D4.6 – Third Status Report of the Pilots

Grant Agreement	676547
Project Acronym	CoeGSS
Project Title	Centre of Excellence for Global Systems Science
Topic	EINFRA-5-2015
Project website	http://www.coegss-project.eu
Start Date of project	October 1, 2015
Duration	36 months
Deliverable due date	30.09.2018
Actual date of submission	16.10.2018
Dissemination level	Public
Nature	Report
Version	2 (after internal review)
Work Package	4
Lead beneficiary	GCF
Responsible scientist/administrator	Sarah Wolf
Contributor(s)	Margaret Edwards, Steffen Fürst, Andreas Geiges, Jette von Postel, Michele Tizzoni, Enrico Ubaldi
Internal reviewers	Ralf Schneider, Patrik Jansson
Keywords	Health Habits, Green Growth, Global Urbanisation, Synthetic Information System
Total number of pages:	55

The CoeGSS ("Centre of Excellence for Global Systems Science") project is funded by the European Union. For more information on the project please see the website http:// http://coegss-project.eu/

Version History

	Name	**Partner**	**Date**
From	Sarah Wolf	GCF	
First Version	for internal review		9/2018
Second Version	for second review		10/2018
Third Version	for submission		10/2018
Reviewed by	Patrik Jansson Ralf Schneider	Chalmers HLRS	9/2018 10/2018
Approved by	ECM		10/2018

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABM	Agent-Based Model
BAU	Business as usual
CO	Combinatorial Optimization
CoeGSS	Centre of Excellence for Global System Science
COVISE	Collaborative Visualization and Simulation Environment
CPU	Central Processing Unit
EA	Evolutionary Algorithm
EV	Electric Vehicle
GB	Gigabyte
GSS	Global Systems Science
HD	High Definition
HPC	High Performance Computing
IPF	Iterative Proportional Fitting
KBA	Kraftfahrtbundesamt (German Federal Office for Motor Traffic)
MOEA	Multiobjective Evolutionary Algorithm
MoTMo	Mobility Transition Model
SIR	Susceptible-Infected-Recovered
SIS	Synthetic Information System
SME	Small and Medium-sized Enterprises
SP	Synthetic Population
SR	Synthetic Reconstruction
TB	Terabyte
WP	Workpackage

Abstract

This deliverable presents the status of the three pilot studies of the Centre of Excellence for Global Systems Science – Health Habits, Green Growth, and Global Urbanization. The pilots finalised HPC-based synthetic information systems for a policy related question each in their respective fields: smoking habits and tobacco epidemics (Health Habits), the evolution of the global car fleet and its emissions (Green Growth), and the two-way relation between transport infrastructure decisions and price mechanisms, particularly concerning real-estate (Global Urbanisation). Further, the scoping task on Future Applications worked on identifying needs and opportunities for future HPC applications in view of global challenges. Progress made throughout the third and last project year is presented together with a look back on the whole project to present challenges encountered in this work and lessons learned from it.

1 Introduction

This document presents the progress of the project work of workpackage 4 (WP4) of the Centre of Excellence for Global Systems Science (CoeGSS) in the third and last project year. It reports on the development of the three pilot studies T4.1 Health Habits (Section 2), T4.2 Green Growth (Section 3) and T4.3 Global Urbanisation (Section 4). Each pilot addresses an example global challenge: smoking as an “epidemic” in terms of health habits, the diffusion of electric vehicles in the global car fleet in view of a sustainable mobility transition, respectively, the two-way relationship between transport infrastructure and real estate pricing as an important element in global urbanisation.

Digital decision support for addressing these (and similar global) challenges needs to take into account that

- the underlying global systems are complex: the system behaviour emerges from interactions of many heterogeneous actors that can influence each other in complex network structures. Moreover, they interact in a common but spatially differentiated environment that they can also influence and be influenced by;
- knowledge and values play equally important roles as the challenges are characterised by uncertainty (the causal structures underlying the complex system behaviour are not well established) and ambiguity (system inputs and outcomes are valued differently by different stakeholders), in other words, anticipated consequences of certain decisions or actions may be controversial, and assessments of these consequences may be contested (Renn & Schweizer, 2009)

While the first point requires sophisticated modelling for a better understanding of the system behaviour, the second point implies that participatory processes linking model scenarios with narratives are also needed (Mielke & Geiges 2018), or, as Dum & Johnson (2017) formulate it: “Policy and societal action is as much about attempts to understand objective facts as it is about the narratives that guide our actions.”

The synthetic information systems (SISs) developed by the pilots (see D4.1, D4.4, and D4.5) and finalised throughout this last project year, are tools for this kind of iterative model-stakeholder interaction; the relevance of large computing power in this endeavour is further discussed in Section 6. This section summarises the experience of three years of work at a previously basically non-existing intersection between the emerging research field of Global Systems Science (GSS) and the world of High Performance Computing (HPC) in form of common lessons learned from a point of view that spans all pilot studies. Before, however, Section 5 reports on the work and results of T4.4, the Future Applications task. Section 6 also plays the role of conclusion. As far as it could be made freely available, the code developed by the pilots can be found at the CoeGSS repository (https://github.com/CoeGSS-Project).

2 Status of the Health Habits pilot

In the third year of the project, the Health Habits pilot focused on two main research lines:

1. developing the code to generate a hierarchical synthetic population based on Eurostat data and a limited list of national surveys (Section 2.1);
2. developing and finalising the code to simulate an agent-based model that integrates the above synthetic population structure with a SIR-like (Susceptible, Infected, Recovered) dynamical model (Section 2.2).

These two components represent the final product of the Health Habits research pilot.

2.1 Synthetic population generation

This section reports a high-level overview of the synthetic population generation tool developed by the Health Habits pilot.

2.1.1 Motivation

In the literature, we can find two kinds of synthetic population generation procedures: Synthetic Reconstruction (SR) techniques and Combinatorial Optimization (CO). The first set of procedures requires a micro-sample (a table reporting a survey on several socio-economic indicators of a small sample of households and individuals in a given area) to calculate the attribute table, usually implementing an Iterative Proportional Fitting (IPF) procedure to generate the joint distributions of the quantities one wants to reproduce in the population.

On the other hand, in the CO approach, which is rarely implemented, one divides an area into many sub areas for which the marginal distributions of the traits to be reproduced are given. Then a sub-sample of the general population is used to fit these marginal distributions and to represent the population of the sub area.

Both of these methods produce populations of individuals grouped in households and organized accordingly to administrative areas (of different granularity levels), reproducing with good accuracy the traits of the population. However, neither of these procedures scale well when the number of traits to reproduce increases and when one wants to generate a population replicating not only the age structure of the agents but also their organization in households. In other words, if one wants to constrain both the agents and the household marginal distributions, the procedures quickly get cumbersome and the amount of data required to carry out the generation procedure suddenly increases. Moreover, all of these procedures require a micro-sample to initialize the contingency table (joint distribution) of the traits to be replicated. These samples are not always available, and when available, it may be for only a small part of the geographical region of which one wants to generate a population.

Work has been done to avoid the requirement of a large and detailed micro-sample. However, the procedure requires data about joint distributions in the country that may not always be available and the procedure cannot then be generalized.

Moreover, in epidemic modelling one is interested not only in the household arrangements of the agents but also in which workplace or school they go to and in which district of a given city they live, so that one can infer the set of people interacting with a specific agent during an ordinary day. Moreover, many ABMs require a hierarchical organization of the population and of the workplaces. This is especially true when dealing with multi-node implementations of ABMs, where many computing nodes simulate the system under investigation. Within this approach it is therefore essential to be able to split the system among the different nodes at an arbitrary level of the hierarchy and to add a custom number of levels to group agents in local clusters, thus refining the most disaggregated level of the census data or administrative boundaries. For example, we may have a region where municipalities are the finest subdivision of the territory. Let us assume that we want to further split agents in smaller groups based on their household location within the municipality. For example, we may want to group them in districts of about 5,000 people, then split in communities of about 800 people which are split in neighborhood groups of ~90 people (about 30 households) in turn. This last step can be used to insert an additional layer of interaction between the agents in the system besides the household and workplace contexts.

To address these problems, our procedure builds on previous works of synthetic population generation and improves them with the following features:

- it provides a general, step-by-step procedure to generate a synthetic population (potentially comprising all of Europe) using only open data coming from Eurostat and from a limited array of national surveys;
- it generates a population with the agents organized in households and workplaces. All these entities are arranged in a hierarchical structure with an arbitrary number of levels and with the possibility to specify the size of each additional local level.
- it arranges the agents' household and workplace locations so as to reproduce realistic commuting data.

Given these features, our generated synthetic populations are suitable to be used as input for epidemics-like ABMs or contact driven models where the daily contact patterns of one agent lead the dynamics of the underlying process under investigation (see Section 2.2).

2.1.2 Data sources

The data used in the synthetic population generation mainly come from the Eurostat database at different geographic definitions. Data is aggregated using the Eurostat NUTS and LAU classification. The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU for the purpose of the collection, development and harmonisation of European regional statistics.

There are 3 major NUTS levels:

- NUTS 1: major socio-economic regions.
- NUTS 2: basic regions for the application of regional policies.

- NUTS 3: small regions for specific diagnoses.

Along with NUTS, Eurostat maintains a system of Local Administrative Units (LAUs) compatible with NUTS. These LAUs are the building blocks of the NUTS, and comprise the municipalities and communes of the European Union[1].

In particular, we retrieve information on:

- fertility and mortality rates by age and NUTS 2 region to reproduce demography;
- participation rates by age group and the education attainment level of the 25-64 age group to establish the education level;
- the employment rate by gender, age and education level at the NUTS2 level to determine the employment status of each agent;
- the population by sex and age in the NUTS3 regions to generate the correct number of agents for each sex and age pair;
- families by type, size and NUTS3 region to get the marginal distribution of households.

While the Eurostat data cover the socio-economic traits and the household structures to be reproduced in the synthetic population, we need to combine these data with additional sources reporting the size of workplaces and schools, the commuting, the spatial density of the population and the administrative boundaries of a country.

These additional resources are:

- the SEDAC adjusted population count, reporting the population count for each cell of about 1 km squared in the whole world;
- the 2012 PISA primary and secondary school size distribution;
- the Open Street Maps (OSM) database for LAU1 and LAU2 boundaries;
- the Italian statistical office ISTAT (commuting and fraction of commuters);
- the ISTAT and UK ONS data on workplace size distribution.

2.1.3 Synthetic population generation pipeline

The code pipeline that generates the synthetic population is structured as follows:

Data pre-processing and import. The pre-processing procedure is aimed at converting the raw tables downloaded from the Eurostat database to Python Pandas dataframes that can later be used in the synthetic population generation procedure. Pre-processing is done through a series of IPython notebooks.

[1] https://ec.europa.eu/eurostat/web/nuts/local-administrative-units

Geo-database. The Health Habits pilot created two collections in a MongoDB database storing the cells of the 2015 SEDAC population count raster[2] and the boundaries of European countries as found in the NUTS scheme[3].

The two collections are organized so as to provide the following features:

- hierarchical structure of the spatial boundaries to replicate the European Commission's NUTS geographical division;
- store census and national health agencies data for each region;
- fast access to the SEDAC raster cells falling within a region to generate the simulation input rasters for the observable under investigation.
- fast identification of the spatial boundary containing a specific cell of the raster.

The cell collection stores the information about the number of people living in a 1x1km area around the world and cells are stored as GeoJSON polygons (rectangles in the latitude-longitude coordinates system) whose 'properties' field reports the population count.

The boundary collection also contains polygons delimiting the NUTS at their different levels. The id value of these entries is set to the NUTS code which naturally provides a hierarchical organization of the documents.

*Database Interface***.** Besides the database, which is intended as a permanent and consistent datastore, the Health Habits pilot also developed a high-level interface to insert (retrieve) data to (from) each region, aggregate them using different schemes, import the simulation output, compare simulation results with empirical time series, and easily visualize results on a map. The interface leverages all the database features and allows for a quick interaction with data.

Algorithm. The first step is to load from disk all the statistics and information about the system. We then proceed generating all the entities that are part of the system, in the following order:

- first, the code generates households by sampling the household size and composition distributions by spatial area units.
- second, the code generates the number of commuters working in each area unit using a gravity model and it assigns workers/students to workplaces and schools by size.
- third, the agents, the households and the workplaces/schools are hierarchically clustered in space using a k-mean clustering method.
- the final output is generated and it is composed by 3 tables in h5 format. The first table contains information of each agent (id, gender, age, household_id, education,

[2]http://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-adjusted-to-2015-unwpp-country-totals
[3]http://ec.europa.eu/eurostat/web/gisco

employment status, income). The second table contains information of each household and workplace (kind, size, location). The third table contains information about the demographic information (year, mortality and natality rates) that can be used to make the population evolve in time.

Figure 1 displays an example of a synthetic population generated for the Piedmont Region (Italy) which is a NUTS 2 region identified by the code ITC1 in the Eurostat nomenclature. A population of 4.3 million agents is divided into households of size *m*, which varies according to the household type. As shown in the Figure, the distributions of household sizes by household kind generated by the model (in orange) closely match the distributions reported by census data (blue).

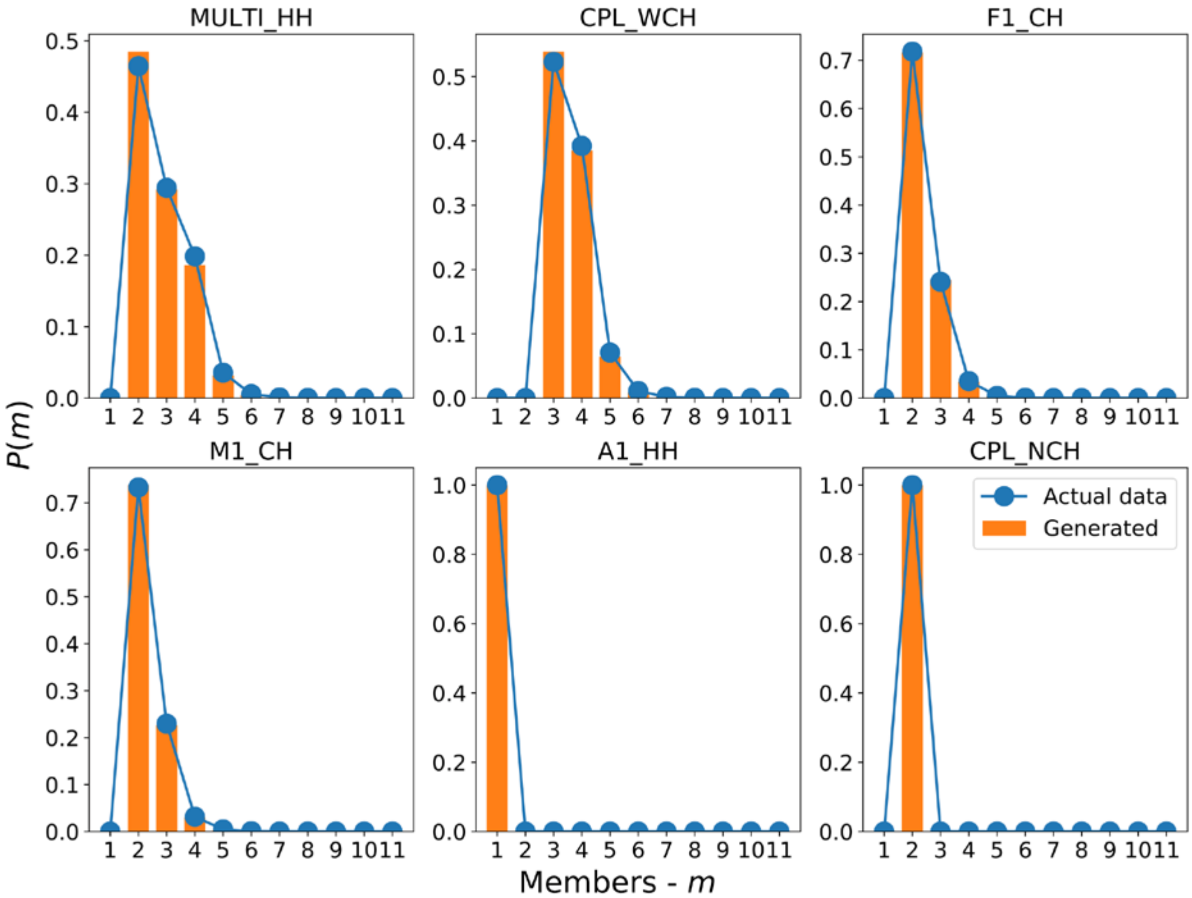


Figure 1. The size distribution P(m) of the actual data and of the generated population for all household kinds in the synthetic population of the Piedmont Region (ITC1, Italy). Household kinds are: couples with children (CPL_WCH), couples without children (CPL_NCH), single adults (A1_HH), single male with children (M1_CH), single female with children (F1_CH), multi-kind household (MULTI_HH).

