CoeGSS

Centre of excellence

D4.4 – First 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	30.09.2016
Actual date of submission	30.09.2016
Dissemination level	Public
Nature	Report
Version	2 (after internal review)
Work Package	4
Lead beneficiary	GCF
Responsible scientist/administrator	Sarah Wolf
Contributors	Marion Dreyer, Margaret Edwards, Steffen Fürst, Andreas Geiges, Jörg Hilpert, Jette von Postel, Fabio Saracco, Michele Tizzoni, Enrico Ubaldi
Internal reviewers	Ralf Schneider, Patrik Jansson
Keywords	Health Habits, Green Growth, Global Urbanisation, Synthetic Information System
Total number of pages:	106

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		11.09.2016
Second Version	for submission		27.09.2016
Reviewed by	Patrik Jansson Ralf Schneider	Chalmers HLRS	19.09.2016 19.09.2016
Approved by	Coordinator	UP	28.09.2016

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 – at the end of the first project year. The pilots are working on 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 population and its emissions (Green Growth), and the two-way relation between transport infrastructure decisions and price mechanisms, particularly concerning real-estate (Global Urbanisation). Progress is presented both in common WP4 work accross pilots, for the synthetic information system of each pilot, and for the Future Applications task that completes WP4. It includes an architecture of synthetic information systems for Global Systems Science, an evaluation of two HPC frameworks for agent-based modelling, the identification of features of agent networks to be modelled, a literature study on autonomous driving as well as a modelling literature review, conceptual model specifications, collection and pre-processing of data, model implementations, simulations run, both on local and on high performance computers, simulation analysis, and finally the identification of two possible future applications: financial stability and the potential of the blockchain technology in addressing global challenges.

Table of Contents

List of Figures

1 Introduction

This document presents the first year project work of the pilot studies in the Centre of Excellence for Global Systems Science (CoeGSS) – Health Habits (HH), Green Growth (GG), and Global Urbanisation (GU) – that brings together High Performance Computing (HPC) and Global Systems Science (GSS).

The pilots are developing HPC-based synthetic information systems for these three global challenges aiming at identifying examples for turning global challenges into business and policy opportunities. As prototypical applications, the pilots derive requirements to steer the evolution of the centre and together they develop the architecture for synthetic information systems suitable for the study of global challenges. The present document first outlines common work between and around pilots in Section 2 and then goes into detail for each pilot in Sections 3 – 5. While some points presented for each pilot are of relevance also to other pilots, the text on each pilot should also be readable as a complete description of the pilot's status at M12. Therefore, we refrained from cutting into the single texts to draw together all pieces of common relevance in one place. The reader will hopefully excuse the resulting slight overlap in content between different pilot chapters in a few points.

In addition to the three pilots, which constitute tasks 4.1 – 4.3, the workpackage (WP4 – 'Pilots') contains a fourth task that is of a different nature: it aims to identify needs and opportunities for future GSS-HPC applications. Its progress is briefly summarized in Section 6.

While the pilots are the topic of WP4, work on and around the pilots also comprises interaction with other workpackages, in particular WP3 – 'Methods and Tools for GSS on HPC' and WP 5 – 'Centre Operation', e.g. on scalability and on data pre-processing.

2 Status of common work across pilots

In the ongoing interaction and discussions between the pilot studies, commonalities have been carved out in the motivation, and in mechanisms relevant for the questions under consideration. Sections 2.1 and 2.2 sketch these, respectively. Then, Section 2.3 summarizes the common approach of the pilots: building a synthetic information system (SIS), and starting with a simple version to add complexity in a step-by-step manner. An agent-based model (ABM), initialised with a synthetic population, is the backbone of a GSS-SIS, wherefore an HPC-framework for agent-based modelling is an essential tool. Section 2.4 sketches why the Pandora framework for agent-based modelling has been chosen as a central tool. Within the ABM, the networks in which agents interact play an important role wherefore modelling these networks is an overarching topic for the pilots. Section 2.5 presents first steps made in this direction. A further activity in WP4, crosscutting over pilots, is a literature study on autonomous driving – a topic of interest both for the Green Growth and the Global Urbanisation pilot. First steps for this literature study are presented in Section 2.6 before Section 2.7 summarizes the common work across pilots.

2.1 Expected benefits of using HPC for GSS

GSS tackles global challenges in complex social systems. Figure 1 illustrates how HPC can improve computational modelling for GSS (using the Global Urbanisation example):

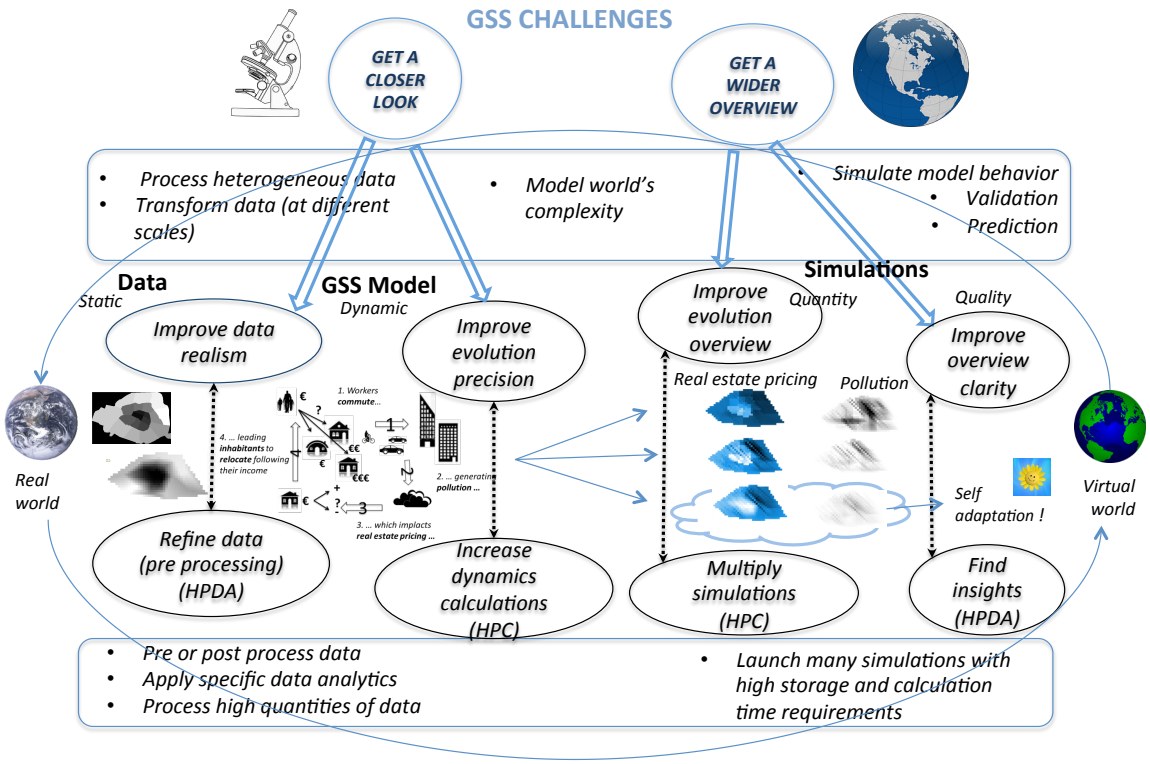

Figure 1: Benefits of HPC in GSS

In particular, GSS modelling aims at

- getting a closer look at basic processes by tackling finer scales and simulating heterogeneity, particularly concerning interactions. Larger detail in the representation of a system can discover dynamics that would not be visible using aggregate scale models. To decide upon optimal strategies, it is essential to identify and understand in a system, every element of influence, which can help decrease social and economic cost while increasing efficiency.
- getting a wider and clearer overview of the dynamics of a system to enhance its understanding by simulating not only usual use cases in a limited time frame, but exploring the range of possibilities to look out for possible risks to be avoided, discover innovative solutions, and foresee long term evolutions and consequences, particularly by evaluating high level system indicators such as resiliency.

