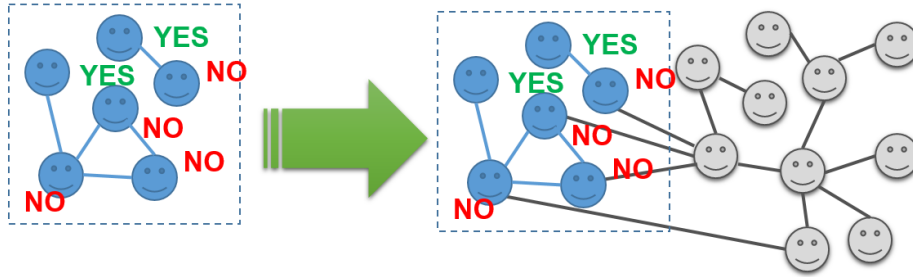


Fig. 1. Using data on platform users (blue agents), edges between agents who don't use the platform (grey), as well as edges between agents who don't use the platform and agents who use the platform are reconstructed (these are represented by solid grey lines).

Steps 3 and 4.

We assume that the opinion $o(v,r)$ can have a threefold value, i.e. -1, 0 or 1. For each agent v model M_o predicts three probabilities (with the sum of 1), corresponding to each of the three possible values of $o(v,r)$. Let us denote these probabilities by $p_v(x)$, where $x \in \{-1,0,1\}$. Model M_o is considered as an *a priori* opinion constructor in the sense, that it does not take into account any information on the network structure (on connections between agents), but it uses only information on an agent's characteristics, as contained in $d(v)$. If data in $d(v)$ is indeed discriminative with respect to $o(v,p)$, a model M_o can be a good predictor for opinions of agents in V^{NS} , i.e. for platform non-users.

As an opinion constructor we use a 3-nomial logistic regression model of the form:

$$y_{v,p}(k) = \frac{\exp(\gamma_k x_{v,p})}{\sum_{k \in \{-1,0,1\}} \exp(\gamma_k x_{v,p})}$$

where: $y_{v,p}(k) = P(o(v,r) = k)$ denotes a probability that agent's v opinion in round r on post p is k , where $k \in \{-1,0,1\}$, which, in the process of estimation, is approximated by:

$$y_{v,p}(k) = \begin{cases} 1, & \text{if } o(v,p) = k \\ 0, & \text{if } o(v,p) \neq k \end{cases}$$

i.e. by a binary variable which indicates what opinion is formed by an agent $v \in V$ on a post p , and:

$$x_{v,p} = f(d(v))$$

where $f: R^n \rightarrow R^{m_o}$ is a function, which constructs an m_o -dimensional vector of explanatory variables for agent v . More specifically, we assume that:

$$f(x) = (x^T, 1)$$

and that we have $m_o = n + 1$ and $\gamma_k \in R^{m_o}$, $k \in \{-1, 0, 1\}$, are vectors of parameters to be estimated using a maximum a posterior estimation method.

To attach an agent her/his opinion, we repeatedly simulate the primary opinions according to the estimated probabilities.

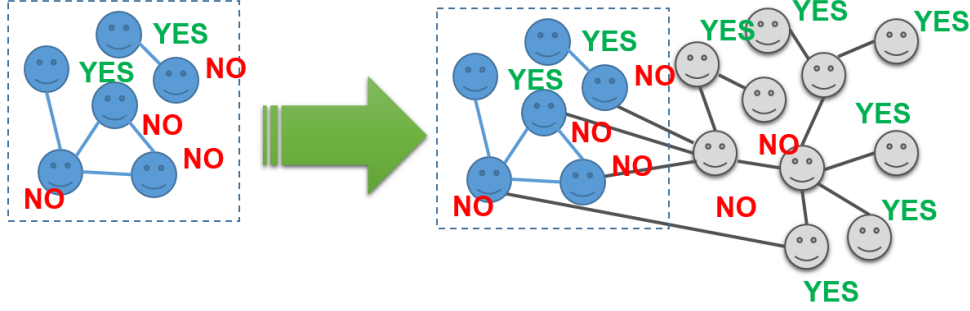


Fig. 2. Using data on platform users (blue agents), initial opinions of agents who don't use the platform (grey) are reconstructed.