2.2 Agent-based model

The agent-based model represents the final output of the Health Habit research pilot. It is a tool that can be used by the GSS community to simulate the dynamics of spreading processes on a large scale, over synthetic populations, and using parallel computing, thus exploiting the capabilities of HPC.

The model uses the output of the synthetic population generation (Section 2.1) and it adds a dynamic interaction between agents to simulate an SIR-like epidemic process that is amenable to several applications, such as the spread of diseases or the spread of health related habits.

The dynamics is composed by 4 steps:

- daily step: agents interact in their home, work or travel destination community and in their workplace with co-workers (if they are at work) or at home with other members of the household (if both at home);
- night step: agents interact in their home community with all the agents living in the same community and in the same household with the other household members. People travelling act as if they actually live in the destination community (see next step);
- travel step: for each agent *A*, we let *A* travel with a given probability and we select the length and the kind of travel (i.e., either a leisure or business travel) from a given travel length distribution and with a given business/non-business travel kind. When *A* is travelling, *A* chooses a destination community at random and gets assigned to a household and a workplace (if this is a business trip) in the target community. The agent *A* then acts as living and working in the assigned community until the end of the journey;
- demography step: in the demography step every agent *A* dies with a probability given by the death probability for its age and sex. If the agent dies, then it is removed from the home and work communities, i.e., from its household and, if working, from its workplace. Regarding the natality step, for each female in the system that covers the role of adult (i.e. she is not a child) can give birth to a baby with the probability given in the demography table. The newborn child is then assigned to a kindergarten in either the home or work community of the mother. If no kindergarten is present in these two communities, the child is set to stay at home during the day.

The code then allows to set the parameters of the SIR model under study and run the stochastic simulations for a specified number of runs.

The output consists in an hdf5 file with the following structure:

```
* root
|
|-- * processor
|  |
|  |-- * run_number
|  |  |
|  |  |-- * time_step
|  |  |  | id | sex | age | status | infection_time | infection_source | demo_status | death_cause |
|  |  |  | 0 |  1 | 18 |   2  |      12       |        0        |      0     |     0      |
|
|-- * timeSteps
|  |
|  | stepNumber |  date   |
|  |     00     | YYYMMDD |
```

2.2.1 Results

To give an overview of the results of the agent-based model, Figure 2 shows the results obtained by running the model with a simple SIR epidemic dynamics. The figure highlights the high level of detail for the dynamic process that can be acquired by analyzing the simulation results.

In this specific case, the model is initialized with the synthetic population of the Piedmont Region and 200 stochastic realizations of the model have been simulated. The epidemic is set to begin on January 1st, 2015, with 10 infected individuals. Simulation runs have been analyzed with pySpark.

The total mean incidence by day of simulation is shown in the top left panel. In the top right panel, a sample of 21 runs is shown, to display the stochastisicty of the dynamic process.

The bottom panels display the final size of the epidemic, according to the individual context where the infection has been acquired (household, workplace/school, general community, traveling) and the source of infection (adult or child). Such results reflect the underlying demographic structure that is incorporated into the model through the synthetic population.

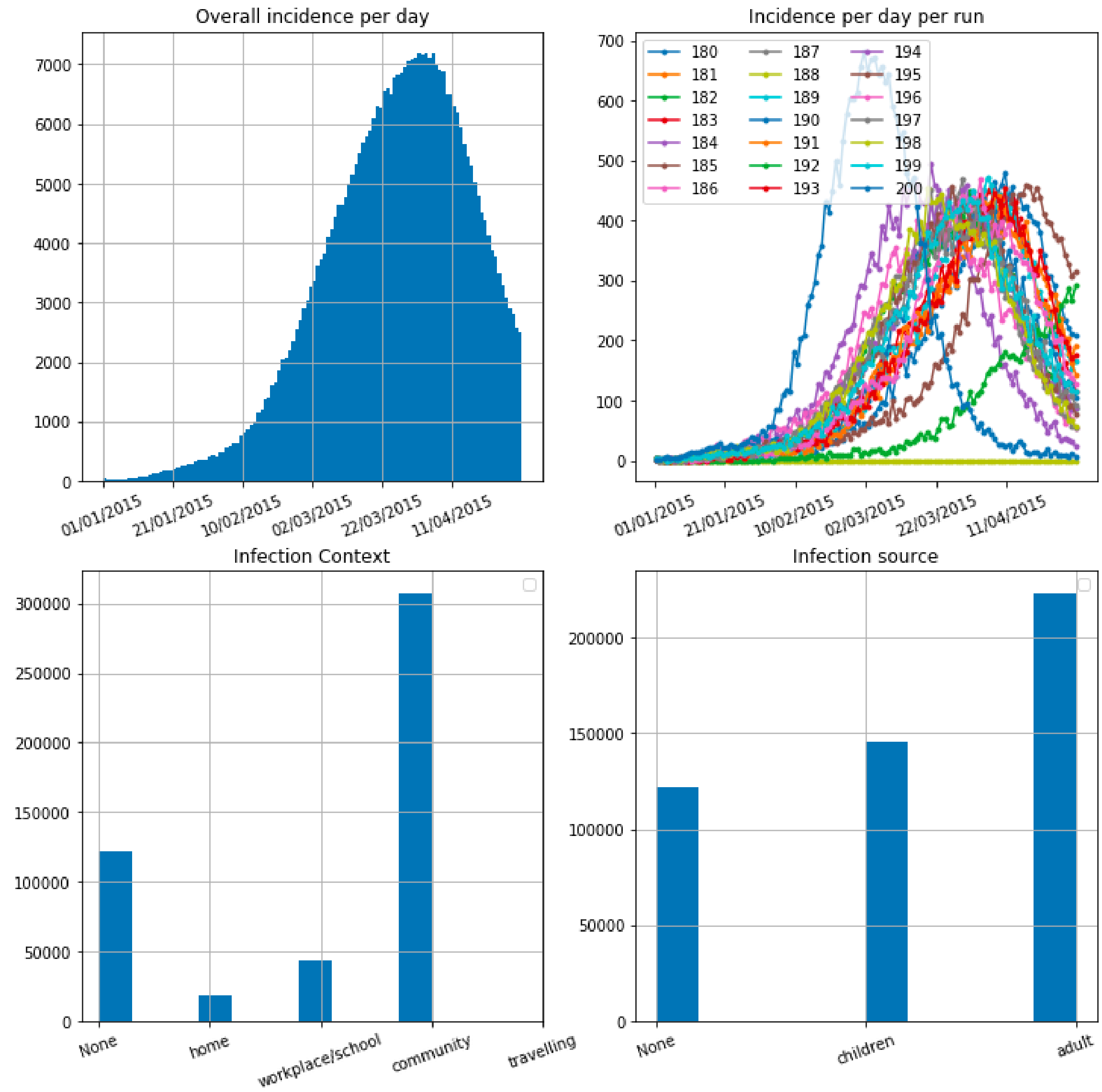


Figure 2. Analysis of 200 realizations of a SIR dynamics with the agent based model. The mean incidence of the epidemic in the population (top left panel). The incidence of 21 realizations of the model (top right panel). The final epidemic size by context of infection (bottom left panel). The final epidemic size by infection source (bottom right panel).

3 Status of the Green Growth pilot

Throughout the third project year, the focus of the Green Growth pilot's work was on the Mobility Transition Model (MoTMo), that has already been described in the previous status deliverable, D4.5. In particular, the model has been extended in several directions (see Section 3.1), calibrated (see Section 3.2), and a hypothetical case of policy scenarios was run (see Section 3.3). Also, the pilot work was presented on several occasions; these are listed in the parallel deliverable D6.7, in the report on dissemination activities.

3.1 Further development of MoTMo

Model extension benefited from synergies with another project, the Energiewende-Navigationssystem (ENavi)[4], that studies the German Energy Transition. Therefore, the previously presented model for the regions of Niedersachsen, Bremen, and Hamburg could be extended to cover the whole of Germany. The model is based on a synthetic population representing age, gender, household size and income distributions of the German population. The generated and fully anonymous agents are located taking into account the different averages in income and household sizes per region.

Additionally, more detail could be added:

- The number of mobility types agents can choose from was increased from three (brown, green, other) to five by further subdividing the category "other" into public transport, car sharing, and non-motorized. Details on these modes are described in (Mielke & Geiges, 2018).
- Agents are equipped with mobility profiles, which describe their number of journeys between 0, 0.5 km, 2.5 km, 10 km, 50 km and more than 50 km in a time-step, but do not give beginning and end locations nor the time needed for single trips. These profiles were developed based on data from a survey by the German Aerospace Center "Mobility in Germany" (DLR, 2008).
- The concept of mobility memes, meaning sets of information containing the mobility decisions of a person, has been introduced. These are further described by Mielke & Geiges (2018).

Since infrastructure is an important factor for mobility, future development of mobility choices needs to take into account the development of the related infrastructure. In particular, the development and the extension of charging infrastructure for EVs is currently very dynamic and likely to play an important role in the current transition phase. While we use data on present road infrastructure, modelling its further evolution would be a very

[4] https://www.kopernikus-projekte.de/enavi

difficult task, since it depends on many local details. Given the question under study, this effort is not warranted here.

Thus, a preliminary model for the development of charging infrastructure was added in the spatial extent of the model. It considers the evolution of the total number of charging stations installed as given; Section 3.3 below shows how scenarios represent policy choices for the deployment of charging stations via different exogenous inputs. Then, the spatial distribution is modelled accounting for three factors: 1) the demand related to electric cars in the surrounding cells, 2) the road infrastructure measured in km of road per cell, and 3) the number of charging stations already present in a cell. The last factor is based on Hotelling's law: places that already have charging stations attract more of them.

The vector of charging stations $nStat$ for all cells is considered with a power of a factor $f1$ which controls its influence. Similarly, the vector of road km per cell is raised to the power of $f2$, which controls the influence of the existing road infrastructure. To avoid too strong self-enforcing, a damping factor d reduces these two effects in case that the supply of charging stations exceeds reasonable limits. In this scenario, we assume that 10 electric cars require at maximum one station.

$$e_{imi} \sim nStat^{f1}$$

$$e_{infr} \sim roadKm^{f2}$$

$$d = max\left(\frac{demand}{10 \cdot nStat}, 1\right)$$

$$p \sim (e_{imi} + e_{infr}) \cdot d$$

The overall probability p is computed according to the equation above and is used to randomly select locations of new charging stations. This preliminary model thus places new charging stations depending on the amount of previous stations in the area and the existing road infrastructure, but limits the growth by a factor taking into account the demand for charging from electric vehicles in the surroundings.

On the technical side, two main achievements regarding the model code were a switch to Python 3 and an integration with the Dakota tool[5] that allows to automatically set up ensemble simulation runs for calibration.

3.2 Calibration

For the calibration, we identified two independent steps that can be approached using different tools. Before the automated calibration of the model to fit the data on electric driving behaviour till 2017, we introduced an initial calibration on mobility data for 2008.

Since there was no electric mobility observed in 2008, the two steps are independent. The survey "Mobility in Germany" (DLR, 2008) serves as a data source about how mobility was

[5] https://dakota.sandia.gov/

used in 2008 dependent on various social, economic and spatial factors. As described in D4.5, agents have priorities for the dimensions convenience, ecology, money, and innovation, represented by person-specific parameters in the utility function. Analysing the behaviour for different age groups, household types, income classes and for different living states, the generation of these parameters for personal preferences is calibrated to match the survey data. This pre-calibrated preference model is then used as a basis for the next calibration step.

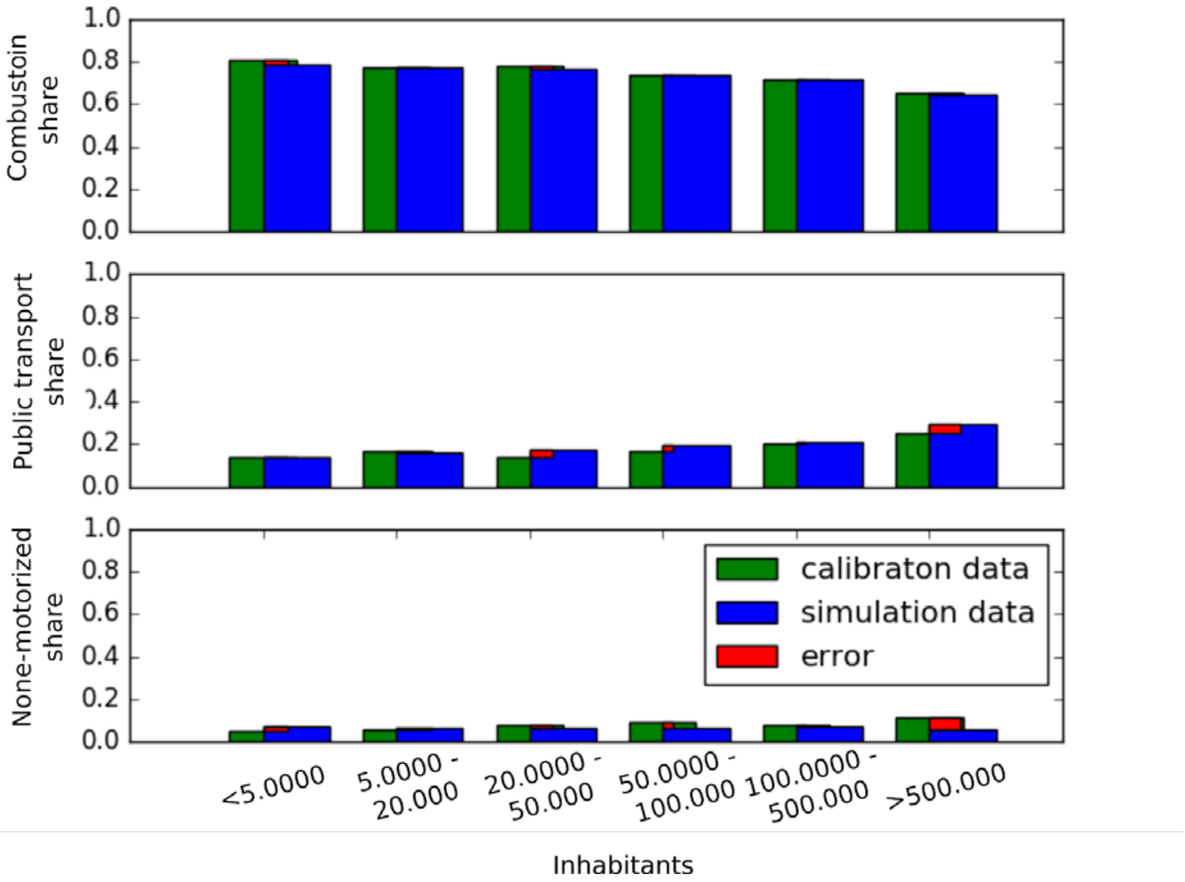


Figure 3 – Intermediate Calibration Results for City Inhabitant Classification

Figure 3 displays that the model covers the mobility change for different city sizes, however needs fine tuning for the behaviour for cities with more than 500.000 inhabitants.