Computing power is relevant to both these challenges, as

- getting a closer look requires
  - improving (static) data realism by enhanced pre-processing, which can benefit from high performance data analysis (HPDA), for instance to calculate statistics from real population data sets to generate realistic virtual ones, accounting for possible variability or uncertainty
  - improving (dynamic) evolution precision, for instance in performing calculations over a higher number of agents or refining aggregate parts of the model, refining time steps, running a simulation over more time steps.
- getting a wider overview requires
  - increasing the number of simulations with fixed parameters to explore uncertainty and with varying parameters to explore a large parameter space
  - applying analytical calculations on simulation results (at various conceptual, spatial, and temporal scales) to help identify possible high level patterns and observe emergence or any other insight on the system's evolution

How HPC and HPDA can be most beneficial to tackle these challenges in GSS computational model by providing what is required is the underlying question of the pilot studies.

2.2 Commonalities of mechanisms investigated

Social habit diffusion is a common basic mechanism in all three pilot studies. Further similarities in fundamental mechanisms have been found in particular between the Green Growth Pilot, which studies the evolution of the electric car market, and the Health Habits pilot, which studies the habit of smoking.

Both pilots strive to simulate the habits of humans related to a rapidly changing market of products involving externalities. The market for mobility is currently revolutionized by the rise and accelerated development of electric cars. While being substitutes for conventional cars (both are used for individual transport of people and small amounts of goods, such as

luggage), electric cars provide substantially different features and benefits (reduced noise, emissions, fewer maintenance needs, etc) but at the same time they suffer from different kind of restrictions (like the limited range). Furthermore, the gap between motorcycles and cars is filled by new vehicle concepts, based on the electric engine, e.g., the Twizy by Peugeot. Thus, the customer reaction to these different newly available options will be key for the development of the future mobility market.

Similarly, the electric cigarette may be changing the market and the habits of smoking. Again, different benefits and features are offered and the reaction of the costumer is hardly predictable. Thus, similar model approaches can be used for both pilot studies, which will foster the synergies of collaborations between pilots.

On the other hand, the externalities of these types of products are rather different. Cigarettes mainly harm the user itself, however usually with a large time delay, and only cause secondary-level effects on others. To minimize these secondary effects, many anti-smoking laws have been applied. Hence, electric cigarettes mainly provide benefits to the consumer. For electric cars, however, emission and noise mainly harm pedestrians, residents and cyclist, whereas the car's air filters protect the driver. The climate change externality involved has a time horizon which mostly concerns later generations. Thus, the decision of buying an electric car seems not to provide such strong benefits for its user, but rather for the community.

A similarity between the Green Growth and the Global Urbanisation pilots arises in the importance of the underlying geographics. The Global Urbanisation pilot explicitly addresses the link with geographical information systems. For the Green Growth pilot, the geographical location of an agent and its features (urban or rural, with the implied commuting distance and transport needs, charging infrastructure availability, etc.) are important components in the agent's decision making.

In all pilots, the object of study is a complex dynamic system, with mutual influences between the interacting agents, characterised by incomplete knowledge of the individuals, learning, communication and decision-making. Synthetic populations shall resemble the real world population in its statistical properties and distributions in social simulations, without revealing the privacy of the actual persons.

2.3 The common SIS architecture

What we mean by a synthetic information system has been described in D4.1 – "First Report on Pilot Requirements". Here, we summarize the basic architecture of an HPC-SIS for GSS. Such a system represents a real-world system on a computer to run simulations for exploring, as in a virtual laboratory, possible scenarios of the system's future evolution and thus help assess possible consequences of decisions. Figure 2 displays its elements in boxes that name the element, describe what it is (if not self-explanatory), and describe related tasks.

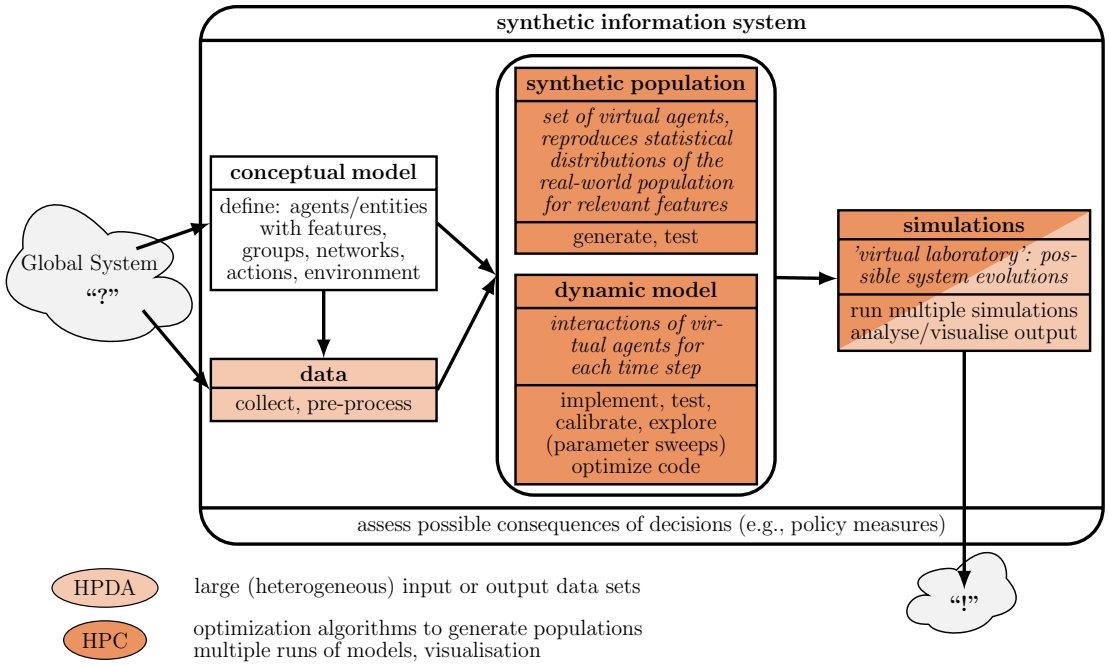

Figure 2: The structure of an HPC-SIS for GSS. Colours indicate where HPC and HPDA play a role.

Given a **global system**, a vaguely defined part of the real world, and a (research) question one tries to answer about this system, there are (at least) two ways of grasping the global system: concepts and data.

First, a **conceptual model** is needed to determine the boundaries of the system (what does one include, what is left out?) and define relevant elements (what are the agents, which characteristics of these agents, which interactions and networks between them are relevant?).

With this conceptual model in mind, **data** needs to be collected and possibly pre-processed to be available in the form and at the level of completeness that is needed.

The system representation on the computer can be considered a "synthetic global system". It consists of a **synthetic population** – a set of virtual agents that, for relevant characteristics statistically match the corresponding distributions found in the real-world population – and a **dynamic HPC model**, an agent-based model, which implements the (inter-)actions of agents. Running the dynamic model, i.e. repeatedly carrying out the agents' interactions, the synthetic global system simulates the evolution of the real-world system.