Step 5.

At this stage for all agents in V^P and edges, or lack of thereof, is assumed between any pair of agents in V^P , therefore the structure of G^P has been fully built. Note, that primary opinions for agents in V^S are assumed not using the model M_o , but using empirical data on opinions $o(v, 0)$ for $v \in V$. Also edges between any pair of agents in V , i.e. edges $(v, u) \in E$, are not determined by the model M_E , but are assumed using empirical data on E^S . As a consequence, after the reconstruction of opinions and edges has been carried out, the empirical network structure G^S is extended by a predicted network structure and these two networks become interconnected. Therefore, the implied structure of G^P consists partially of empirical data and partially of data which was predicted or reconstructed by models M_E and M_o .

Once opinions have been assigned to agents and connections between them have been established, we simulate the synthetic population represented by G^P and $o(v, r)$, $\forall r \in 1, \dots, n$ using an opinion diffusion algorithm which represents:

- the way in which an opinion of any agent $v \in V^P$ is influenced by an opinion of any agent $u \in V^P$, $u \neq v$, such that $(v, u) \in E^P$,
- the way in which an opinion of any agent $u \in V^P$ influences an opinion of any agent $v \in V^P$, $u \neq v$, such that $(u, v) \in E^P$.

The algorithm is implemented in 3 different versions (to take into consideration that different opinion updating procedures are possible) according to the following rules, defined below. We also assume that parameter β , such that $\beta \in (0, 1)$ (homogenous for all the agents in the current implementation) represents the weight of the agent's own opinion to the opinions of the neighbours, the agents that the agent is connected to. We simulate for different values of β and different updating rules

- For each agent $v \in V^P$ let $e(v) = \{u \in V : (v, u) \in E^P\}$ denote a set of agents with which agent v is connected in G^P .

- b) Let $\pi(v) = (\frac{n_k}{n}, k \in \{-1, 0, 1\})$ where $n_k = \sum_{u \in \mathcal{S}(v)} \chi_{o(u,r)=k}$ where:

$$\chi_{o(u,r)=k} = \begin{cases} 1, & \text{if } o(u, r) = k \\ 0, & \text{if } o(u, r) \neq k \end{cases}$$

and $n = \sum_{k \in \{-1, 0, 1\}} n_k$. Consecutive elements of a vector $\pi(v) = (\frac{n_{-1}}{n}, \frac{n_0}{n}, \frac{n_1}{n})$ represent fractions of agents in $\mathcal{S}(v)$ which opinion is -1 , 0 and 1 respectively, whereas n_{-1} , n_0 and n_1 represents the numbers of agents (friends) with negative, neutral or positive opinion.

- c) Let us arrange agents in V^P in a random order and let such a generic order be represented by a vector $w = (v_1, v_2, \dots, v_{|w|})$.
- d) For consecutive elements of w , update opinion of agent v_j , $j = 1, 2, \dots, |w|$, according to the following rules:

$$o(v_j, r+1) \leftarrow k^*$$

where:

- i. Agent type I: **Mean neighbourhood opinion**

$$s = \beta \times o(v_j, r) + (1 - \beta) \times \frac{n_1 - n_{-1}}{n} \text{ and}$$

$$k^* = \begin{cases} -1, & s < -0,33 \\ 0, & -0,33 \leq s \leq 0,33 \\ 1, & s > 0,33 \end{cases}$$