In the second step, in turn, a step-wise procedure was chosen. Model runs, usually extending over the period from 2005-2035, were shortened to the period for which data on the numbers of electric and conventional cars are available from the german Federal Office for Motor Traffic (KBA)[6], that is, from 2009-2017.

While for a best fit calibration, multiple criteria would need to be aggregated into one exact value, which raises the question how to weight the different criteria, as a first step, a

[6] https://www.kba.de/DE/Statistik/Fahrzeuge/Bestand/Umwelt/b_umwelt_z.html?nn=663524

categorical calibration was considered more useful. This means that one defines a range of plausible values for each criterion, and selects parameter sets that match all or most of these intervals. In our case, the car fleet of combustion cars and the car fleet of electric cars for the years 2012-2017 were chosen as criteria (because numbers of electric vehicles were too small before that), the interval to match allows a deviation of 20% from the data. So the procedure aims at matching intervals for all the combinations of car type, year and federal state, that is, $2 \cdot 6 \cdot 16 = 192$ intervals in total.

Dakota supports different multiobjective optimization methods; for our case, we tested two different methods: coliny_direct, which is an implementation of the “Division of rectangles” method (Dakota User’s Manual (2014), page 129) and moga, which is an implementation of a multi objective evolutionary algorithm (Dakota User’s Manual (2014), page 137). coliny_direct was combined with a weighting factor approach for multiobjective reductions, where each interval gets the same weight. The parameters to calibrate where the following:

- *innoPriority* is a selected parameter from the generation of agent preferences and controls the importance of "innovation" as a utility component.
- *mobIncomeShare* controls the average share of the income a household is confident to spend for mobility.
- The *memoryTime* controls for how many months old utility values of former mobility decisions are stored and incorporated in the current decisions process.
- The connection radius (*connRadius*) defines the agent’s interaction radius, within which agents can interact with each other.
- *selfTrust* controls how much the agent trusts his own former experiences in comparison with the experiences of his social network.

These parameters were chosen as examples of different types of model parameters, e.g. innoPriority has a direct and comprehensible influence on the agent’s decisions, while others, like memoryTime, are more artificial.

From the results of test runs for a single federal state with about 600 iterations we concluded that coliny_direct performs better for our calibration needs. These results for coliny_direct are shown in Figure 4, which is a parallel coordinate plot, in which each line represents a single simulation run; the bold lines are those runs that miss only two of the twelve intervals for the federal state, as can be seen in the upper plot that shows the different error measures. These are the six intervals for combustions cars, the six intervals for electric cars, the sum of both and additionally a weighted root square error, where the weight is the inverse sum of the electric respectively combustion cars of the KBA data. Selecting the runs that miss only two intervals (see the little box at the bottom of the interval sum axis, c_intervalSum), one can then view the different parameters used for the calibration, see the lower plot.

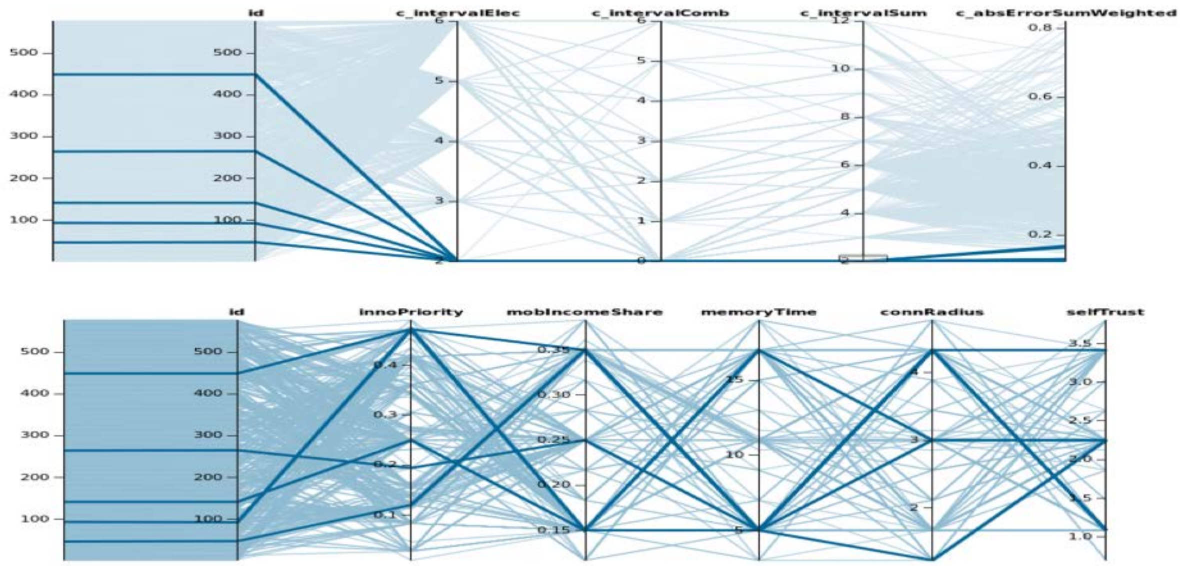


Figure 4 – Parallel Coordinate Plot

In this example, the runs that miss only two intervals do not indicate a best parameter combination. As this may often be the case in GSS modelling, one could then move on to the next step with several parameter sets in parallel, increasing the needs for computation. Here, the next step of calibration, that takes a closer look at error measures as well, was still in progress at the time of writing of this document.

3.3 Interactive Visualisation of Example Scenarios

In decision support, models of global systems help explore potential future evolutions of the system in question, and assess possible consequences of alternative actions. Thus, they need to be able to represent different policy alternatives, or scenarios. The results of (sets of) model runs relating to these scenarios then need to be accessible in an interactive visualisation tool, so that they can be shown as required to support discussions in real time.

The Green Growth pilot considered the example question "What are possible effects of different decisions on the deployment of charging infrastructure on the uptake of electric vehicles in Germany up to 2035?". This choice was motivated by the facts that Germany has declared goals for electric mobility (1 mio EVs on the road by 2020, and 6 mio by 2030 (Deutsche Bundesregierung, 2011)) which by now seem hard to reach, that there is a close connection between charging infrastructure and EV numbers (Hall & Lutsey 2017), and that charging infrastructure can be considered "part of a regional ecosystem" with significant spatial variability both in charging infrastructure density and in EV uptake (Hall & Lutsey 2017). The following illustrates how MoTMo can be used in discussing alternative policy scenarios and thus presents the final output from the Green Growth pilot.

Two example scenarios were represented, depicted in Figure 5: a business as usual (BAU) scenario, in which the given trend of data on charging infrastructure deployment is continued linearly, and a more ambitious "alternative scenario", in which investment is increased rapidly to reach roughly a million charging stations by 2035.

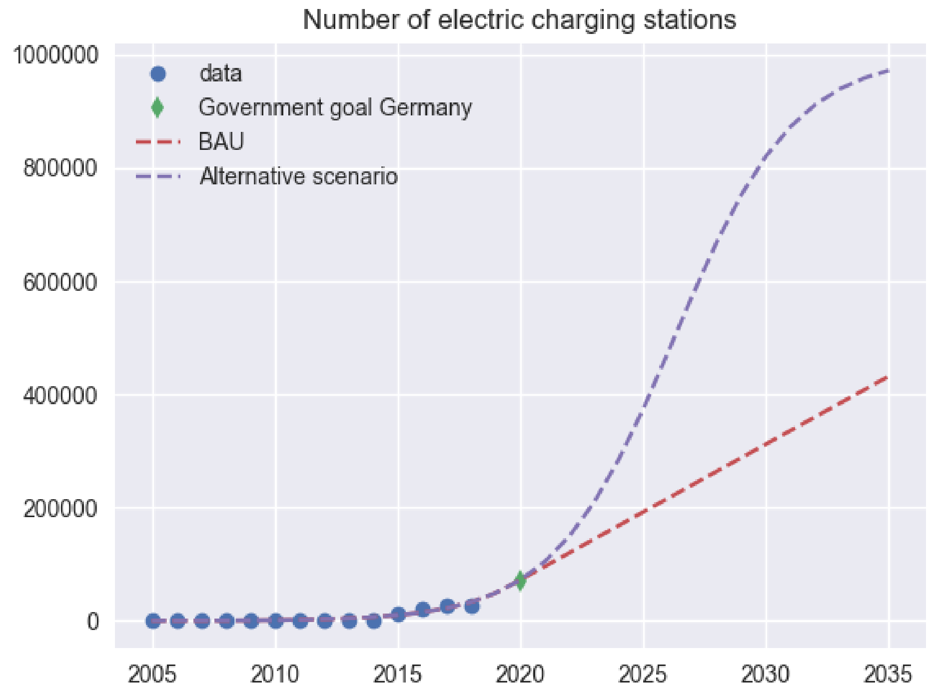


Figure 5 – Scenario Presentation

The preparation of tools and results for interactively showing potential consequences of these two scenarios included several steps. First, MoTMo was run with the respective inputs for both scenarios (preliminarily, 10 simulation runs for each scenario have been computed, requiring about 63 core hours each run, producing 18 GB output data each run). Second, MoTMo output was post-processed to obtain information about aggregate values and variation over the 10 runs per scenario as well as to aggregate output values per agent per time step along different dimensions that are of interest in decision support (e.g., spatial subunits of Germany, different types of households, etc.). So far, the post-processing was carried out via Python scripts without interaction with the visualization tools. Only pre-computed and extracted data is therefore shown as visualized output. Third, the visualisation tool was set up to show a) time-series (aggregates with uncertainty, shares of mobility modes chosen, outputs from several single runs), b) direct comparison of time series for the different scenarios, c) comparisons of maps for spatial distribution of results ("heat maps", 2D) and d) comparisons of 3D maps on a globe with coloured spikes that can display two variables in results at the same time via height and colour. The open source visualisation code is based on R and the shiny package for "interactively telling data stories" (shiny, 2017).

With the thus prepared data and toolchain, MoTMo helps to consider effects of enhanced deployment of charging infrastructure in several dimensions and views. In the following sections, we present several aspects of both the visualisation options and qualitative preliminary results that could be discussed with stakeholders.

3.3.1 Uncertainty

Figure 6 shows mobility developments for the BAU scenario for all of Germany. One can depict many runs within one graph or their mean including the variation per mobility mode. As all model output data pictured in this section is from a dataset prior to the calibration process, the purpose here is rather to show qualitative effects and visualization options than to present accurate numbers. Qualitatively, these overall results for Germany show a tendency that electric vehicles pick up at the cost of public transport. This needs to be taken with caution as public transport is not, as yet, implemented in a sophisticated way, or calibrated against data. In a decision support context, it is nevertheless a point that may be worth discussing: policies for enhanced charging station deployment may need to be complemented with policies strengthening the attractivity of public transport.

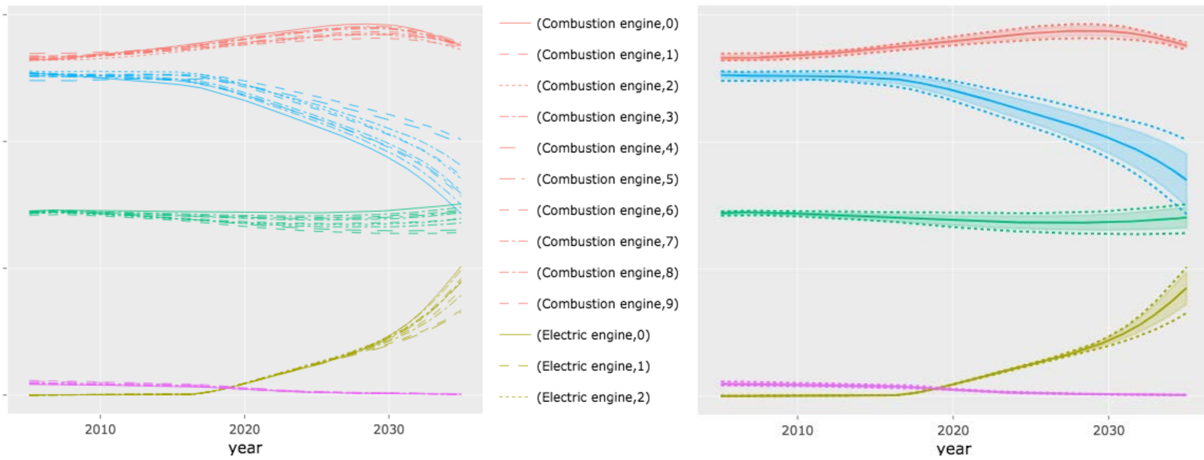


Figure 6 – Mobility Developments for Germany

3.3.2 Different spatial areas and social groups

Figure 7 shows mobility demand for two German federal states, Niedersachsen and Berlin. It illustrates that regional differences can be presented, based on an aggregation of data to the respective level. The differences seen here are intuitive: Berlin is a city state where public transport plays a larger role while Niedersachsen contains large rural areas so that conventional cars are the first means of mobility. As above, the decrease of public transport occurring together with the increase of electric vehicles needs to be checked and/or discussed.

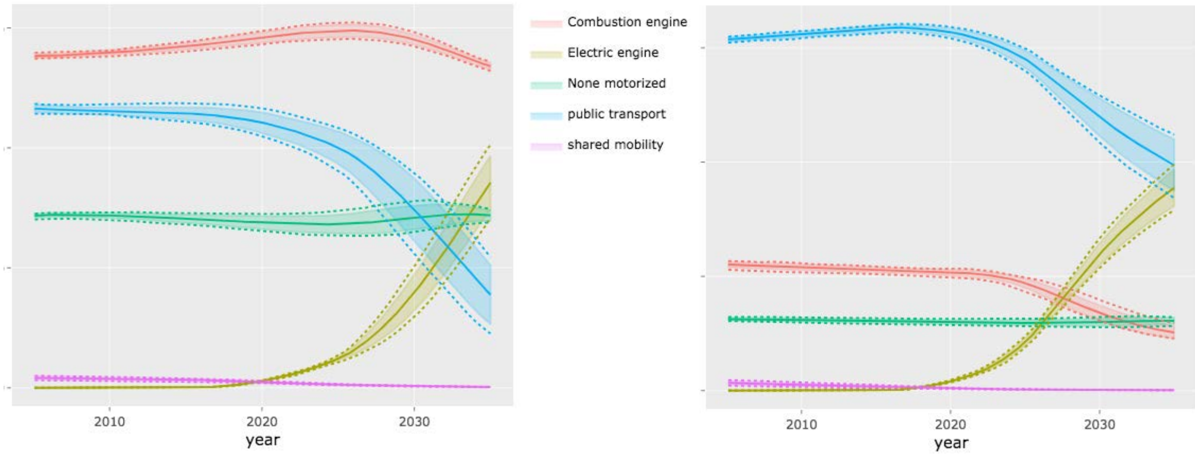


Figure 7 – Mobility Demand for the Federal States Niedersachsen (left) and Berlin (right)

Similarly, Figure 8 shows the mobility demand aggregated for different social groups, e.g. a young single person household (on the left) versus an elderly couple (on the right), so different outcomes across social groups can also be shown very easily. Content-wise, the results point to a larger innovative propensity of younger people. Discussions at this level of granularity become important, when policy / decision makers want to target specific groups of people.