Sets of **simulations**, reflecting uncertainty in global systems, are explored with the help of analysis and visualisation tools to answer the given question. Compared to traditional simulation workflows, agent-based simulation in social sciences requires new modelling approaches for model development, as well as analysis tools. In ABMs, dynamics are implemented at the level of the agents, the "micro" scale. This scale allows mimicking the behaviour and decision making of people directly and by simple and comprehensible rules. Observation data of the system, however, are usually available as aggregated statistical data, surveys and general behavioural patterns. Thus, the micro dynamics and interaction of all

agents together need to match the observed dynamics of the full system at the macro scale (see Epstein 1999). This is comparable to molecular simulations, where e.g., all forces between molecules add up to the macro behaviour of a fluid like the surface tension. In contrast to molecular simulations or any physics-based simulations, in social science, the atomic rules of agents on the micro scale for the agents are not given by widely agreed upon physical laws. Instead, agents' (re-)actions change over time, are different for each individual, their living situation and education and, thus, are constantly subject to discussion. Therefore, the development process of a conceptual model in social science is intrinsically different from approaches in natural sciences. Building a SIS usually requires an iterative process. Many steps will reveal a need to go back to previous steps and make changes until one obtains a SIS that may confidently be used to explore, for example, the effects that different policy inputs may have on a global system.

Accounting for this fact, the pilots' strategy from the beginning has been to start simple and increase complexity step by step in order to quickly gather experience with all components of a SIS and to begin using HPC quickly to identify challenges and opportunities. Also, the pilots shall engage stakeholders as soon as possible. As a basis for fruitful dialogues, a "demo" SIS is needed to render the idea – such a dialogue cannot be based on plans for a SIS alone. At the same time, it must be possible to shape a more complex SIS together with stakeholders if one wants to address users' needs, providing another motivation for starting with a simpler SIS that can then be refined.

2.4 HPC-frameworks for ABM

Deliverable 3.1 lists frameworks for agent-based HPC modelling. Out of these, the open-source code Pandora (Rubio-Campillo, 2014) has been used for model implementation of the Green Growth pilot. For comparison, Repast HPC has been given a closer look from the pilot perspective.

From the modeller's perspective, the most immediate features of a modelling framework are different aspects of its ease of use: documentation and examples for getting started, flexibility of what kind of model elements and structures can be implemented with the help of a framework (e.g. in terms of communication between agents, ease of importing data and usability of the output produced by simulations), code quality when looking at the details, ease of implementing a model using the framework, and availability of tests as well as a low propensity for implementation errors.

From the HPC perspective, performance and scalability issues of the code implemented and generated with the help of the framework are important issues. Finally, for the development of sustainable business applications in CoeGSS, the licence of the software used needs to be considered as well.

In the process of reviewing different frameworks, different solutions provided by the individual frameworks were evaluated. On the one hand, this allowed reflecting upon common requirements of ABM, applied on HPC. On the other hand, being aware of the

multitude of different approaches and implementations will allow us to choose the appropriate solutions for future development and extensions in the Pandora code. The following sections cover the most relevant requirements investigated.

2.4.1 Documentation and examples

The documentation (for the model implementer) is much better in Repast than in Pandora, but still incomplete. Pandora comes only with a short tutorial, Repast has a detailed tutorial and additionally a 43 page long manual.

On the other hand, Pandora provides a much larger set of examples to see the features the framework provides in practice. This helps potential users to get familiar with the Pandora code. CoeGSS could provide well-documented GSS examples via the portal as an appetizer for potential users.

2.4.2 Communication

Pandora was originally developed for historical simulations that do not need telecommunications. Therefore, there is a maximum interaction range for agents located in a spatial grid. Repast, on the other hand, has implemented support for different communication patterns between agents:

- In principle each agent can always get the information of any other agent (using calls of Repast functions), independent of their processes.
- On top of this, so-called projections can be defined. A projection can have a network structure or a spatial structure. Like in Pandora, agents occupy positions in a 2-dimensional space.[1]
- The Repast documentation suggests that Repast's spatial structure also has a limited communication range.
- In Repast all communication across processes is "read-only". The state of an agent projected into a process is updated when the state is changed by the process that owns the agent, not when the state of the copied (called borrowed in Repast/ghost in Pandora) agent is changed. This is different from Pandora, where also changes in the copied agent are synchronized with the original agent.

Thus, a current challenge is to include a long range communication mechanism in the Pandora code. We strive for a general and efficient solution that does not raise too many restrictions to the user, while allowing an efficient parallelization as well. Thus, this task is approached in close cooperation with WP3, which allows to carefully consider all available

[1] http://repast.sourceforge.net/hpc_tutorial/RepastHPC_Demo_03_Overview.html

technical options and libraries. Currently, a general specification of such a communication mechanism is developed and will be part of later deliverables.

2.4.3 I/O for the modeller

In Pandora, GeoTIFF files can be used to initialize spatial data, and GIS tools can read the dynamic spatial data output from simulations. Repast on the other hand, does not have any GIS-support, and any kind of spatial support for storing an attribute value per cell seems to be lacking. Results can be written using the NetCDF or csv formats; how this is done is not clearly documented.

Both frameworks do not have any explicit interface for reading a pre-generated synthetic population, that is, reading agents with individual, exogenously given, attribute values. At the current stage, we plan to use HDF5 for all different types of input data. HDF5 is flexible enough to implement the same functionality as GeoTIFF for the geographical referencing. Further, its dynamic data structure allows storing synthetic population data. Thus, we will create a general interface for the efficient storage and transfer of arbitrary populations in cooperation with WP5. The efficient communication of such data between the computing nodes presents another challenge that will be tackled together with WP3.

2.4.4 Quality of source code

The source code of Repast looks better-structured, decoupled and written more generic than Pandora. However, for the modeller, this presents a trade-off between flexibility and simplicity. It was easy to understand the program structure of Pandora and perform small changes. This cannot be expected to be the case in Repast.

Some decisions in the Pandora code seem disputable, e.g. using a concatenate string to represent the agent type and id. While a final statement about this cannot be made on Repast from the pilot side, the impression is that the decisions taken by the Repast team are generally more thoughtful.

2.4.5 Flexibility vs. complexity

In general, Repast has a higher flexibility than Pandora, not only for the communication patterns mentioned above. For example, it works with schedulers (not to mix up with the class `Scheduler` in Pandora) that call a registered function at a given step in the simulation[2], while in Pandora there are just predefined functions that could be overwritten by the model implementation (e.g. World::stepEnvironment). However, the flexibility of Repast comes at the cost of higher complexity. For example, to register a new scheduler event, a function call with the following form has to be written:

[2] http://repast.sourceforge.net/hpc_tutorial/RepastHPC_Demo_00_Step_08.html.

```
runner.scheduleEvent(1, repast::Schedule::FunctorPtr(new repast::MethodFunctor<
        RepastHPCDemoModel> (this, &RepastHPCDemoModel::doSomething)));
```

The Repast code sometimes seems to be more complicated than necessary; this might be because it does not use C++11 features. The `scheduleEvent` function could also use a `std::function` as a parameter, so that the model implementer would have to write only:

```
runner.scheduleEvent(1, std::bind(&RepastHPCDemoModel::doSomething, this));
```

Also, the more flexible communication in Repast comes with the disadvantage that the model implementation is responsible for the code that creates the package with the agent information, which is sent to another process[3], while in Pandora this code is generated automatically, the model implementer must only notify the attributes that should be sent with a `// MpiBasicAttribute` comment and add the agent to the project compile configuration (`SConstruct`) file. On the other hand this means that in Pandora the implementer is limited to the attribute types that are supported by Pandora.