- ii. Agent type II **Dominating opinion of neighbours**

$$s = \beta \times o(v_j, r) + (1 - \beta) \times o(max, r)$$

$$o(max, r) = \begin{cases} -1, & n_{-1} = n_{max} \wedge n_{-1} \neq n_1 \\ 0, & n_0 = n_{max} \vee n_{-1} = n_1 \\ 1, & n_1 = n_{max} \wedge n_{-1} \neq n_1 \end{cases}$$

$$n_{max} = \max(n_{-1}, n_0, n_1)$$

$$k^* = \begin{cases} -1, & s < -0,33 \\ 0, & -0,33 \leq s \leq 0,33 \\ 1, & s > 0,33 \end{cases}$$

- iii. Agent type III: **Polarizing opinion**

$$k^* = \text{sign}(\beta \times 10 \times o(v_j, r) + n_1 - n_{-1})$$

The steps a)-d) are repeated for each round r .

Step 6.

At this step we compare simulated opinions dynamics (for each value of parameter β and each of opinion updating versions generated opinions observable on the $v \in V^S$).

Similarity measures can be defined in the following way:

$$\rho_1 = \sum_{v_j \in V^S} 1(o(v_j, n) = o^{observed}(v_j, n))$$

Or

$$\rho_2 = \sum_{r=1, \dots, n} \sum_{v_j \in V^S} 1(o(v_j, r) = o^{observed}(v_j, r))$$

Where $1(.)$ represents standard identity operator and $o^{observed}(v_j, r)$ denotes the opinion expressed by the citizen v_j and observed on the platform in round r .

A version of opinion dynamics and parameter value β is chosen that maximizes the similarity measures ρ_1 or ρ_2 . So we choose such opinion dynamics that the pair wise concordance of the observed opinions and the simulated opinions observed on the sample of SPOD users is the highest.

For this particular dynamics on the whole synthetic population $v \in V^S$ we take $o(v_j, n)$ - which represents final reconstructed opinion as the generalization of the opinions observed in the sample.

10 BIBLIOGRAPHY

1. Acemoglu, D. and A. Ozdaglar, (2011) Opinion Dynamics and Learning in Social Networks, *Dynamic Games and Applications*, p.3-49.
2. Ashraf, Q., Gershman, B. and Howitt, P. (2011). Market Organization and Macroeconomic Performance: An Agent-Based Computational Analysis. NBER Working Paper No. 17102.
3. Axtell, R.L. (2007). What economic agent do: How cognition and interaction lead to emergence and complexity. *Review Austrian Economics*, 20, 105-122.
4. Barrat A., M. Barthelemy, A. Vespignani (2008) *Dynamical Processes on Complex Networks*. Cambridge University Press.
5. Barthelemy J., P. Toint (2012) Synthetic Population Generation Without a Sample, *Transportation Science*
6. Barton R. R., Metamodels for simulation input-output relations, in: J. Swain, D. Goldsman, R. Crain, J. Wilson (Eds.), (1992) *Proceedings of the 1992 Winter Simulation Conference*, IEEE, 1992, pp. 289-299.
7. Bratley P., Fox B.L. (1988); Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator, *ACM Transactions on Mathematical Software* 14, 88–100
8. Butts C. (2003) Network inference, error and informant (in)accuracy: a Bayesian approach, *Social Networks* 25, pp.103-140
9. Darley V., Outkin A. (2007), *Nasdaq Market Simulation: Insights on a Major Market from the Science of Complex Adaptive Systems*, World Scientific Publishing Company.
10. DeGroot, M. H. (1977) Reaching a Consensus. *Journal of the American Statistical Association*, 69, 118–121
11. Diao S-M., Y. Liu, Q-A Zeng, G_X Luo and F. Xiong, (2014) A novel opinion dynamics model based on expanded observation ranges and individuals' social influences in social networks *Physica A*, 220-228
12. Dosi, G., Fagiolo, G. and Roventini, A. (2006). An Evolutionary Model of Endogenous Business Cycles. *Computational Economics*, 27(1), pp. 3-34.
13. Fagiolo, G. 1998. Spatial interactions in dynamic decentralized economies: a review. In: P. Cohendet, P. Llerena, H. Stahn, and G. Umbhauer, ed., *The Economics of Networks: Interaction and Behaviours*, Springer Verlag, Berlin - Heidelberg.
14. Fagiolo, G., Windrumy, P. and Monetaz, A. (2007). A Critical Guide to Empirical Validation of Agent-Based Economics Models: Methodologies, Procedures, and Open Problems. *Computational Economics*, 30(3), pp. 195-226.
15. Farine D., Strandburg-Peshkin (2015), Estimating uncertainty and reliability of social network data using Bayesian inference, *Royal Society for Open Science*
16. Farmer D.J., D. Foley (2009), The economy needs agent-based modelling. In *Nature*, vol. 460, pp. 685-686.
17. Fischhoff B., Manski C. F. (2000). Elicitation of Preferences. In *Journal of Risk and uncertainty*, vol. 19: 1-3.
18. Frank O. (1974) Survey sampling in graphs. *Journal of Statistical Planning and Inference*, vol. 1, pp. 235–64
19. Frank O., (1981), *Survey of Statistical Methods of Graphs Analysis*, *Sociological Methodology*
20. Frick M., K. Axhausen (2004), *Generating Synthetic Populations Using IPF and Monte-Carlo Techniques: Some new Results*, Conference Paper STRC 2004