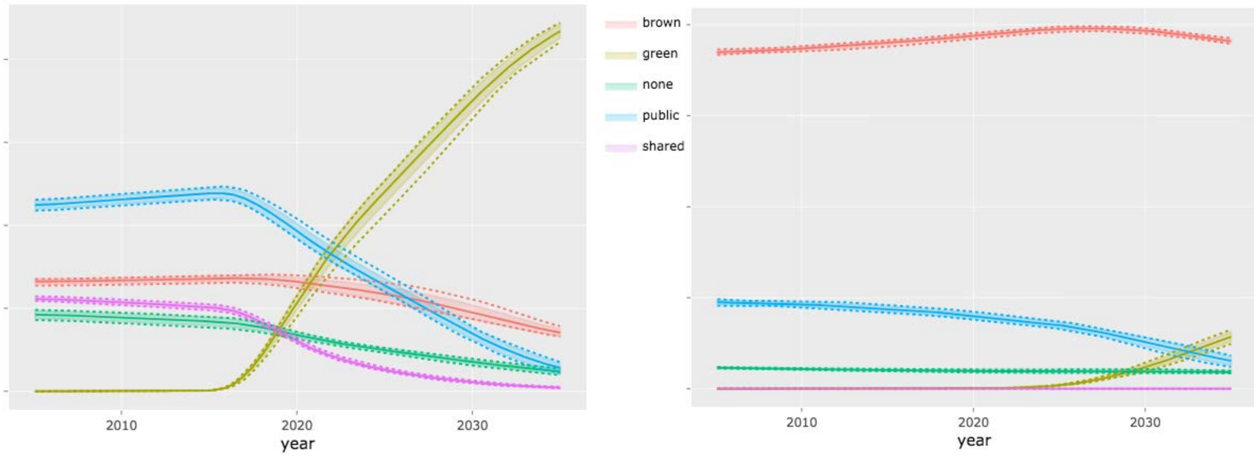


Figure 8 – Mobility Demand for Young Single Person (left) and Old Couple Household (right)

3.3.3 Different development of external factors

Figure 9 presents a visualisation for comparing the outcomes of different assumptions on the development of external factors; in this case, the assumptions concern the two policy scenarios on charging infrastructure deployment described above. The BAU scenario is represented by solid lines, the ambitious scenario by dashed lines. The partly more, partly less diverging lines show how differently the assumptions affect the mobility modes.

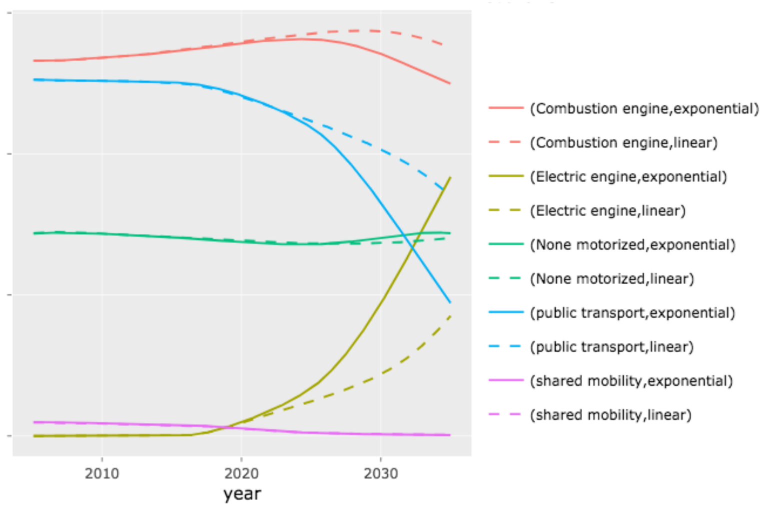


Figure 9 – Mobility Demand for Different Scenarios

Different assumptions can also be included in the model at many other levels, from uncertain future exogenous influences (e.g., different stakeholders may have different views on probable developments of prices for electricity relative to prices for motor fuels) via different assumptions on the effects that certain decisions may have on the modelled agents (in our case, e.g., will digitalization make car sharing more convenient, and how so?) to different representations of decision making by actors in the system. Comparing model output under different assumptions helps to understand dynamics of the system under consideration (Mielke & Geiges, 2018). As the concrete assumptions to be tested and compared will evolve together with discussions about the global challenge at hand, the underlying model should be flexible enough to allow for implementing and testing different assumptions at rather short notice. Here ABM are useful: since mechanisms are implemented at the micro-level, a behavioural assumption can be replaced by another one much more easily than this would be the case for aggregate models.

3.3.4 Spatial representation and different focal points

A focus on one or several pilot regions is often very useful when dealing with alternative decisions in view of global challenges. For any sustainability transition, one aims for a future that is fundamentally different from the given situation and unknown; real-world “labs” for observing effects of certain actions are often useful. It needs to be possible to present model output for decision support relating to regions in the form of maps.

Figure 10 shows two visualisation examples: values for one variable, here the number of EVs, are shown in a 2D map on the lefthand side. The righthand side presents two variables in a 3D map: the height of the spikes depicts the number of charging stations and the colour of the spikes the number of EVs (the brighter, the more). Both maps show data for 2035 from an ambitious scenario run.

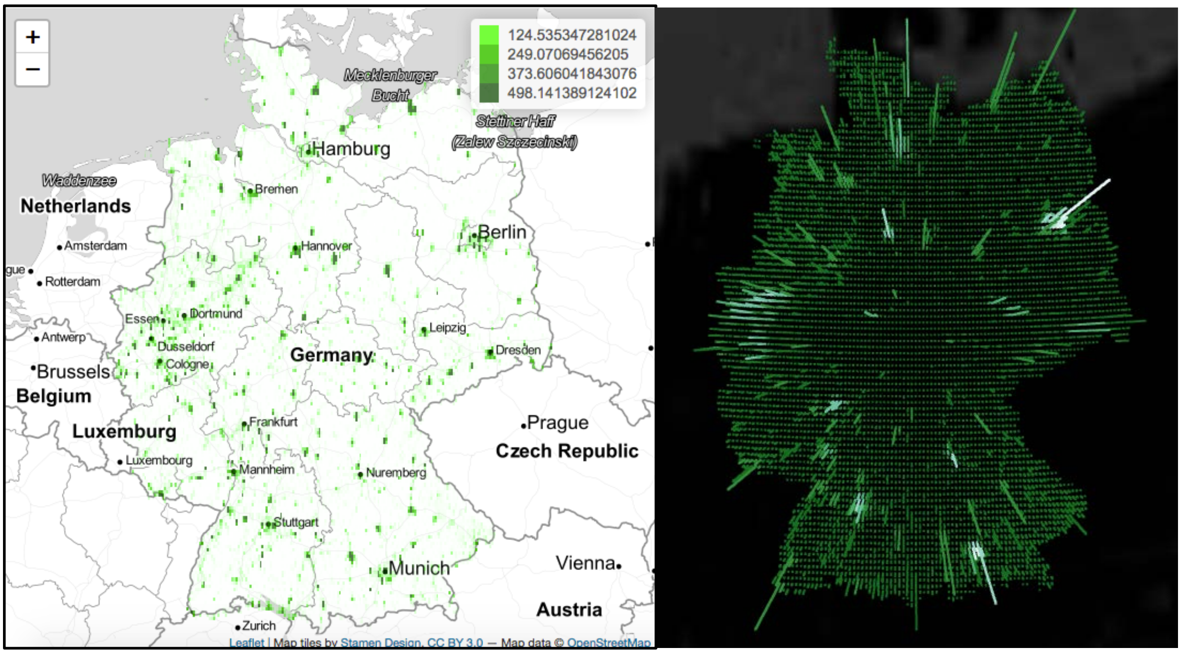


Figure 10 – Visualisation of Various Variables in Maps

Both maps show that electric mobility is likely to develop more strongly in larger cities. The right map shows a few places where charging stations are present but electric vehicles are not. As mentioned, the data used here was preliminary, so, again, this needs to be checked. Should this feature remain in a calibrated model version, it is again a point for further research and discussion: “what are the properties of places where installing charging stations does not have the desired effects?” and similar questions are relevant to decision makers who need to decide on charging station placement.

Further, the visualisation tool allows to zoom in; the left part of Figure 11 presents a zoom into a region (Nordrhein-Westfalen with parts of Niedersachsen), based on the map on the left in Figure 10. On the right, the same region is shown, but with data from a simulation run for the BAU scenario. What one sees was to be expected: the more ambitious scenario for charging infrastructure investment leads to larger numbers of electric vehicles. The spatial distribution can also be compared for these examples.

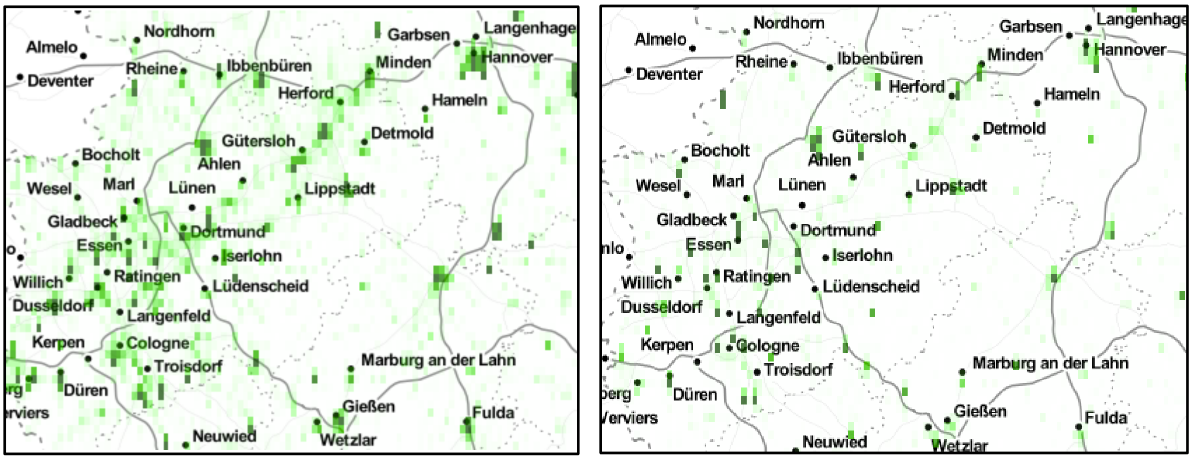


Figure 11 – Zoom on a region to compare scenarios

3.3.5 Variety of model outputs

While the previous pictures have focused mostly on shares in mobility demand of the 5 mobility types represented in MoTMo, the model allows for collecting many further items of information: for example, emissions (total and per mobility type), or electricity demand from EVs. Again, the ABM approach allows to flexibly add and aggregate features or further consequences of what has been modelled based on interactions with stakeholders.

Figure 12 shows the emissions per mobility type, for the BAU and the ambitious scenario in a time series as an example of what the amount of model output offers. Here, one can see that at the beginning, the emissions from “brown” (conventional) cars are higher in the ambitious charging station deployment scenario. Such “surprise” effects show that models of complex systems can be useful to point out possibilities one would not otherwise have thought of. The mechanism at play here might be a feedback between numbers of cars and technical progress that reduces the emissions, even of conventional cars, over time. Apart from the previously mentioned caveat that this particular model version was not yet calibrated, this would again be a point for further research and discussions.

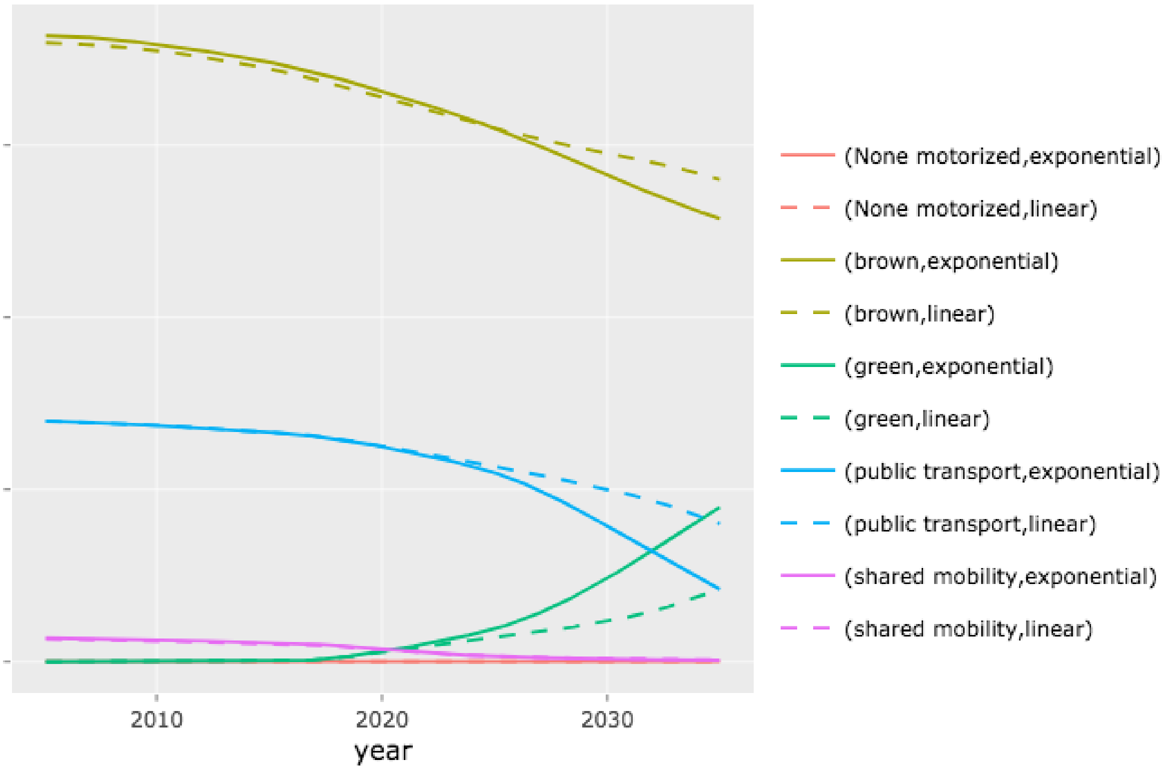


Figure 12 – Emissions per Mobility Type

In conclusion, the Green Growth pilot has obtained a SIS that can be used in initial stakeholder dialogues and further developed according to needs identified, to support an iterative process of digital decision support in view of a global challenge, enabled by large computing power.

4 Status of the Global Urbanisation pilot

Throughout the project, the Global Urbanisation pilot has explored different forms of possible synergies between GSS and HPC, in particular from the point of view of an SME . These appear in the graph hereunder (Figure 13). Indeed, HPC (lower part) allows GSS (higher part) firstly to refine the scale of data and agent (left side), secondly to get a wider overview (right side), along the life cycle of a model, from the real-world observations to the model scenario simulations (left to right).

Over the first two project years, the aspects we studied were: refining the data and agent granularity, getting a wider overview by multiplying the simulations. Refining data is represented by the two maps of Paris real estate prices, per district and interpolated (in Figure 13 on the left side). The influence of agent granularity is shown in the table displaying an example of evolution of real estate pricing for different levels of agent granularity and a key parameter (number of car commuters). Finally, the wider overview is depicted by the picture at the bottom, which shows the level of pollution as following from ecological awareness and adaptability of public transport to demand.