On a different level, Pandora is more flexible than Repast without adding complexity: "designed to deal with the varied needs of modellers and fill the gap between prototyping and advanced simulations" (Rubio-Campillo, 2014), it comes with a twin programming interface in C++ and Python, the latter being useful for easy prototyping.

At the current point of the project, there is an ongoing valuable discussion between the HPC and GSS groups about the flexibility of the code versus the parallel code efficiency. Obviously, easy access and hiding the complex layer of parallelization will attract a larger community of potential users, in particular, those users not familiar with parallelization. Furthermore, object oriented-programming allows for a better code structure and easier model development. Epecially for agent-based models, object-oriented code appears straightforward at the first look. However, efficient parallelization suffers from this type of code. Thus, one long-term challenge of the project is to come up with new solutions in the parallelization that do not limit the users in their programming styles. We believe that only such user-friendly HPC-codes will allow a sustainable use of HPC applications in the GSS community in the future.

[3] For details, see http://repast.sourceforge.net/hpc_tutorial/RepastHPC_Demo_01_Step_07.html and http://repast.sourceforge.net/hpc_tutorial/RepastHPC_Demo_01_Step_08.html. In short: even for a simple agent this involves 60 lines of code. It might, however, be possible to adjust the Pandora code generator for use in Repast.

2.4.6 Tests and propensity for errors

In our work with Pandora, we found some bugs and got the impression that it could at some points be developed in a more orderly fashion. For Repast, statements cannot be based on work with the framework here, but some points are nevertheless worth mentioning:

- In contrast to Pandora, there are no tests written for Repast.
- Repast has the tendency to be fault-prone by design, for example,
    - in the case that the implementor forgets to add a random seed into the configuration file, it can happen that the random generator constructs the same sequence of numbers for each process[4].
    - in the effects of certain routines, as the tutorial warns: "This sequence of four calls (balance, synchronizeAgentStatus, synchronizeProjectionInfo, and synchronizeAgentState) is subtle. The 'balance' method must be called before the 'synchronizeAgentStatus' method; these two are usually called one after the other. But you may find it useful to call 'synchronizeProjectionInfo' in a different place – e.g., before the 'play' method is called. In this simulation, 'synchronizeAgentStates' probably doesn't need to be called after 'synchronizeProjectionInfo'; however, there are some instances where 'synchronizeProjectionInfo' does not result in all non-local agents having updated copies of their local originals' states. Generally this has to do with network projections, not spatial projections, so this is probably not a concern here. However, it is a good idea to keep in mind that the synchronize routines do not necessarily leave the simulation in a complete and consistent state; only by using all of them together are you guaranteed to leave the simulation in a consistent state."

The test functionality in Pandora is quite extensive. Thus, the testing methods have been investigated (and where necessary corrected), so that the quality of future model developments can be surveyed consistently. With the addition of new features, accordingly new testing routines will be added as well.

2.4.7 I/O from the HPC perspective

In Repast, parameter files can be read by only one process and then sent to the other processes via MPI. Similarly, results (in NetCDF or csv format, see above) are written via MPI communication with a single process wich is responsible for creating and writing the file. Also, Repast includes some functions that use MPI's reduce functionality to aggregate data across processes and write the result to single files[5].

---

[4] http://repast.sourceforge.net/hpc_tutorial/RepastHPC_Demo_01_Step_04.html

[5] See http://repast.sourceforge.net/hpc_tutorial/RepastHPC_Demo_01_Step_17.html

In Pandora, all processes read the complete GEOtiff-file at the beginning, and each process writes its own output file. The output is written as HDF5 output files, which are very suitable for parallel writing routines, when it is done right. Thus, WP3 has already begun to address the challenge of optimizing Pandora's I/O.

Together with WP3, state-of-the-art functionalities like MPI-collective read/write routines have been implemented in Pandora without requiring major changes. In the next implementation phase, HDF5 will also be used for reading the input, which unifies the code and allows relying more on the efficient features of the MPI library.

2.4.8 Partitioning and scalability

A general remark is in order: Murphy (2014) states that a "simulation created in Repast HPC will never compete with a special-purpose application. Once a simulation's dynamics are established, a pure MPI version will be possible and will run faster (probably much faster)" and "However, the MPI version will almost certainly lack the flexibility of Repast HPC; it will be an Agent-Based Model, but it will not help with Agent-Based Model*ing*."

This remark seems transferable to any useful ABM framework and has implications for the pilot work in CoeGSS. Being pilot studies carried out to gather experience and develop tools for HPC-GSS applications, in the inevitable trade-off between generic solutions, or flexibility, and performance, flexibility always needs to play an essential role. That said, a few things can be stated about scalability of models implemented in Pandora or Repast.

- In our work with Pandora, especially with the GG pilot model, a challenge was identified in the distribution of agents to parallel processes: the partitioning of the model currently acts on the spatial domain via the underlying spatial layer that is evenly divided across processes. Using a world map processes with parts of the domain representing only water or lightly populated domains naturally caused load balancing problems.
- In Repast, the 2-dimensional space projection has a lot of similarities to the `SpacePartiton` in Pandora, in particular they share this problem of space being evenly distributed across processes and the resulting load balancing problems in our case, since also in Repast, the space projection is used to balance the agents across the process. No balancing/partition methods for graphs are implemented; in models with only graph structure, previous partitioning with external tools seems to have been applied.
- Acknowledging that "heterogeneity of problems to be modelled [in agent-based form] suggest that there is no optimal scheduling algorithm [to manage the order of execution of agent and environment updates]", Pandora allows to use any scheduler with any model and provides an interface for advanced users to develop their own schedulers if needed (Rubio-Campillo, 2014).

Scalability tests with the preliminary pilot model, implemented in Pandora, have been carried out (see Section 4.1.4). For scalability of Repast, some information is provided by two publications:

- "Large-Scale Agent-Based Modelling with Repast HPC: A Case Study in Parallelizing an Agent-Based Model" (Collier, Ozik and Macal, 2015)
- "Computational Social Science and High Performance Computing: A Case Study of a Simple Model at Large Scales" (Murphy, 2011)

Collier and colleagues parallelized an epidemiological model and simulate scenarios for Chicago. The resulting graph contains 1.2 million vertices and approx 1.9 million edges. METIS is used to partition the graph. The authors report a speedup of 1350% compared to the non-parallel version when running the simulation on 128 processes with 16 threads each[6], which implies a parallel efficiency of 0.007.

Murphy uses a simple simulation[7] for his scaling tests. The biggest simulation for this publication contains approx 68 billion agents distributed on 32768 processes. The results shown are really nice, but they are produced with an implementation “in which the code managing the exchange of agent information was written using low level C/C++ and directly invoking code from the Message Passing Interface (MPI) library that provides the cross process communication in the parallel HPC environment.” (p.3) Additionally, the paper mentions that the usage of Repast's File Output system[8] “slows the simulation performance down to unreasonable levels” (p.6), and therefore HDF5 was used instead. The paper does not mention any numbers for the Repast implementation without these two modifications.

2.4.9 Licence

Pandora’s licence is LGPL. Repast does not use a standard licence, but the licence seems to be permissive.

* Repast for High Performance Computing (Repast HPC)

* Copyright (c) 2010 Argonne National Laboratory

* All rights reserved.

* Redistribution and use in source and binary forms, with

[6] In one section, they mention an OpenMP implementation, so it is likely that the 16 threads per process are caused by a mixed implementation of MPI and OpenMP.

[7] Each agent selects two agents and moves between these two or behind one of those (in the line of the two agents).