21. Gaffeo, E., Gatti, D.D., Desiderio, S. and Gallegati, M. (2008). Adaptive Microfoundations for Emergent Macroeconomics, *Eastern Economic Journal*. 34(4), pp. 441-463.
22. Gajdos T., Tallon J.M., Vergnaud J. C. (2008). Representation and Aggregation of Preferences under Uncertainty. In *Journal of Economic Theory*, vol. 141, iss. 1, July 2008, pp. 68-99.
23. Gilbert, N.: *Agent-Based Models*, SAGE Publications (2008)
24. Giovanni, D., Fagiolo, G., and Roventin, A. 2010. Schumpeter Meeting Keynes: A Policy-Friendly Model of Endogenous Growth and Business Cycles. *Journal of Economic Dynamics and Control* Volume. 34(9), pp. 1748–1767.
25. Guo J., C. Bhat, (2007) Population synthesis for microsimulating travel behaviour, *Transportation research record*
26. Haug Z., P. Williamson (2001) A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of the small-area microdata, Working Paper 2001/2, University of Liverpool
27. Hongming Xi L., M. Yong D. (2015) An evidential opinion dynamics model based on heterogeneous social influential power. *Chaos, Solitons & Fractals* 73, 98–107
28. Kamiński B., (2015) Interval metamodels for the analysis of simulation Input–Output relations, *Simulation Modeling Practice and Theory*, 54, s. 86-100
29. Kamiński, B. (2012). Multi-agent approach for market modelling. Methods and applications (in Polish: Podejście wieloagentowe do modelowania rynków. Metody i zastosowania) SGH Warsaw School of Economics Press.
30. Kirman A.P. (1992). Whom or What Does the Representative Individual Represent?. In *The Journal of Economic Perspectives Publication*, 6(2), pp.117-136.
31. Kirman, A.P. 1997. The Economy as an Interactive System. In: W.B. Arthur, S.N. Durlauf and D. Lane, ed., *The Economy as an Evolving Complex System II*, Santa Fe Institute, Santa Fe and Reading, MA, Addison-Wesley.
32. Kleijnen J. P., R. G. Sargent, A methodology fitting and validating metamodels in simulation, *European Journal of Operational Research* 120 (1) (2000) 14-29
33. Krause U., (2000), A Discrete Nonlinear and Nonautonomous Model of Consensus Formation, in *Communications in Difference Equations*, S. Elaydi, G. Ladas, J. Popena, and J. Rakowski (eds.) Gordon and Breach, Amsterdam, 2000.
34. Law, A.: (2006) *Simulation Modeling and Analysis*, McGraw-Hill
35. Leijonhufvud, A. 2006. Agent-based macro. In: L. Tesfatsion and K. Judd, ed., *Handbook of Computational Economics*, volume 2 of *Handbooks in Economics*, North Holland, Amsterdam, pp. 1625-1637.
36. Lengnick. 2013. Agent-based macroeconomics: A baseline model. *Journal of Economic Behavior & Organization*. 86, pp. 102–120.
37. Lenormand M., G. Deffuant (2013). Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods, *Journal of Artificial Societies and Social Simulation*
38. Leskovec, J. & Faloutsos (2006) C. Sampling from Large Graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 631–636
39. Lorenz J., (2005) A Stabilization Theorem for Dynamics of Continuous Opinions, *Physica A*, vol. 355, pp. 217-223
40. Lovasz L.,(1993) *Random Walk on Graphs: A Survey*, Bolyai Society Mathematical Studies
41. Miller, J.H., Page, S.E.: *Complex Adaptive Systems*, Princeton University Press (2007)
42. Newman M., A-L Barabasi, and D. Watts (Eds.). *The Structure and Dynamics of Networks*. Princeton University Press. 2006.