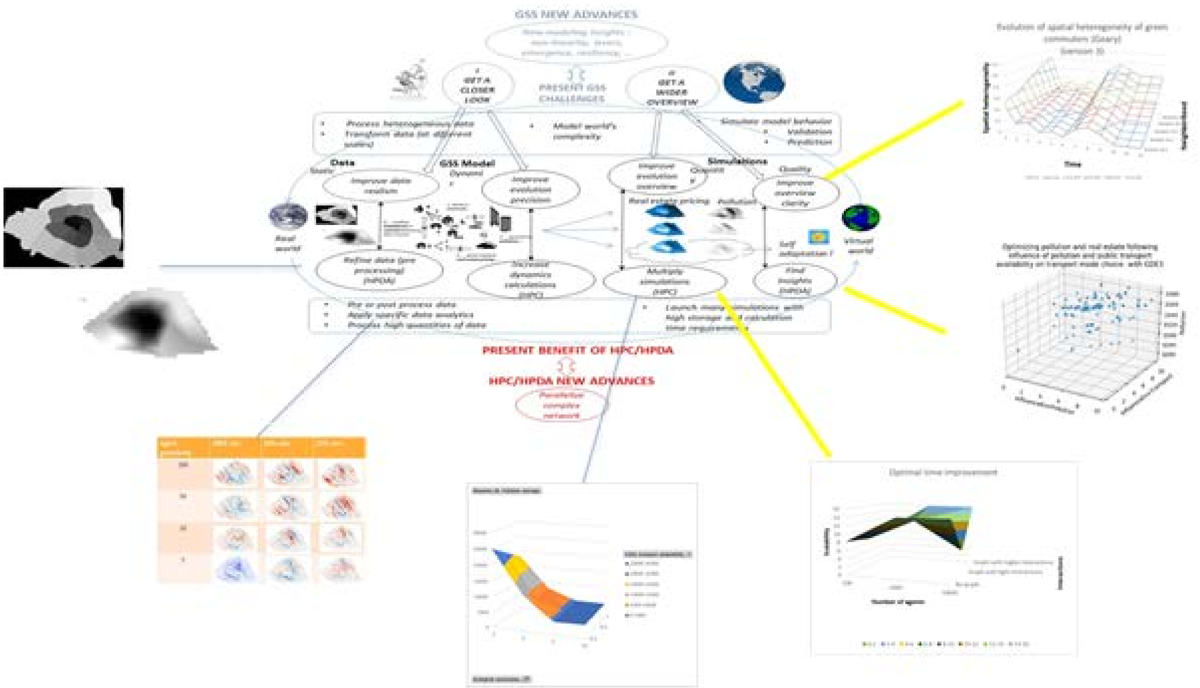


Figure 13 – Overview of HPC-GSS possible synergies investigated by the city pilot

This year, after post-processing further results by calculating the evolution of their spatial heterogenenity, we have focused on the parallelization of one and of a set of simulations (highlighted by yellow lines in Figure 13). The post-processing study is represented on the top right in the graph showing the evolution over time of the heterogeneity of the percentage of green commuters, calculated for different sizes of neighbourhood (depth). The parallelization of one simulation (bottom yellow line) is depicted by a graph showing the non-trivial speedup for different kinds of interactions between agents. The parallelization of a set of simulations (middle yellow line), called by a set of optimization algorithms, is displayed by a cloud of objective values. These are calculated by an optimization algorithm following the value of two parameters.

In the first section (4.1), we summarize the results of a post-processing study concerning the evolution of heterogeneity and final state of various indicators (pollution, real estate, green commuters and public transport offer). This shows how computing resources can help refine insights on models, while keeping high level clarity.

In the second section (4.2), we study the influence of different parameters on the speedup of a parallelized CoSMo model. After putting into light more sensitive parameters, it shows how speedup can improve with the challenge difficulty (calculation complexity, number of agents, intensity of their interactions).

In the third section (4.3), we compare the performance of parallelized multiobjective evolutionary algorithms over optimization problems of increasing difficulty. We put into light how the optimal algorithm varies with the targeted area of the parameter space and the specific problem. This highlights how valuable the parallelization of these algorithms is for models with GSS level of complexity, by allowing to find a better solution than if running a single algorithm.

4.1 Data post-processing

In this section, we calculate the spatial heterogeneity of various observables (pollution, real estate, percentage of „green“ commuters (preferring public transport over their car), public transport offer). Spatial heterogeneity allows to observe diffusion processes with more precision while opening to high level questions stakeholders might ask (such as equity).

Our study aims to show how this indicator refines aggregate indicators by giving not only an average value but characterizing it spatially. It appears discriminating, and therefore pertinent, since the values calculated vary with the observables and the model versions. It therefore provides new insights on the evolution and final values of the various observables (pollution, real estate, green commuters), and on the similarities and differences between the different versions of the model. These two different versions of the model (version 2 and 3) are detailed in the section 5.4.1 of deliverable D4.5. For version 2, we tested two values of a key parameter defining the decision to choose public transport: either this decision depends on the environment and observed pollution level (“all environment”) or it depends on the public transport offer (“all transport”).

We first (4.1.1) compare final spatial heterogeneity of these indicators for different values of a key parameter (the citizens‘ ecological awareness) and different versions of the model.

We then (4.1.2) observe, again for these indicators, the evolution of heterogeneity over time.

4.1.1 Calculating final spatial heterogeneity

We calculated the spatial heterogeneity of various indicators following Moran's and Geary's C indicator. In the table below, we show the results for Geary's C indicator.

$$C = \frac{(N-1)\sum_i \sum_j w_{ij}(X_i - X_j)^2}{2W\sum_i (X_i - \bar{X}_j)^2}$$

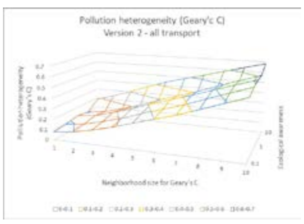
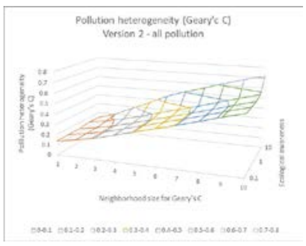
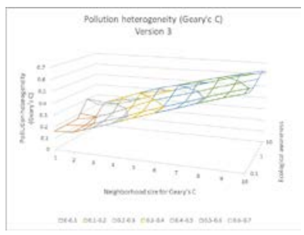
where X_i are the values, w_{ij} is a weight defining whether *i* and *j* are neighbouring cells (and depends on the size of the neighbourhood chosen), *W* is the sum of all weights, *N* is the total number of values.

This study (see Table 1) shows differences of heterogeneity for various observables (pollution, real estate and green commuters) and various versions of the model. Every graph shows the heterogeneity of the indicator depending on the value of a key parameter (the ecological awareness) (depth) and the size of the neighborhood taken to calculate Geary's C (horizontal).

Expectedly, the observed heterogeneity increases with the size of the neighbourhood considered (left to right), but more or less quickly depending on the indicators.

The heterogeneity varies for the different indicators. Pollution and real estate provide the smoothest curves, due to only small local variations. At the opposite the green commuters show more elaborate patterns of spatial heterogeneity. This heterogeneity appears to vary more sharply with the ecological awareness. Indeed, depending on the levels of ecological awareness, green behaviours spread more or less widely among commuters, leading to varied levels of heterogeneity.

Finally, the table also highlights the differences between the model versions in terms of heterogeneity: increasing the version and therefore the feedback and interdependence of model elements (particularly pollution, public transport availability, real estate), can lead to more complex sensitivity of the heterogeneity to the ecological awareness defining the readiness of citizens to prefer public transport.

Observable	Examples (Geary and Moran) of heterogeneity visualization following different neighbourhoods (x) and different levels of ecological awareness (y, depth)		
	Version 2 – all transport	Version 2 – all environment	Version 3
Pollution 			

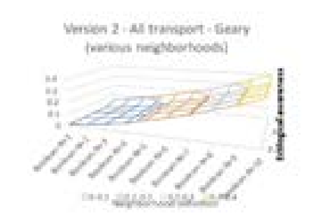
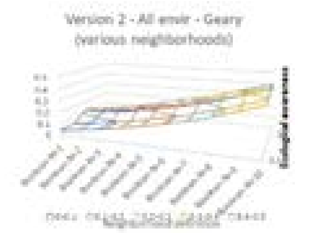
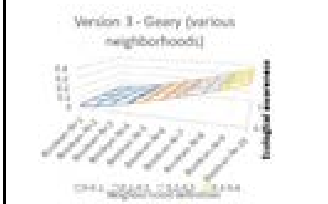
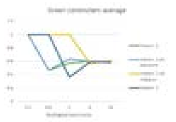
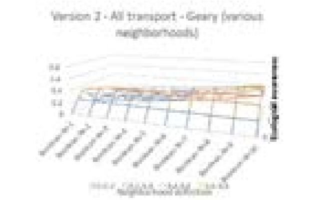
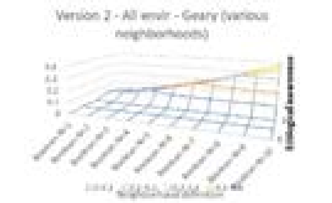
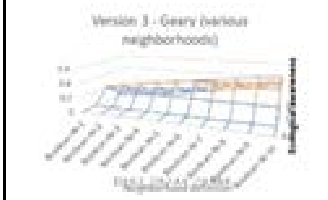
Real estate 			
Green commuters 			

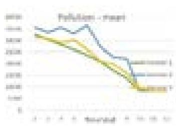
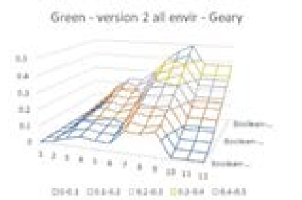
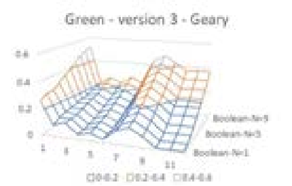
Table 1 – Examples (Geary and Moran) of heterogeneity visualization for pollution levels, real estate pricing and percentage of green commuters, following different neighbourhoods (x) and different levels of ecological awareness (y, depth)

4.1.2 Comparing the evolution of heterogeneity over time for different versions of the model

Here (Table 2) we calculate the heterogeneity over time (horizontal) depending on various sizes of neighborhood (depth).

The heterogeneity varies here too for the different indicators. Again, the real estate and the pollution show the least variation of heterogeneity, here over time. The heterogeneity of green commuters varies more sharply over time, increasing in the diffusion process before decreasing again when they become the majority.

This highlights more clearly the differences between the models than when observing only the final heterogeneity or other aggregate indicators.

Observable	Examples (Geary) of heterogeneity evolution visualization following time (x) and different neighborhoods (y, depth)		
	Version 2 – all transport	Version 2 – all environment	Version 3
Pollution 			

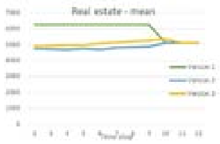
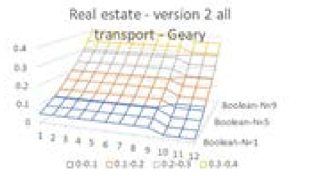
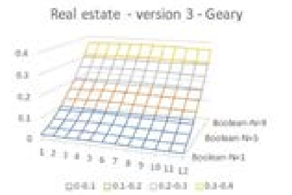
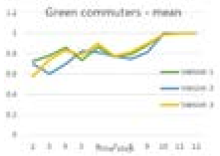
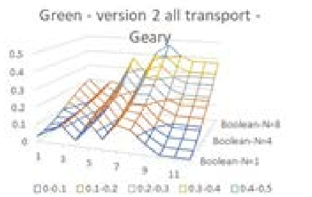
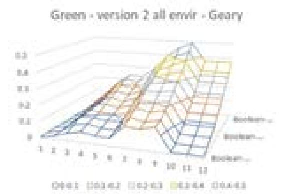
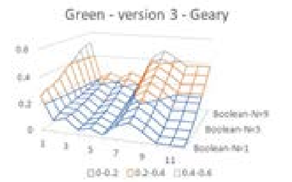
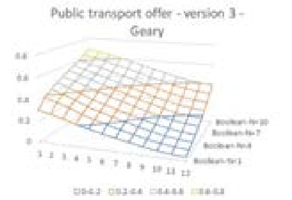
Real estate 			
Green commuters 			
Public transport offer 			

Table 2 – Examples (Geary) of heterogeneity evolution visualization for pollution levels, real estate pricing and percentage of green commuters, following time (x) and different neighbourhoods (y, depth)

4.2 Studying the parallelization of one simulation

4.2.1 Purpose

The purpose of this study was firstly to study how well a parallel CoSMo model can speed up and secondly, how increasing the calculation challenge raised by various features can influence this speedup. We define the speedup by the ratio between the run time with different numbers of processing units. If the first question is quite specific, the second concerns more generally GSS models.

4.2.2 Overview

Our study shows that CoSMo models can speed up quite well, with the speedup varying with the specific calculation challenges tackled. Indeed, speedup appears mostly best when calculations raise a real challenge. We have tested four different kinds of features: the unit calculation difficulty and time required, the number of time loop iterations, the number of agents and the level of interconnectivity of their interactions (see Figure 14).

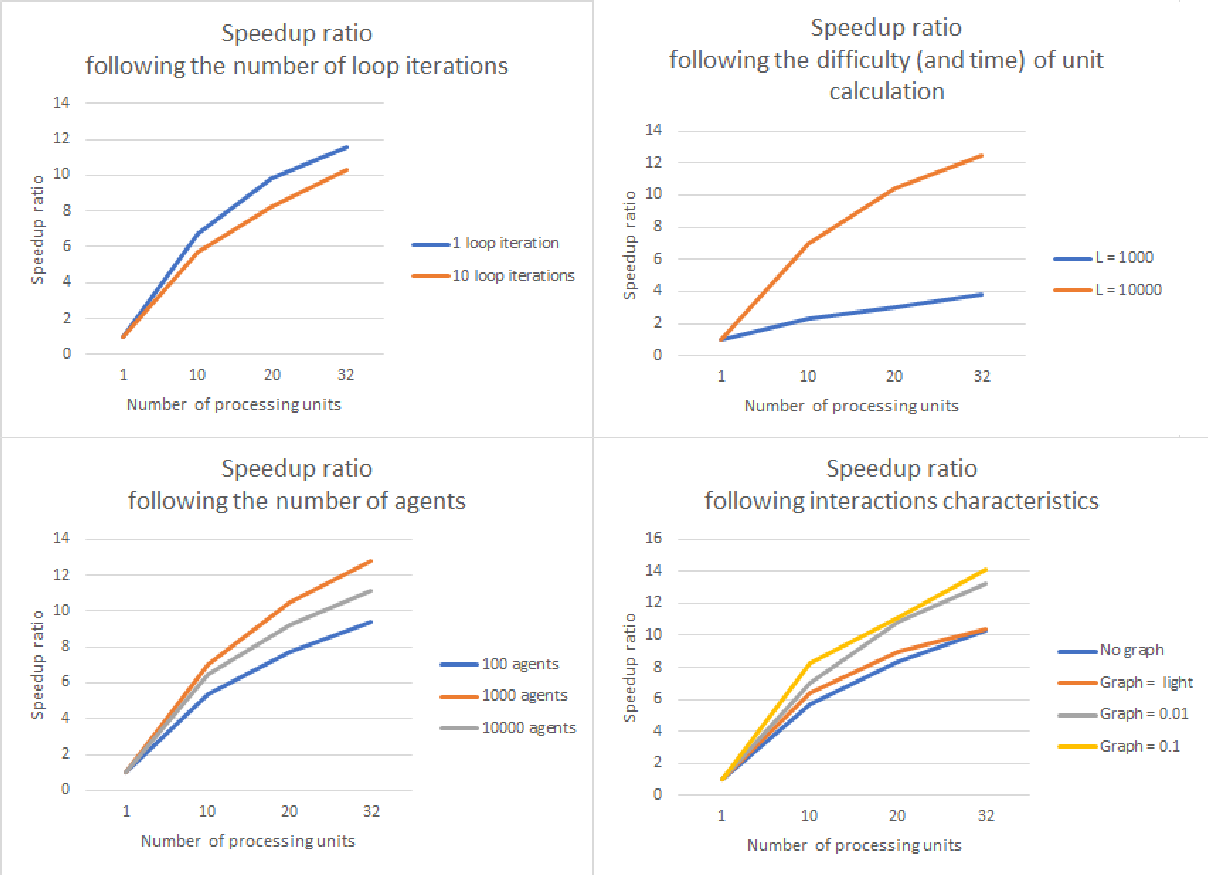


Figure 14 – Influence of various features on the speedup of a CoSMo model: number of loop iterations, complexity and time of unit (per agent) calculation, number of agents, agent graph interconnectivity (no graph, graph with light interactions (calling for light calculation), graph with interactions calling for significant calculations, and interconnectivity of 1% and 10% (every agent connected to respectively 1% and 10% of other agents)

4.2.3 Influence of unit calculation time and number of time loops

Expectedly, the benefit of parallelization appears best when the unit calculation increases in difficulty and time required (being multiplied by 3), minimizing the share of time dedicated to dispatch of calculation and retrieving results. At the opposite, the number of time loops has little influence on the speedup.