[8] In the paper Murphy does not mention that the described output strategy is exactly what is implemented in Repast, so we assume here that the described initial attempt is using the functionality delivered by Repast.

2.4.10 Final requirements

For global systems studies, presently no HPC tools exist that immediately supply all desired functionalities or requirements for an all-purpose ABM platform (see also Rubio-Campillo, 2014):

- Rapid prototyping in a dynamic programming language
- An alternate interface providing access to an efficient version of the same functionality
- The users should be able to analyse their models with a wide range of analytical tools, including spatial analysis, statistics and visualization.
- Parallel execution, not only of different runs but also of a single, CPU-demanding, simulation. The switch between sequential and distributed executions should not translate into a re-implementation of the model or in learning the complexities of parallel programming.
- Representation of the spatial vicinity (basic feature of ABMs, see Epstein, 1999) and network structure by a general graph framework.

Having considered various aspects of Pandora and Repast HPC in comparison, Pandora will be our main choice for future simulations. The main reasons that led to this decision include: Pandora's elegant way of hiding the MPI complexity from the modeler, its small and readable codebase, and its Python interface for rapid prototyping. Choosing one common tool facilitates CoeGSS work regarding scalability, load balancing, data management etc. Enhancements to Pandora have to be made, but other enhancements would have to be made for other frameworks.

This does not, however, mean that we will restrict our attention to Pandora only. As CoeGSS should also be of use to customers who use other tools, like Repast or Flame, in their own

projects, it would be desirable to develop the related know-how. If a pilot has reasons to use another tool they should do so. However, since resources of CoeGSS are limited, there are not the same resources for enhancements of other tools.

2.5 Networks

The contagion of ideas and diffusion of habits is an important issue in addressing the research questions of all pilots. In order to model how, for example, healthier or "greener" behaviour percolates through a part of society, for example, a city, the pilot models need to take into account the interaction between agents. For an agent, three layers can be distinguished that contribute to decision-making:

- individual layer: agents decide, in part, based on the individual benefits they attribute to a decision. The agent evaluates these benefits based on the available information and on individual agent properties. This may be modelled as maximization of individual utility, as a propensity to keep doing what an agent was already doing, or in many other ways.
- social layer: agents interact or exchange information with other agents directly. The term "contact network" is mostly used for agents (as nodes of the network) and face-to-face meetings (as its edges); the term "social network" often refers to internet based virtual networks (e.g., facebook and twitter), where nodes are agents, or, more precisely, their accounts, and edges correspond to the types of links possible between agents (friendship, sending messages, "liking" posts etc.). There are further relations between agents, such as kinship, that may play a role. The term "agent network" shall designate a general network of agents with links between them, be these links face-to-face interaction, internet communication, or any other relation between the agents.
- environment layer: agents obtain information, or are subject to rules and norms from their environment. This may happen via the media, advertisement campaigns, the law, policies and regulation, etc. Further, agents' decisions may be influenced by conditions they find in their environment (e.g., the availability of charging stations may influence the decision to buy an electric car).

The networks between agents constitute an essential difference to models using representative agents, which are wide spread for example in economic modelling. There, all agents are assumed to be identical and can hence be aggregated into a single agent. With networks being an explicit and important component, models can no longer be reduced to a single representative agent.

In the literature, most cited features of physical agent networks (e.g., Granovetter, 1973) are:

- heavy tail in the degree distribution, which means that the number of connections of the nodes (agents) are not distributed according to a Gaussian distribution, but there are more nodes with extremely large degrees.
- high clustering coefficient, that is, a high number of closed patterns of edges like triangles. This characterizes resilient networks: a high presence of closed patterns means that cancellation of a single node rarely implies isolation of other nodes.
- assortativity among vertices, that is, nodes have a tendency to link to similar nodes.
- community structure, that is, the presence of sets of nodes much more linked to elements inside the same set than to the ones outside.
- hierarchical structure, that is the presence of subclusters (and subclusters of subclusters and so on) inside clusters.

These features have been noticed upon studying agent networks at different scales and should therefore also be found in the networks between the agents of the pilots' ABMs.

Given a set of virtual agents for use in a pilot's ABM, they need to be equipped with a network. The network can be considered part of the synthetic population, whether it is generated with the agents, or added in a post-processing step after the generation of the agents has still to be evaluated. To construct such a network, pilots will work together with the network experts in CoeGSS to choose, apply, and potentially develop further the most promising methods and techniques.

As only partial information is usually accessible, reconstructing networks from partial data is often necessary. The literature provides network generation and reconstruction techniques, developed for example for economic networks. In order to make unbiased predictions on the network to be reconstructed, an entropy-based method has been widely used. This method rests upon Shannon entropy maximization, while imposing some constraints. The latter represent the information one has access to, while everything else is random (for a theoretical overview, see Park and Newman, 2004).

Some networks can be satisfactorily reconstructed by using the information encoded into the degree sequence, i.e. the information on the number of neighbours of each node (Squartini, Fagiolo and Garlaschelli, 2011). Sometimes, there even is no need to know the whole degree sequence (i.e. N node-specific pieces of information) but only one aggregate, global quantity (the total number of connections (Garlaschelli and Loffredo, 2008); (Cimini, Squartini, Musmeci, *et al.*, 2015)). Further, there are models defined in terms of non-topological quantities; these are called "fitness models" (Cimini, Squartini, Garlaschelli, *et al.*, 2015); (Squartini, Caldarelli and Cimini, 2016); (Caldarelli *et al.*, 2002).

Such approaches are pretty general and can be implemented on social networks too: the methodology developed by (Saracco, Di Clemente, *et al.*, 2016) and (Gualdi *et al.*, 2016) is currently being applied to the social network of Facebook (Saracco, Cimini, *et al.*, 2016) by

the network experts in CoeGSS. In the bipartite network of posts (by argument) and users, there is a link if the selected user "liked" the chosen post. Using a projection algorithm, different communities of users can be sketched, based on commonly liked posts; the application of (Saracco, Cimini, *et al.*, 2016) is faster than (Saracco, Di Clemente, *et al.*, 2016) and (Gualdi *et al.*, 2016) and permits to infer communities of users with common interests. This allows to consider, in studying "contagion" of a community with a new idea, whether the idea is close to the interests shown by the community, and how an idea propagates inside a community as well as between communities.

In order to simulate a society, as is the aim of the CoeGSS pilots, one could adopt the model proposed by Granovetter (1973). Generally speaking, one should think of many dense communities linked by weak links holding them together. In more quantitative terms one can:

- impose a community structure consisting of many disconnected blocks that represent communities;
- such disconnected blocks should be quite dense in order to ensure that nodes within have a high clustering coefficient. Using a fitness model, one can create blocks with nodes having a power-law degree distribution;
- then one can add links among blocks, in such a way as to create a small-world network and account for the presence of individuals acting as hubs;
- according to the level of detail wanted, one can repeat such a procedure so as to create a hierarchical structure.

2.6 Literature study on autonomous driving

With the developments in autonomous driving, contemporary societies are moving towards another global "mobility revolution" which is expected to have profound effects in various dimensions including safety, convenience, ecology/fuel consumption, urban planning and urban structure. Accordingly, autonomous driving is closely related to the global challenges analysed by the two pilot studies, Green Growth and Global Urbanisation.

Within WP4 common work, a literature study on automous driving is being carried out. Automated vehicles (AVs) can be defined as vehicles, in which "[…] at least some of the safety critical control functions (e.g. steering, throttle, or braking) occur without direct driver input" (Wadud, MacKenzie and Leiby, 2016). AVs can be classified in relation to different levels of automation. They may drive partially or fully themselves and may ultimately require no driver at all. While the literature study addresses to some degree the different levels of automation, it is mostly concerned with the highest level of automation (full automation) where all decisions are taken by the system. The analytical focus is on the potential benefits, potential risks and individual and societal acceptance aspects of AVs.