43. Nooy W. A. Mrvar, and V. Batagelj. Exploratory Social Network Analysis with Pajek. Cambridge University Press. 2005.
44. Oechslein, C., Klügl, F., Herrler, R., Puppe, F.: (2002) UML for Behavior-Oriented Multi-agent Simulations, Lecture Notes in Computer Science, vol. 2296/2002 pp. 742-743
45. Oeffner Marc (2009) Agent - Based Keynesian Macroeconomics - An Evolutionary Model Embedded in an Agent-Based Computer Simulation. MPRA Paper No. 18199, posted 31. October 2009
46. Pyka, A. and Fagiolo G. (2005). Agent-based modelling: A methodology for Neo-Schumpeterian economics. University of Augsburg, Discussion Paper Series No. 272.
47. Ribeiro D., B. Towsley (2010) Estimating and Sampling Graphs With Multidimensional Random Walks, IMC'10 Melbourne
48. Santos I. R., P. R. Santos, Simulation metamodels for modeling output distribution parameters, in: S. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. Tew, R. Barton (Eds.), Proceedings of the 2007 Winter Simulation Conference, IEEE, 2007, pp. 910-918
49. Shang Y., (2014), Consensus Formation of Two-Level Opinion Dynamics, Acta Mathematica Scientia, 1029-1040
50. Tesfatsion L. (2002) Agent-Based Computational Economics: Growing Economies From the Bottom Up. In Artificial Life, vol. 8, no. 1, MIT Press Journals, pp. 55-82.
51. Wang G. G., S. Shan, Review of metamodeling techniques in support of engineering design optimization, Journal of Mechanical Design 129 (4) (2006), 370-380
52. Wasserman S., K. Faust (1994) Social Network Analysis: Methods and Applications. Cambridge University Press. 1994.
53. Wasserman S., P. Pattison (1996), Logit Models and Logistic Regressions for Social Networks: I An Introduction to Markov Graphs and p^* , Psychometrika 61, 401-425
54. Windrum, P. (2005). Heterogeneous preferences and new innovation cycles in mature industries: the camera industry 1955-1974. Industrial and Corporate Change, 14(6), pp. 1043-1074.
55. Windrum, P., Fagiolo, G., and Moneta A. (2007). Empirical Validation of Agent-Based Models: Alternatives and Prospects. Journal of Artificial Societies and Social Simulation 10(2), 8.
56. Zhou X., B.Chen, L. Liu, L. Ma and X. Qju, (2015), An Opinion Interactive Model Based on Individual Persuasiveness, Computational Intelligence and Neuroscience