4.2.4 Influence of number of agents

The number of agents and the graph interconnectivity play a more complex role. It is nice to observe that speedup (and benefit of parallelization) appears to increase often with the difficulty, even when agent interconnectivity densifies.

The speedup first increases with the number of agents from 100 to 1000 but then doesn't change much when increasing a further factor of 10. We can firstly hypothesize that too low a number of agents to be parallelized leads to an overhead which limits the speedup; and secondly, that over a given threshold and reached level of speedup, the overhead becomes difficult to reduce.

4.2.5 Influence of the graph interconnectivity

As concerns the level of graph interconnectivity, we tested two characteristics. Firstly, different levels of calculations triggered by interacting entities: none, light, or significant (corresponding to the labels graph = 0.01, and graph=0.1). Secondly, for higher calculation time linked to interactions, we tested two levels of interconnectivity: every agent interacting with 1% (=0.01) or 10% (=0.1) of the other agents. Here again the speed up (and consequently the benefit of parallelization) appears to improve with the difficulty.

Let's finally look at the best speedup following both the number of agents and their level of interactions (see Figure 15). Here, as previously, we consider different kinds of interactions triggering different levels of calculations: none, light, or significant (with 1% of the other agents).

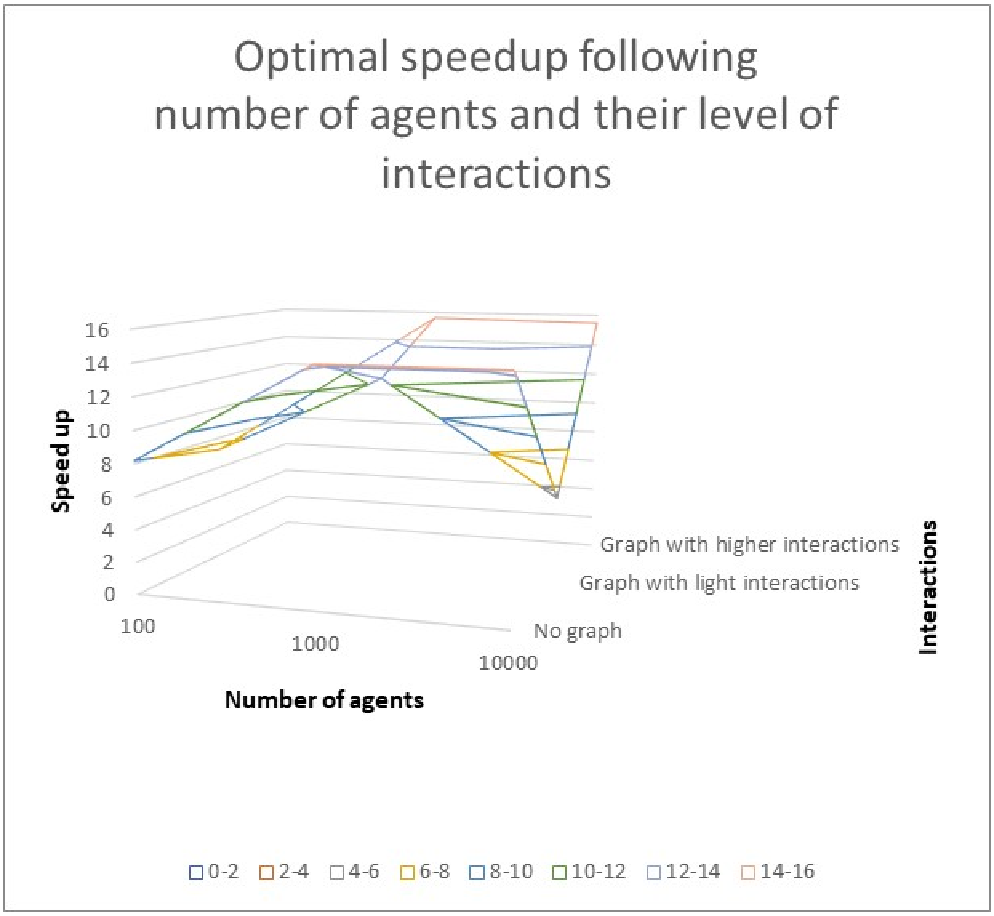


Figure 15 – Optimal speedup (Time for minimal number of processing units / Time for maximal number of processing units) following the number of agents and their level of interactions.

With no interactions, speedup increases nicely with the number of agents.

With light interactions, speedup increases before decreasing again with the number of agents. We can hypothesize that here handling the interactions adds to the overhead, without the calculation cost of a single interaction being high enough to observe a global benefit from parallelization.

Finally, with consequent interactions the speedup increases again nicely with the number of agents, performing even better than with light interactions, and comparably to the case with no interactions.

To summarize this study shows how speedup seems to often improve with the difficulty (increasing the calculation time, the number of agents or the intensity of their interconnections).

4.3 Studying the parallelization of a set of simulations (over multiobjective evolutionary algorithms)

Finding a potentially multi-objective optimum proves challenging, especially when the optimum is to find over the multi-dimensional response surface of a complex GSS model.

Classical approaches assuming a certain level of regularity of these response surfaces might therefore fall short. We consequently chose to explore the possible benefit of evolutionary algorithms. Quoting their Wikipedia page[7]: "In artificial intelligence, an evolutionary algorithm (EA) is a subset of evolutionary computation, a generic population-based metaheuristic optimization algorithm. An EA uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions (see also loss function). Evolution of the population then takes place after the repeated application of the above operators. Evolutionary algorithms often perform well approximating solutions to all types of problems because they ideally do not make any assumption about the underlying fitness landscape."

More specifically, we chose to explore the possible benefit of multiobjective evolutionary algorithms (MOEAs), which tackle the optimization of more than one objective.

To do so, we used more specifically a python library, Platypus, completed with its Experimenter module, which allows for parallelization.

Quoting the documentation of Platypus[8]:

"Platypus is a framework for evolutionary computing in Python with a focus on multiobjective evolutionary algorithms (MOEAs). It differs from existing optimization libraries, including

[7] https://en.wikipedia.org/wiki/Evolutionary_algorithm
[8] https://platypus.readthedocs.io/en/latest/

PyGMO, Inspyred, DEAP, and Scipy, by providing optimization algorithms and analysis tools for multiobjective optimization."

Over this we have used the experimenter module, which allows to run and test in a parallel way different evolutionary algorithms on a given problem. Indeed, the most appropriate ones can vary with the problem (number of variables and objectives), but also with the specific response surface of the model and the characteristics of the evolutionary algorithm. Here parallelization proves valuable not only in terms of the number of calculations, but of the quality of the optimum reached over not only one but various algorithms.

Still quoting the documentation

"There are several common scenarios encountered when experimenting with MOEAs:

1. Testing a new algorithm against many test problems
2. Comparing the performance of many algorithms across one or more problems
3. Testing the effects of different parameters

Platypus provides the experimenter module with convenient routines for performing these kinds of experiments. Furthermore, the experimenter methods all support parallelization."

In Figure 16 we show how this parallelization allows to improve calculation efficiency.

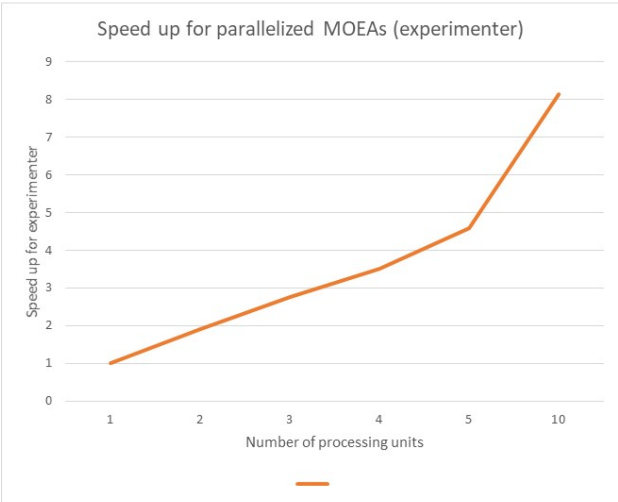


Figure 16 – Speedup of parallelized MOEAs over Experimenter module.

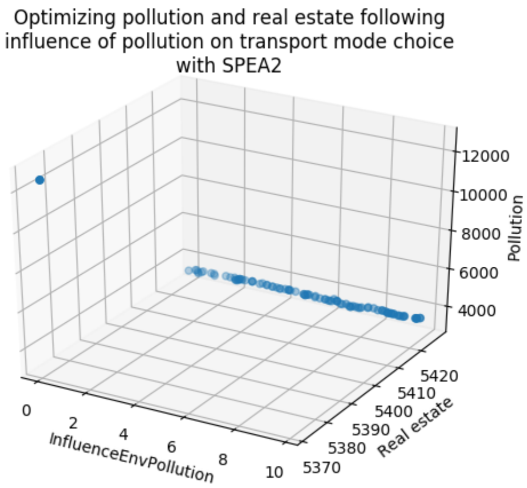
In the following table (see Table 3) we summarize a study of problems of increasing difficulty.

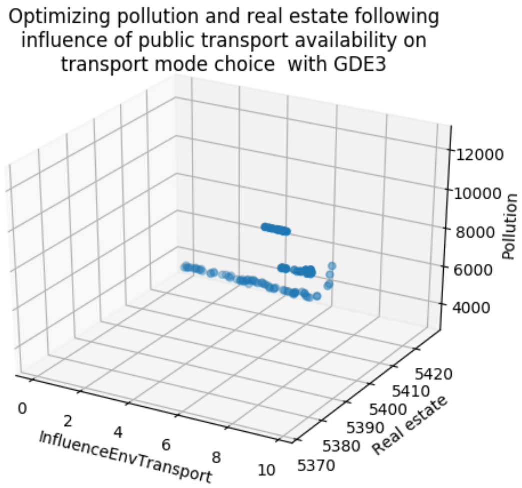
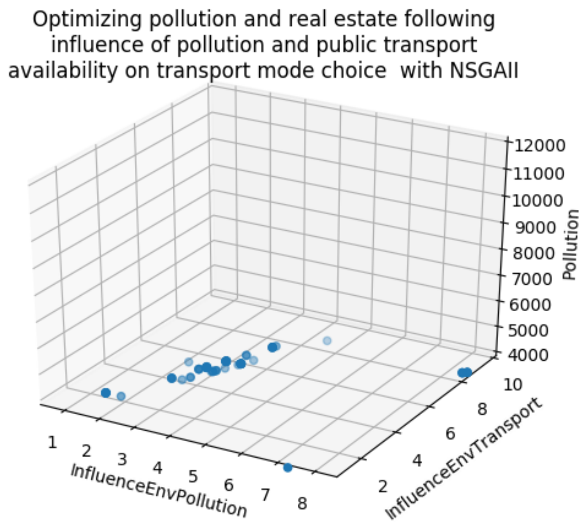
The graphs show objective values found by the algorithms for different parameter values.

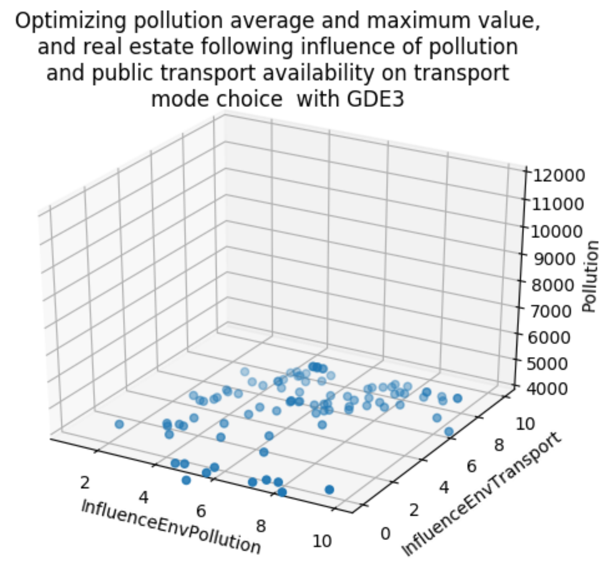
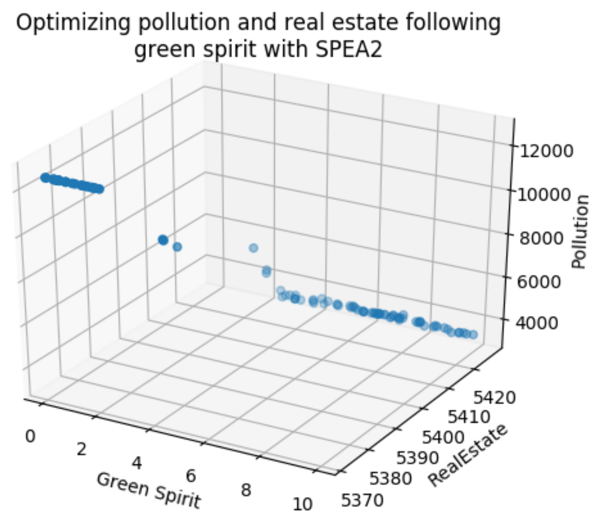
When there is only one parameter, it appears on the x (horizontal) axis, while the objectives correspond to the y (depth) and z (vertical) axes. When there are at least two parameters, they correspond to the x and y axes, with the pollution objective value appearing on the z axis.

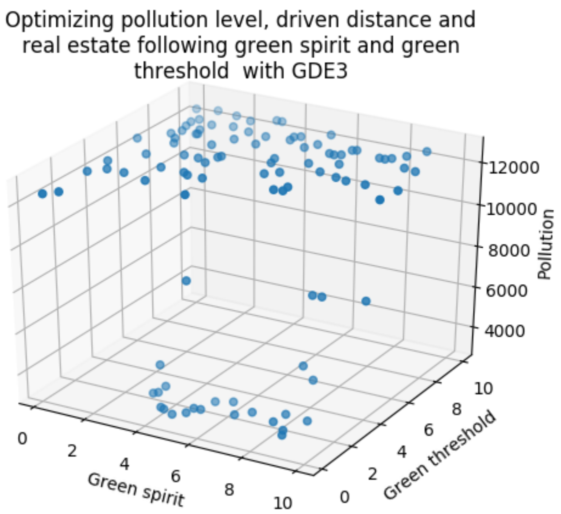
For a given problem, we call 'optimal' the algorithm(s) which find(s) the solution with the highest maxima and/or lowest minima values for the objectives. In the following examples, different algorithms appear to find the optimal result(s), depending upon the problem to solve. In some cases, this is not a single point in the objective space, and the best objective values are not all found by a single algorithm, highlighting even more the benefit of using more than one optimization algorithm.

Indeed, to summarize the conclusion of this study: due to the complexity of GSS models, different algorithms prove optimal following the specificities of every problem. This puts into light the benefit of associating them over parallelized runs to find the optimal solution, for which HPC proves valuable.