The results of the literature study will provide information that can be useful for the two pilot studies Green Growth and Global Urbanization mainly at a later stage in the project

when the models and decision-making situations applied grow more complex and broader "background information" relating to agent decision-making on mobility behavior is relevant.

In the first project year, activities on the literature study have included collection of relevant studies and papers; definition of categories for the analysis of the studies and papers; analysis of the studies and papers according to these categories; and drafting of first parts of the study.

2.6.1 The literature study approach

For the search of relevant studies and papers the following databases were used:

- IEEE Xplore: http://ieeexplore.ieee.org/
- Science Direct: http://www.sciencedirect.com
- Springer Link: http://link.springer.com
- Wiley Online Library: http://onlinelibrary.wiley.com
- Taylor & Francis Online Library: http://www.tandfonline.com
- Deutsche Nationalbibliothek (German National Library): http://www.dnb.de

The search terms included: Autonomous Car/Vehicle/Driving, Autonomes Fahren, Autonomes Auto, Autonome Fahrzeuge, Neue Mobilitätskonzepte, Automatisiertes Fahren, Automated Vehicle/Car/Driving, Intellegent Car, Fully Automated/Connected Driving, Automating Automobiles, Automation Vehicle, self-driving, implication/impact, driver assistance, Automation in Driving, Computing Driving Car/Vehicle.

The search resulted in more than 90 papers and studies. These were analyzed with a system of categories including five main categories and 22 sub-categories. This system was developed in an iterative manner by recurrent loops of reading, (provisional) category building, and verification of categories. In an excel file, the studies and papers were assigned to the relevant (sub-)categories, mostly with a quote or short summary of the key arguments. A study or paper was assigned to a (sub-)category when it dealt with the topic at some length and analytical depth. The categories are listed in the following; the frequency of studies and papers dealing with the categories is shown in the tables.

- Current state of autonomous vehicles
- Potential benefits
    - (more) safety (e.g. reduction of crashes)
    - (enhanced) comfort
    - travel time (use of travel time for other purposes)
    - efficiency (e.g. efficiency gains at the system level – less congestion, less fuel consumption)
    - equal access (progress in equal access for the elderly, blind etc.)
    - costs
    - other

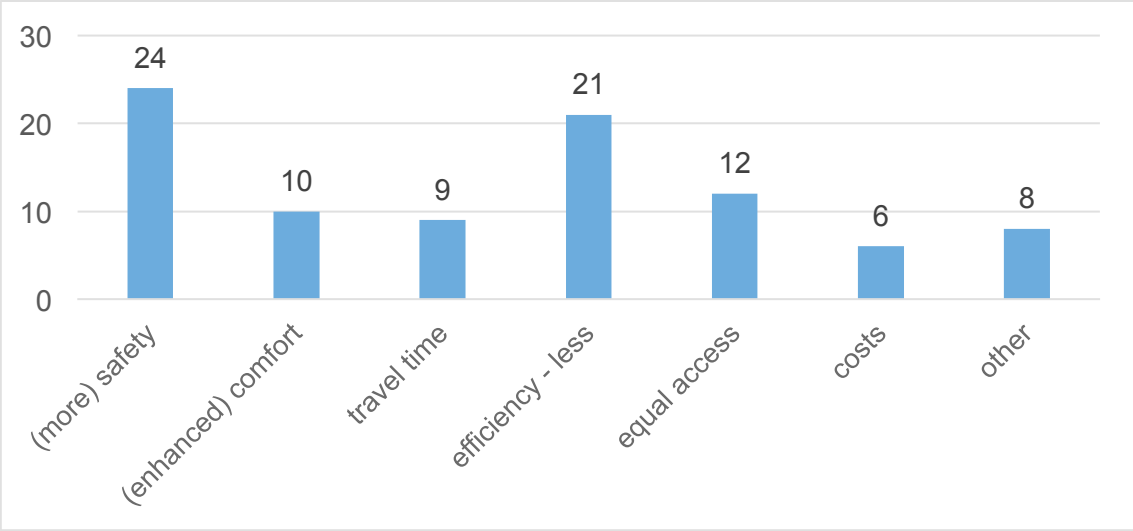

Figure 3: Frequency of studies/papers dealing with potential benefits

- **Potential risks** (categories draw on Grunwald 2016)
  - Accidents (i.e. safety: e.g. mixed systems of AVs and non-AVs)
  - Travel system (e.g. risks of complex technology and software together with unintended human behaviour might aggregate into unintended system problems)
  - Investments (e.g. Germany's economic dependence on automobile industry, or software errors and mass media scandalization (lack of return on investment))
  - Job market (e.g. taxi drivers, employees of logistic and delivery enterprises)
  - Equal access (e.g. higher costs, exclusion of less affluent)
  - Privacy (e.g. movement profiles)
  - Dependence (of societal mobility needs from system functioning; systemic risks, e.g. cyber war, and loss of competences, e.g. of driving)
  - Other

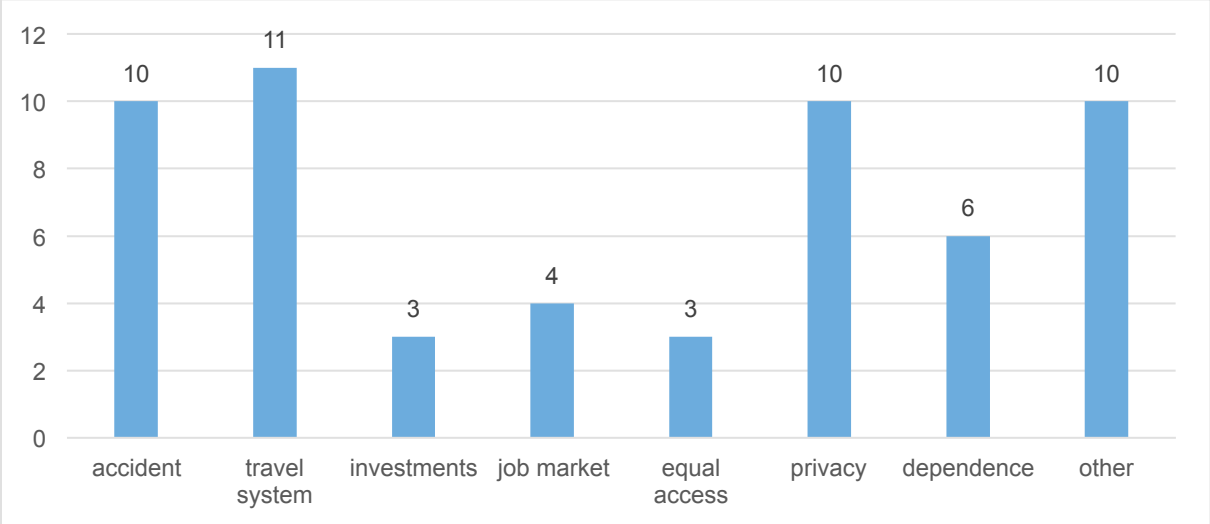

Figure 4: Frequencies of studies/papers dealing with potential risks

- Other potential impacts
  - Urban structure (e.g. density, mix of uses, urban design/planning)
  - Social impacts
- Barriers to implementation – open issues
  - Acceptance/perception (User and societal acceptance/perception - relevance of trust; behavioural models, users' adoption aspects)
  - Regulation (e.g. regulation/AV certification)
  - Liability (liability issues – increasingly liability of service providers or car producers)
  - Ethics (ethical/responsibility issues)
  - Others

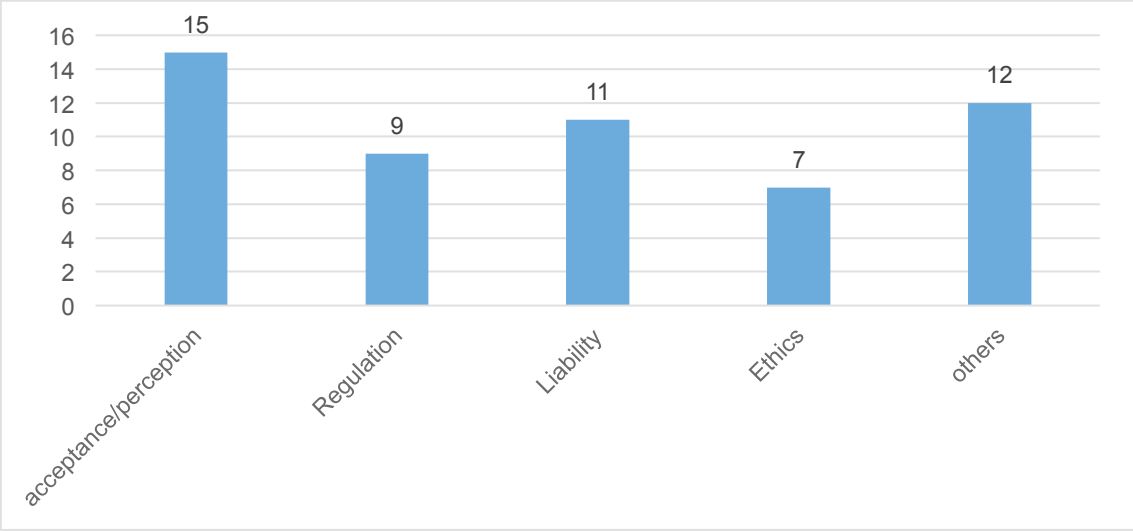

Figure 5: Frequencies of studies/papers deaing with barriers of implementation / open issues

2.6.2 Next steps

The next steps will be to provide an overview of potential benefits and risks and aspects of user and public acceptance as discussed in the literature. This overview will include conclusions from the studies and papers included on the possible role of AVs in a "Green Growth" scenario in terms of business and policy opportunities and conclusions on factors that may influence future individual and societal decision making on autonomous driving.

2.7 Outlook

While not assigned an extra task in the workpackage, common work at the intersection between pilots and in general support of the pilots will be carried on throughout the coming project years. This concerns the work on agent networks and the literature study, as described above. Other topics may be added as necessary, if they arise from the exchange between pilots.

Together with WP3 and WP5, supporting work required across pilots includes data pre-processing, simulation output analysis and visualisation tools, and enhancements to Pandora. Currently, work on using a graph-based structure instead of the spatial structure is ongoing involving WP3 and the pilots.

On the whole, the development of useful large-scale models of complex social systems and of efficient HPC software for these is not an easy task, yet possibly offers high benefits if we are successful. The flexibility that GSS applications require poses many new challenges for the straightforward efficiency thinking in the HPC world. Moving agents, dynamic changes of network structures and global connectivity will provide a uniquely different simulation structure of GSS applications, compared to any other HPC field. This flexible interaction structure of ABMs, in contrast to technical simulations, heavily complicates the use of common HPC approaches and negates the benefits of static load balancing approaches. Thus, only dynamic load balancing approaches will most likely succeed in GSS applications. Yet, a lot of solutions can be adapted from e.g., from molecular dynamic simulations and existing dynamic balancing approaches.

Together with WP3, the next challenge will be to identify the most powerful set of tools that exist and extend it by the missing essential features. In WP5, great effort will be required to transform the above used tools into services, that attract new users and enlarge the overlapping between GSS and HPC. Only by offering such a useful set of services, sustainable business concepts can successfully emerge in cooperation with WP2.

3 Status of the Health Habits pilot

The modelling of smoking habits and tobacco epidemics has been selected as the global system under investigation for the Health Habits pilot in Deliverable 4.1.

The top priority of this study is due to the devastating world-wide consequences of tobacco consumption, which is know to cause around 700.000 deaths every year in Europe alone, where at least 13 million people are suffering from smoking-related diseases. The impact of this particular class together with other diseases related to bad health habits (e.g. obesity, overeating, sexually transmitted diseases etc.) on the healthcare system is heavy as the estimated annual cost of just tobacco consumption to the European economy is of more than half a trillion euros (4.6% of EU GDP (Jarvis *et al.*, 2008)).

Given this burden on the welfare and healthcare systems, the development of a SIS able to monitor and forecast the spread of such behaviours is a top priority task. Indeed, the continuously increasing average age of the advanced countries' population may further increase the load on healthcare systems due to the growth in conditions associated to risk behaviours. On the positive side, recent projections indicate that "if most of the future gains in life expectancy are spent in good health and free of disability, this could offset more than a half of the projected increases in spending due to an ageing population" (Centola, 2011).

Addressing effectively the tobacco epidemic requires understanding and modelling how smoking behaviour is transmitted. This is a challenging task that has been tackled at the individual level in health psychology (Schwarzer, 2008) and with the development of ad-hoc mathematical models describing smoking dynamics (Sharomi and Gumel, 2008);(Levy, Bauer and Lee, 2006).

However, difficulties arise when one wants to measure and model how individuals of an interacting population update their behaviour. Research works on this subject have shown that the interactions in each individual's social networks regulate the spread of health habits in the general population (Przywara, 2010). In addition, one also has to consider the notion of complex and social contagion so as to correctly describe the dynamics of such behavioural changes (Christakis and Fowler, 2007). The investigation of such mechanisms and their framing in an HPC-compliant SIS architecture would considerably help policy makers and stakeholders in the planning and optimization of campaigns, allowing for a maximization of their effects on the population (e.g. by more effectively lowering the smoking initiation rate among youth).

Given the complexity of both the system and the dynamics under consideration the first aim of our work has been to review the existing literature of the field, inspecting all the so far developed models so as to distillate the relevant properties that an HPC compliant model implementation should feature. Such a SIS featuring an Agent-Based-Model is under development in close collaboration with the other partners. Here we present the preliminary specification of the model and the open issues found. In Section 3.1 we present the

background studies on smoking behaviour from a psychological point of view together with the relevant source of data to use on in the modelling task. Then in Section 3.2 we present the mathematical and computational models previously developed in the literature, listing their features and the open issues in the field. Then we also give a detailed description of the relevant mechanisms that can be framed in the decision-making process of the agents in the SIS under development. In Section 3.3 we then present the exploratory work done in the direction of developing a HPC compliant SIS. Finally, we present in Section 3.4 our outlook on the next steps and on the issues to tackle in the work to follow.

3.1 Problem definition and background

3.1.1 Smoking and health habits

Despite being long investigated, the modelling of tobacco-related habits remains a challenging task due to the complexity of the elements it involves. For instance, while smoking initiation or the adoption of un-healthy tobacco-related habits (such as pipe, tobacco chewing etc) can be roughly thought as an infection process, several differences are found between a classical epidemic model and the tobacco one. Firstly we are dealing with a so-called complex contagion process, where the medium of the contagion process still includes the physical social network (i.e. all the contacts that an individual experiences in the real-world) but also encompasses other layers of interactions as we show later in Section 3.2.3.