N objectives	N para meters	Algorithms finding best optima values	Comments and graphs
1. Optimizing pollution and real estate following the sensitivity to pollution of citizens when choosing their transport mode			
2	1	**CMAES > SPEA2**	This is a simple problem. Here most algorithms provide similar results. 
2. Optimizing pollution and real estate following the sensitivity of citizens to public transport offer when choosing their transport mode			

2	1	**GDE3** ~ MOEAD	This second problem is quite similar to the first one, except the parameter to optimize is now the sensitivity of citizens to public transport offer when choosing their transport mode. The influence of this parameter is less clear, since its influence depends also upon the effective evolution of the public transport offer. 
3. Optimizing pollution and real estate following the sensitivity of citizens to both pollution and public transport offer when choosing their transport mode			
2	2	~ NSGAII, NSGAIII, GDE3, **IBEA**, MOEAD	This problem is a synthesis of the two previous ones. Here a few algorithms perform equivalently well (NSGAII, NSGAIII, GDE3, IBEA, MOEAD) with real estate prices similar to the two previous cases, and pollution levels lower, thanks to the increased degree of freedom provided by the second parameter. 

4. Optimizing average pollution, maximal level of pollution and real estate following the sensitivity of citizens to both pollution and public transport offer when choosing their transport mode			
3	2	GDE3	This problem is similar to the previous one while introducing an additional objective: the maximum level of pollution observed. 
5. Optimizing pollution and real estate following the green spirit			
2	1	**SMPSO** > SPEA2	This is again a simpler optimization problem, studying the influence of the green spirit parameter, which defines how ecologically minded citizen are, and prone to choose public transport over using their car. 

6. Optimizing pollution, driving distance and real estate following the green spirit and green threshold				
3	2	N	GDE3	Now we complicate a little further the problem by adding another objective to minimize, the driving distance, and another parameter, the green threshold (green behaviours are favoured by a high green spirit value and a low green threshold). OMOPSO finds the optimal driving distance, but coupled with poor values of average pollution and real estate, and otherwise rather average values. GDE3 does quite well for the average pollution (finding its minimal value observed) coupled with a real estate close to the best one found, but a driving distance pollution in the average. Here the complexity of the problem doesn't allow to identify any solution as completely outperforming the other ones, which once again underlines the benefit of the varied approach and multiplicity of solutions proposed. 
7. Optimizing pollution, driving distance and real estate following the green spirit and public transport adaptability				
3	2	N	**MOEAD**	This case is similar to the previous one concerning the objectives and the number of parameters, however less insidious here since they are independent. Here MOEAD proves best, finding optimal pollution and real estate, coupled however with a driving distance over the average.

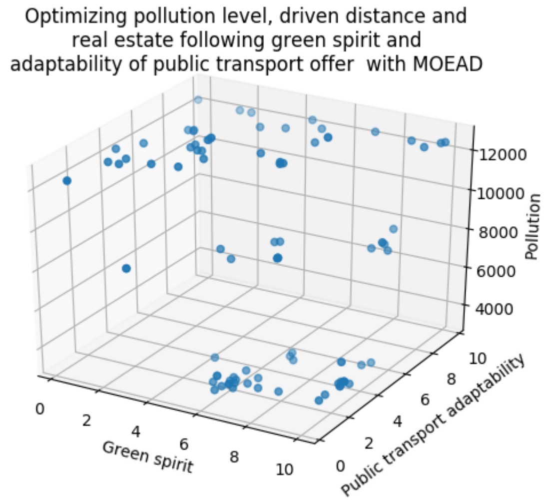
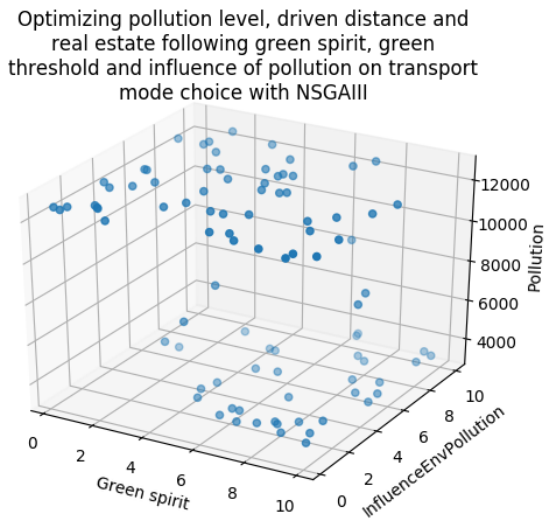
				
8. Optimizing pollution, driving distance and real estate following the green spirit, green threshold and sensitivity to pollution of citizens when choosing their transport mode				
3	3	N	**NSGAIII**	This complicates a little the previous problem by adding a third parameter, two of them being linked as previously (which makes it tricky). The best solution is found by NSGAIII for the pollution, real estate (close to optimal) and driving distance a little above the average. 

Table 3 – Comparing the performance of different MOEAs to solve city pilot optimization problems of increasing difficulty

4.4 Conclusion

The city pilot aimed to put into light and illustrate specific GSS needs calling for the computing resources of high performance computing centers. They concern the whole modeling cycle, starting with data pre-processing to results data post-processing and analysis. But they concern also the simulation itself, not only by allowing to refine agent granularity, which appears valuable for GSS models, but by allowing to multiply the simulations to explore a complex and unpredictable variation of observables in the parameter space, and parallelize optimization algorithms in a challenging quest to find better optima.2

Particuarly, these studies have allowed to develop a new city pilot centered on the two-way relationship between transport and real-estate pricing with different levels of interdependence and feedback. They have shown some scientific challenges of GSS modeling and how a complex systems approach can tackle them while benefitting from HPC computing resources. The studies have provided the opportunity to explore the benefit of analytics for pre-processing (by aggregating or disaggregating available data) and post-processing of data (for instance by calculating spatial heterogeneity). Finally they have allowed to explore the model behavior over multiple optimization approaches.

5 Future Applications

5.1 CPGC conference

The task on future applications benefitted from the open conference “Computing Power for Global Challenges” that is more closely described in D6.7.

A new topic in this conference that presents a field where CoeGSS methods and tools can usefully be applied is the question how digitalisation will play out in a world in which democracy is in danger as it was not for a long time; the conference contributed thoughts and discussions, that also linked the topic with the state of the European Union, as a starting point.

In the field of sustainable finance, that has been identified as a future application earlier (see D4.4 and D4.5), a concrete example was given by Camilo Mondragon from the International Finance Corporation from the World Bank Group. He introduced the IFC’s twin goals of ending extreme poverty and reducing inequality, and their activities of providing loans, investing in equity, and providing advisory services to private companies across all sectors of the economy, across the developing world. To understand how fostering economic activity in the private sector can help achieve these goals, models have to link private sector development with the distributional goals, thus including both firms, in particular small and medium, and a population of households at high granularity. Moreover, the financial sector needs to be included to represent a large part of IFC’s investment portfolio that is channeled through banks which in turn lend to small and medium companies.

While data on households is available in surveys, there is a lack of data for adequately modelling the private sector, representing planning cycles of firms, that for employment and investment decisions use all the information they can get at each point in time, but need to work also with expectations. A particular problem is given by the fact that in many of the relevant countries, large parts of the economy are informal. Synthetic populations of firms, that use the patches of data (some business surveys, sometimes average data) that are available and reconstruct or generate formal and informal sectors, so that models using these populations can then be calibrated, could help to address this modelling challenge.

More broadly, in the closing session, several participants indicated needs for future work and suggestions on how this can succeed. Social sciences, having to work in large part with the notion of an unstructured and moreover changing grid, pose challenges to current HPC architechtures in that proximity relations of the processors are not necessarily convenient and constant. To deal with this challenge, effort needs to go also into more abstract and conceptual work on the assumptions about the best abstractions for computing global systems. For example, what an agent is can be represented in many ways, and a change in representation may change how they are computed; thus changing abstractions could change computing characteristics.

A long-term vision on what one wants to achieve, taking account of what is feasible, is needed to converge on those paths that provide the most potential for solving given problems, that is, for addressing global challenges. In order to develop this, the background and concepts underlying the computing applications need to be shared between GSS- and HPC-experts. Progress, both in terms of problem solving and in terms of computational technology, happens when people interactively focus on applications at the edge of what is currently feasible, with a vision of what may be possible in the future. It also helps to imagine a future in which much more computation is possible, and “look back” on tools and techniques used today for studying social systems. Comparing a list of assumptions made in order to be able to solve models with how the real world works, computation provides a way to add more realistic features to models one on top of the other. Finally, this work takes commitment and patience.

5.2 Finance and Economics

The field of economics and finance having been discussed (both in previous deliverables and here above). As a sideline of the work in T4.4, the CoeGSS synthetic network tool, in particular, the Lin similarity, was applied to a dataset of small companies to detect the effects of Italian "contratto di rete" agreements. In a "contratto di rete" (literally: network agreement), companies agree to take part in a consortium and share their resources therein. This activity also served to test the network algorithm on smaller datasets. Restricted to the set of Tuscan firms for the industrial sectors of Leather Factories and General Manufacturing (the ones in which the network agreement is most frequent), no performance improvements were observed when analysing firms in the network agreement with respect to the others. The work shall be continued to enlarge the set of firms to the national level.

5.3 Delphi questionnaire

For a more systematic analysis, a Delphi exercise was conducted. The Delphi method is a systematic procedure for soliciting the advice of a number of experts and forging a consensus from that advice; it uses formalised questionnaires and, maintaining anonymity of single answers, establishes a statistical answer of the group. The group is then provided with feedback about that answer and a repetition of the survey allows participants to update their views based on the new information given by the feedback. Anonymity makes it easier for experts to revise their ideas, and avoids that group dynamics influence the results. Starting out from a general question - in our case “What are needs and opportunities for future HPC applications in view of global challenges?” - this question needs to be operationalised for eliciting experts’ views. The CoeGSS consortium did this by first defining dimensions and elements of interest, listed in the table below, in a group discussion (teleconference).

computation / technical dimension (c)	o scale of application (c1) o scalability of application (c2) o mechanisms used (c3) o machines it runs on so far (architecture, processors) (c4)

	o format of input and output data (c5) o use of data during computation (c6) o processing time (c7)
GSS application fields (a)	o health / epidemics (a1) o sustainability / climate policy / green growth (a2) o urbanisation (a3) o financial stability / financial networks (a4) o other (a5)
models (m)	o model types: ABM, other simulation models, statistical, economic models (CGE, DSGE, etc), network models (m1) o interaction networks in models (regularity, evolution) (m2) o maturity of models (m3)
data	o input (d1) o output (d2) o pre- and postprocessing (d3) o homogeneity / heterogeneity (d4) o sources (d5) o privacy (d6)
limiting factors (l)	

Table 4 – Dimensions and elements of interest in identifying needs and opportunities of future HPC-GSS applications

Then, members of the consortium proposed questions – keeping them as simple as possible – for exploring these dimensions and elements with the experts. An example, that combines the elements l, c6, m3, d1, was "Which of the following do you consider a limiting factor in computational work on global challenges? Please rank the limiting factors: computing time, data availability, model availability / maturity / implementation."

The so obtained questions were assembled into an online survey, that, in addition, contained questions geared to the interest in and willingness to pay for services, prepared by WP2, after initial questions about the background of respondents (field of work, methods currently used, type and size of datasets currently used).

The section on future HPC-GSS applications of the questionnaire asked

1. for an example where real-time large-scale simulation can make a difference for decision makers in the coming years
2. whether respondents currently consider computational factors (memory, RAM, # processors, etc.), data availability, and/or the availability, maturity, or implementation of models a limiting factor in computational work on global challenges
3. for estimates of time frames in which the fields of health / epidemics, climate policy / green growth, urbanisation, and financial stability / financial networks will produce applications which require high performance computing
4. for an estimate for which model types (such as ABM, statistical models, standard economic models, network models) computational resources required will increase most quickly over the coming five years
5. for an idea about the most relevant reasons for increases in computational resources, such as the number of parameters to be set, the number of agents / entities / nodes in a model, the complexity of the computation for the single agent / entity / node, interaction / network structures in models, the number of observables of interest, or the number of runs needed
6. for an opinion on which steps in modeling work (such as data pre-processing, data post-processing & results analysis, model simulation while designing the model, exploring the high-level behavior of the model, e.g., finding zones of risk or resilience, scenario simulations with the validated model, live input feed (e.g., data on traffic, social media data), or interactive visualization, e.g., live output feed), require the largest computational resources
7. for a judgement on the priority of showing the predictive capabilities of a simpler model vs. enhancing the computational scalability of the model vs. proving the model flexibility and easiness of generalization vs. enlightening the lack of data and showing where new data may improve predictions for the purpose of encouraging stakeholders to collect and release more detailed data (relevant to the development and validation of an agent-based model of a global system)
8. whether the more relevant case for computational work on global challenges is that where data is used for a single computation or the one where the same data is used for many computations
9. which of the following cases is the most relevant for models with an interaction network: to explicitly represent the interaction network between agents in a single object shared among the processes, to compute the interaction network at every

possible interaction using the agents' internal variables, or to use the explicit network only for interactions between groups of agents.

After an internal run with the consortium members, the experts who had been present at the two CoeGSS conferences (ICSP and CPGC, see D6.6 and D6.7) were invited to fill in the questionnaire. Due to limited participation, the second round after feedback was carried out only within the consortium, at the final technical meeting of the project. Results for each question are summed up below. In general, the second round gave more distinct results on some of the questions, confirming the Delphi method's underlying idea that a reflection of results from a previous round can elicit a sharpened view on future issues.

Answers to the first question spanned a wide range of global challenges, including those that are addressed by the pilot studies or closely related work (such as in epidemiology). Other than these, banking supervision, financial crises, logistics, real-time management of smart energy grids, disaster management and early warning, flooding event simulations, genomics, and digital humanities were suggested. In the second round, the topics migration, urgent computing for example in disaster simulation, earth observation as input for land-use, climate change and others, as well as terrorism were added. The breadth of the spectrum covered by these examples indicates that HPC for GSS could become a large field. A "meta-level" answer recommended that in interactive dialogues with decision makers, real-time adaptation of model output could be very useful to make scientific content available and more transparent.

For question 2, data availability was identified as the most urgent limiting factor in computational GSS modelling; about 80% of respondents marked this, while model availability (about 54%) and computational factors (about 38%) were not considered as important in limiting computational work on global challenges. The second round showed an even stronger focus on data availability (91%), and a reduced consideration of computing power as a limiting factor (38%).

Answers to question 3 show that the field of health and epidemics is seen at an advantage compared to the other fields that were listed (see Figure 17), with about two thirds of respondents considering 0 to 2 years plausible, while for the other three fields, about the same number of respondents (slightly less for financial stability), consider 3 to 5 years a plausible time frame for applications to require HPC – specified in this case as e.g., 500 CPU for a single run, 5000 for ensemble simulations, 2 GB of memory per CPU, 100GB of data to be pre-processed, 1 TB of output data to be analysed, and/or Full HD times 25 frames per second for visualisation.

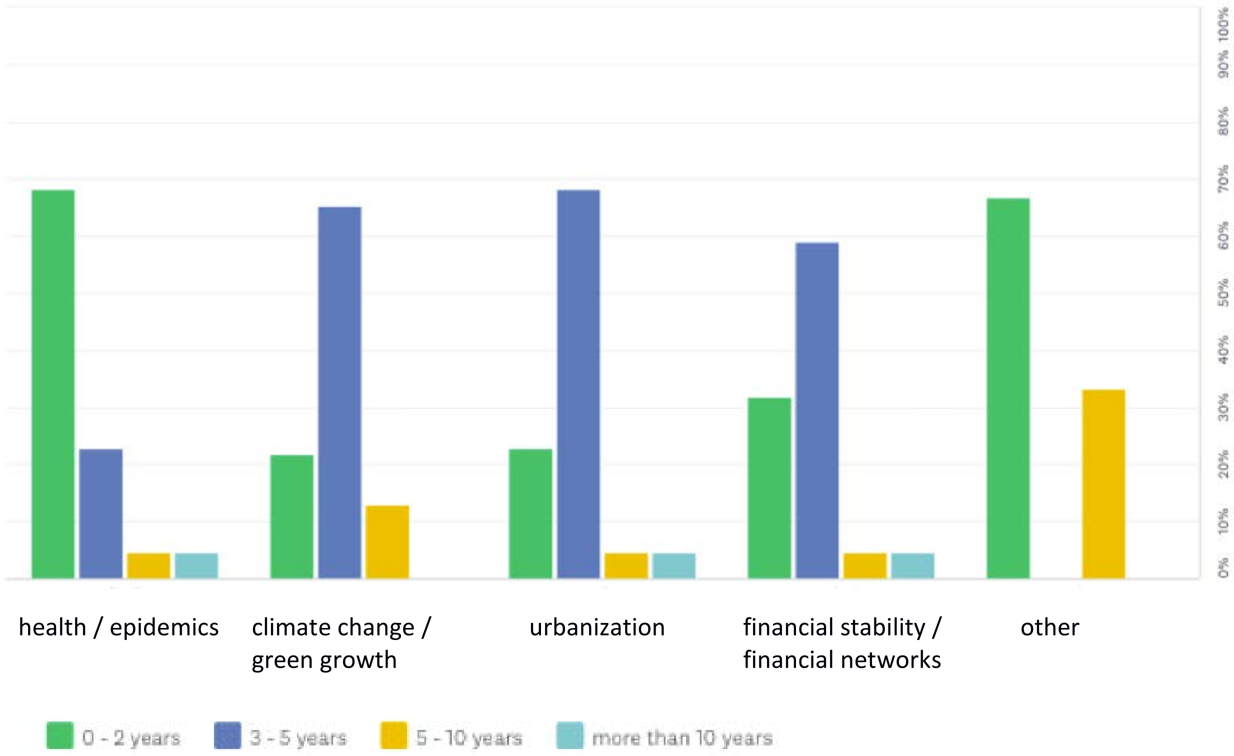


Figure 17 – Estimated timeframes for fields producing HPC applications

Answers from the second round were slightly more pessimistic on these timeframes: the majority estimates were 3-5 years for health and epidemics, and 5-10 years for urbanisation, while the climate policy/green growth estimates were tied between 3-5 and 5-10 years.

On question 4, ABM (82%) and network models (78%) are those for which the largest increase in computational resources required is expected, before statistical models (60%). The second round was even more decisive on ABMs, all participants saw these as requiring large computational resources, while the estimate for network models decreased to 45%.

The model types relate closely with the factors considered the most likely reasons for increases in computational resources needed in question 5: interaction and network structures in models (70%), before the complexity of computations for the single agent or node (60%), and the number of agents (55%). Here, the second round gave more balanced answers but still saw the same reasons as the most important ones.

The ranking of steps in modelling work does not show particular steps considered most relevant, an observation confirmed by the second round. Data pre-processing, model simulation while designing the model, live input feed and interactive visualisation score slightly higher (above 4) than the remaining options (3 for exploring high-level behaviour of the model, being the lowest value).

Similarly, the priority of different tasks for encouraging stakeholders to share data does not show any particularly favored option, in either of the rounds.

For the next to last question it turned out that the options were initially phrased in an ambiguous way; therefore, here only the second round result seems usable: 45% of respondents consider the same data used for many computations the more relevant case.

On the final question the option to use the explicit network only for interactions between groups of agents remains below 10%, while an explicitly represented interaction network between agents in a single object shared among the processes as well as a network computed on the fly are considered the more relevant cases by 38%, respectively 35% of respondents. In the second round, respondents see the on the fly computation of the network as even more relevant.

5.4 Future application in praxis

As a further activity of T4.4, throughout the overall project duration, members of the consortium tried to “recruit” a future GSS application to the CoeGSS computing infrastructure. For example, two participants of the two conferences organised by CoeGSS, whose work revolves around economics and finance, showed an interest in testing their models within the CoeGSS context. Initial conversations were positive, however, in both cases, the extra effort of moving to HPC turned out to be impossible with the everyday constraints research groups face in terms of time and efforts. With a longer planning horizon, and as part of a funded project, this effort would, however, have been possible: both research groups were among the new partners in the follow-up proposal EDGE (Exascale, Data, and Global Evolutions) of CoeGSS.

6 Challenges Encountered and Lessons Learned

This section summarizes challenges encountered and lessons learned in the pilot workpackage throughout the CoeGSS project. Starting out from bringing together HPC and GSS (Section 6.1), we discuss two elements of the pilots' synthetic information systems where using HPC for GSS plays a role (agent-based models in Section 6.2 and simulations in Section 6.3) and conclude with a view on computing power for global challenges, in particular, digital decision support in view of such challenges. Therefore, this section also serves as a conclusion to this document.

6.1 Two communities

As the project brought together two previously unrelated fields, an effort had to be made in order to develop a common language and a shared vision of CoeGSS; the pilot work was a particular "meeting point" between GSS and HPC. While the challenge was known to exist from the outset of the project (see, e.g., D4.1, Sections 2.1 and 2.2), it turned out to be more difficult than expected. In order to create interdisciplinarity, people are required to give up pieces of their "disciplinary identity", as rightly pointed out by members of the CoeGSS advisory board. Since this is not easy, a lesson learned in this respect is to not take this kind of challenge lightly, and in particular, to make it explicit. In fact, this was done with T2.3 in the project. The related work is reported on in D2.4.

6.2 Agent-Based Models

A challenge encountered with ABMs, the "workhorse" in synthetic information systems, was the tension between generic usability and specific codes. As described in D4.2, no ABM framework for GSS models on HPC, that was suitable for all pilots, was available off the shelf; and, moreover, it is clear that there is a tradeoff between generality and efficiency (see, e.g., Murphy, 2014). However, in a field where the modelled systems consist of many entities in complex interactions for which even the micro-level behaviour is far from well-established and generally agreed upon, model development implies an iterative process with many loops of model building, testing and improving or refining. In this process, GSS modellers, who are mostly not experts in parallel programming need to be able to implement parallel prototypes of large-scale ABM. To this end, a framework is needed, so that modellers can arrive at the point where they can decide that an efficient special purpose implementation is worthwhile. Such a framework, in turn, needs to be generic enough to accommodate a variety of GSS problems.

The Global Urbanisation pilot, being led by the SME CoSMo, used the CoSMo Modelling Suite, which is proprietary software; this represents one possible type of GSS modelers wanting to use their models on HPC. Another type, i.e., GSS researchers who want to develop new models with the help of an open source ABM framework, has been the case of the Health Habits and the Green Growth pilots. These implemented first, more basic models using

Pandora (Rubio-Campillo, 2014), but then resorted to tailor-made solutions for more sophisticated models. For the case of the Green Growth pilot, this led to the development of an ABM framework for prototyping parallel GSS-ABM, which is presented in D3.8. This project output can be considered a lesson learned in itself: in developing the Green Growth pilot's model, it turned out that a generic basis of elements (e.g., the definition of locations, agents, etc.) could be extracted and further developed into a framework to be reused for implementing other GSS ABM. To seize this opportunity, this framework was included into the project outputs in this third year. While the tension between generality and efficiency is probably unresolvable, it may be alleviated in the modeller's day to day work by using this tool.

6.3 Simulations

Two sources of uncertainty in GSS modelling and models provided the challenge that any model needs to be run many times: on the one hand, in the process of model development, parameters are defined and need to be calibrated so that the model output best fits available data; on the other hand, models are stochastic, and therefore, any scenario consists of a number of runs, that provide statistical information (often, one looks at averages and variation) on the system's potential evolutions.

In the calibration process, the goal is to systematically explore the parameter space, defined by considering relevant ranges for all parameters. This is done by sampling and may be referred to as a parameter sweep. As the influence any parameter has may depend on the values chosen for other parameters, combinations of parameters need to be probed, which implies quickly increasing numbers of runs to be made for increasing numbers of parameters. There are various sampling techniques, including full factorial design (all by all parameters), simple random sampling, Latin hypercube sampling (dividing the ranges for all parameters into intervals of equal probability and drawing once from each interval combination), and optimisation methods, which iteratively determine the next parameter configurations based on previous results. While some of these help reduce the number of runs needed (e.g., Latin hypercube sampling as compared to full factorial), nevertheless large numbers of runs are needed to explore the parameter space.

Once a model is calibrated, the information one wants to obtain from it is of statistical nature. Stochasticity in the models represents the fact that in any portrayal of a global system, elements unknown to the modeller will necessarily remain – these are summarily represented by stochastic processes. Hence, the presentation of results of models will always need to include a presentation of uncertainty. Sets of simulations are therefore needed also for each scenario that is to be computed with the model.

Together, the simulations in these sets are referred to as ensembles: the single simulations, usually moderately parallel, are run independently from each other. A challenge encountered in the project, that could also have been mentioned in the section on communication, was that such "embarrassingly parallel simulations" are of course not the prime application for

HPC, and the case needed to be made that it was nevertheless a pilot requirement. As data from these runs have to be collected and evaluated across runs, however, there is scope for not moving to cloud computing with this task. In order to meet the ensemble-challenge, the Dakota tool was used (see Section 3.2).

6.4 Computing Power for Global Challenges

In summary, the pilots have produced synthetic information systems that can support discussions with stakeholders, who can be customers of CoeGSS, by visual presentation of model-based results, and that can further evolve according to the inputs from and needs arising in such discussions. Here, large computing power plays a role in several respects:

- Agent-based models come with advantages in this process. A model in terms of the actors in a global system, represented as agents in the model, can be more easily explained to non-modellers than, for example, one in terms of differential equations. Assumptions on how agents will react to certain decisions, for example by policy makers or business, can be discussed. Then, a given ABM can be adapted to reflect these behavioural assumptions, again, more easily than an aggregate model. It is of advantage here if one is able to model a population at full scale, otherwise, aggregation is an intermediate step to be considered, leading to questions like "which model artefacts are created if one model agent represents 10, 100, 1000 real world persons?". Going towards full scale ABM simulations of the evolution of global systems, the required computing power increases per single simulation.
- Spatially explicit simulations and calibration on real data are additional reasons for full scale ABM simulations. Dealing with agents that aggregately represent a given number of real-world individuals means that this is the smallest possible number of individuals that can be looked at. In a spatially explicit model, one may have to deal with small numbers of individuals (such as the number of early adopters in an innovation diffusion process in a small spatial unit), wherefore aggregation may need to be avoided. For example, in the Green Growth pilot's model, if an agent represents 400 people, and buys an electric vehicle, this means that in real-world terms, 400 EVs are bought in the same spatial cell in the model.
- Running many simulations is then necessary as discussed above, and it needs to be done in a time frame that can support decision making. Depending on the relevant problem scales, for pandemics, this time frame may be weekly or bi-weekly, while for finance, daily decisions may need to be supported. This is a further point where computing power plays an important role.
- Visualisation is the final point, since large data sets are created as output of ensemble simulations and need to be drawn on in an interactive environment in decision support discussions.

Building on the work done by the pilots, and using the methods and tools developed by CoeGSS, future applications can benefit from this project's experience and work to draw on "computing power for global challenges", as was the title of the second conference organised by CoeGSS. This event, as the previous one, described in D6.5, showed that there is a general interest in being able to use HPC for addressing global challenges. A lesson learned, in particular from the activities in T4.4 on future applications, however, is that the concrete step to using the computers at an HPC centre requires an initial investment of time and efforts that can be prohibitive. While CoeGSS has made steps to bringing together GSS and HPC, further work will be needed to establish this new combination alongside other application fields that have co-evolved with HPC from the beginning.

7 References

7.1 CoeGSS deliverables

D2.4: CoeGSS Deliverable D2.4; to be released.

D3.8: CoeGSS Deliverable D3.8; Report of framework for prototyping of parallel Agent Based Modelling Systems; delivered to the European Commission; Andreas Geiges (editor), Sarah Wolf, Gesine Steudle, Steffen Fürst

D4.1: CoeGSS Deliverable D4.1; FIRST REPORT ON PILOT REQUIREMENTS; delivered to the European Commission; Sarah Wolf (editor), Margaret Edwards, Steffen Fürst, Andreas Geiges, Alfred Ireland, Franziska Schütze, Gesine Steudle, Daniela Paolotti, Michele Tizzoni

D4.2: CoeGSS Deliverable D4.2; SECOND REPORT ON PILOT REQUIREMENTS; delivered to the European Commission; Sarah Wolf (editor), Margaret Edwards, Steffen Fürst, Andreas Geiges, Luca Rossi, Michele Tizzoni, Enrico Ubaldi

D4.4: CoeGSS Deliverable D4.4; FIRST STATUS REPORT OF THE PILOTS; delivered to the European Commission; Sarah Wolf (editor), Marion Dreyer, Margaret Edwards, Steffen Fürst, Andreas Geiges, Jörg Hilpert, Jette von Postel, Fabio Saracco, Michele Tizzoni, Enrico Ubaldi

D4.5: CoeGSS Deliverable D4.5; SECOND STATUS REPORT OF THE PILOTS; delivered to the European Commission; Sarah Wolf (editor), Andreas Geiges, Steffen Fürst, Enrico Ubaldi, Margaret Edwards, Jette von Postel

D6.5: CoeGSS Deliverable D6.5; FIRST ANNUAL REPORT ON TRAINING, STANDARDIZATION, COLLABORATION, DISSEMINATION AND COMMUNICATION; delivered to the European Commission; Sarah Wolf (editor), Andreas Geiges, Michael Gienger, Fabio Saracco

D6.6: CoeGSS Deliverable D6.6; SECOND ANNUAL REPORT ON TRAINING, STANDARDIZATION, COLLABORATION, DISSEMINATION AND COMMUNICATION; delivered to the European Commission; Sarah Wolf (editor), Michael Gienger, Fabio Saracco, Jette von Postel

D6.7: CoeGSS Deliverable D6.7; THIRD ANNUAL REPORT ON TRAINING, STANDARDIZATION, COLLABORATION, DISSEMINATION AND COMMUNICATION; Sarah Wolf (editor), Sebastian Böhne, Michael Gienger, Fabio Saracco, Jette von Postel

7.2 Literature

Dakota User's Manual (2014), A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.7 User's Manual. Online. Available at: https://dakota.sandia.gov/content/manuals

Deutsche Bundesregierung (2011). Regierungsprogramm Elektromobilität. [Government programme electric mobility] BMWI, BMVBS, BMU, BMBF.

Dum, R., & Johnson, J. (2017). Global systems science and policy. In: Non-Equilibrium Social Science and Policy (pp. 209-225). Springer, Cham.

Hall, D., & Lutsey, N. (2017). Emerging best practices for electric vehicle charging infrastructure. The International Council on Clean Transportation (ICCT), Washington, DC.

DLR (German Aerospace Center) (2010) "Mobilität in Deutschland 2008. [Mobility in Germany 2008]." www.mobilitaet-in-deutschland.de

Mielke, J. & Geiges, A. (2018). Model-Stakeholder Interactions for a Sustainable Mobility Transition. Working Paper 2/2018. Global Climate Forum, Berlin. Available at: https://globalclimateforum.org/wp-content/uploads/2018/09/GCF_WorkingPaper2-2018.pdf

Murphy, J.T. (2014). High Performance Agent-Based Modeling in Repast HPC. Computation Institute, Chicago. Available at: https://www.ci.uchicago.edu/events/high-performance-agent-based-modeling-repast-hpc

Renn, O., & Schweizer, P. J. (2009). Inclusive risk governance: concepts and application to environmental policy making. Environmental policy and governance, 19(3), 174-185.

Rubio-Campillo, X. (2014). Pandora: A versatile agent-based modelling platform for social simulation. *Proceedings of SIMUL 2014 the Sixth International Conference on Advances in System Simulation*, 29–34.

“shiny” (2017). Online. Available at: https://shiny.rstudio.com