Second, while in epidemics processes the infection mechanism is independent of the individual opinion (meaning that an individual will get sick even if she does not want to) and people may tune only their behaviour by applying a more conservative attitude in the presence of a pandemic, in the tobacco case the decision process is far from being understood. The individual indeed ponders her choices based on a non-trivial decision process that takes into account social pressure, personal status and experience.

From the social-sciences point of view the process leading to initiation of, quitting from and relapse to negative health habits (e.g. smoking) has been modelled in different ways. From a psychological point of view two families of models are present, mainly continuum models and stage models (Schwarzer, 2008). While the models belonging to the first class project all the processes governing the adoption of a good health habit (e.g. doing regular physical excercise) in one prediction equation, in the second case the decision process of an individual is modelled through different stages describing the diverse level of adoption of a particular behaviour by the agent herself. For instance, in Figure 6 we show the two phases of Motivation and Volition of the Health Action Process Approach model. The latter introduces different interactions of social-cognitive traits. In the first phase of motivation a person usually develops the intention to act, based on a concern or a perceived risk. After the inclination towards a particular health behaviour is settled, the volitional phase starts. In the latter the “good intention” has to lead to the actual performing of the desired action, and also maintaining it once it has started. This is achieved through a non-trivial interplay of

self-regulatory skills and strategies together with other proximal factors that are not included in the picture.

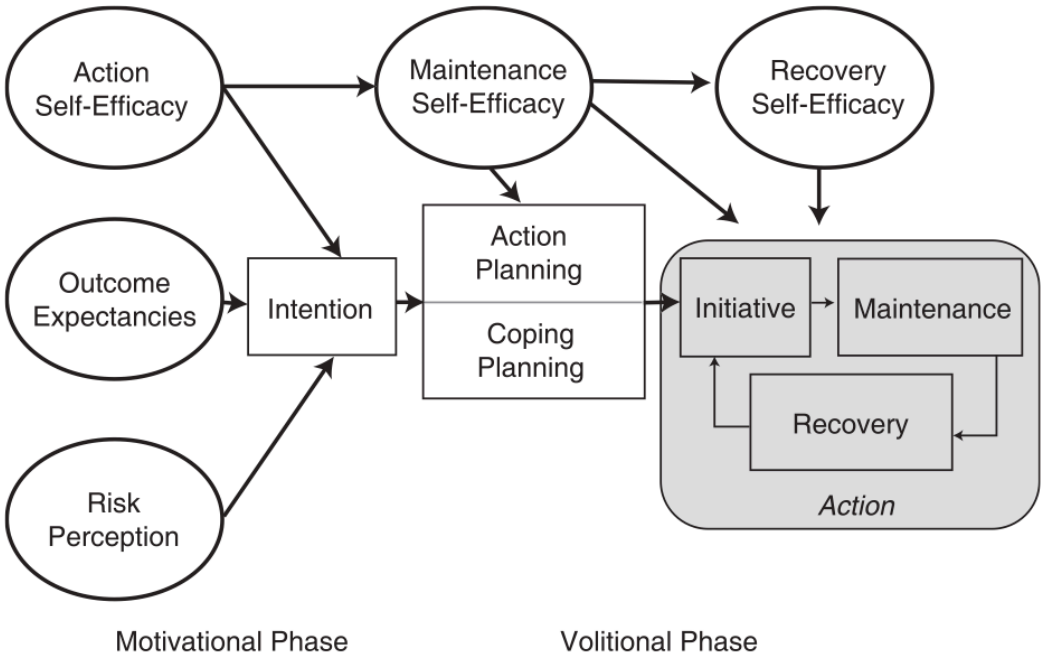

Figure 6: Generic diagram of a stages model of the Health Action Process Approach. Figure credit from (Schwarzer, 2008).

The main difference between continuum and stage models resides in the fact that in continuum models changing the order of the targets of the intervention approach toward an individual does not affect the outcome of such intervention. That is, the predicting equation will return the same evolution of an individual's status either if all the interventions (or policies) are applied at once or in a different temporal order. On the other hand, stage models support the intuition that different interventions are appropriate at different stages of health behavior change. The most popular stage model, i.e. The Transteorethical Model of Behavior Change (TTM) (DiClemente and Prochaska, 1982), proposes five discrete stages of health behavior adoption. These are defined as a function of an individual's past behavior and future intentions. The stages are precontemplation, contemplation, preparation, action and, maintenance. The path of an agent along these steps includes multiple attempts to progress from one stage to the next one as well as relapse that could in general occur at any time. Despite their more detailed modeling scheme, stage models are still hard to implement as the critical factors and processes letting people migrate from one stage to another still need to be clearly identified (Armitage and Arden, 2002).

However, the theory of behavioural change is an open and prolific field of research from which to borrow ideas and guiding principles in the selection and implementation of the mechanisms leading the evolution of the system under investigation, as we will show in Section 3.2.3.

3.1.2 Data

Given the importance of the subject, national health agencies of several countries provide detailed information about smoking prevalence as well as historical records and forecasts both on the health status of the population (e.g. fertility and mortality rates) as well as recorded adoption or cessation rates of smoking. These data sources are of central importance in the development of synthetic populations and in the definition of the population demography (see Section 3.2.4 for details) as they: i) allow for the calibration/fitting of the models under development against empirical data, ii) set the baseline of the simulations and, iii) serve as a reference in the generation process of the synthetic population.

3.1.2.1 Census and population

World wide data at a fairly detailed level on human population density are found in the UN-Adjusted Population Count and within the Population of the World version 3 (GPW v3) provided by the Center for International Earth Science Information Network (CIESIN) at the Earth Institute at Columbia University (Balk and Yetman, 2004).

3.1.2.2 Demography and population traits

As the usual time scales of the epidemics simulations for smoking span more than a decade, detailed information on the projected population and demography evolution are needed. To this end, the EuroStat agency provides data and forecasts on the population level for the EU as the UN does globally in their revision of World Population Prospects. Other sources of information on a more detailed level are from the national bureaus such as the United States Census Bureau. The latter gives estimates of the population status, its projection in future years, estimates and temporal projections of the birth rate and death rates (both cumulative and age-sex-ethnicity specific ones, see for example the US census projections at http://www.census.gov/population/projections/).

A more extensive study on this particular aspect can be found in the Lee-Carter mortality projection and following (Lee and Carter, 1992).

Regarding the population traits, detailed information on both the population structure and the distributions of the relevant traits of individuals are available mainly in the EU and US areas (as well in the most advanced countries of Asia such as South Korea and Japan).

EuroStat, through the CensusHub (https://ec.europa.eu/CensusHub2) gives a detailed review at a country level of:

- Households type and size as well as their structure;
- Employment rates and place of work;
- Education, marital status.

More information on the working teams size distributions, the schools size and the education level may be found through the PIRLS project (http://timssandpirls.bc.edu) and PISA reports from the OECD foundation (https://www.oecd.org/pisa/home/).

3.1.2.3 Smoking prevalence and longitudinal data

Data assessing the historical records of smoking prevalence and the causes leading to the adoption of the habits are available both at national level (with data aggregated on the whole population and cross sections on age, gender and ethnicity) and a detailed level thanks to local surveys and studies. While the latter gives the advantage of a more comprehensisve introspection on the data, they have the obvious downside of being smaller in size.

Nation-wide prevalence data (also featuring a historical track) are available at Euro-Stat and from the National Health Service of most of the developed countries (see for instance the US Bureau of the Census and the National Health Interview Survey (NHIS) by the National Center for Health Statistics of US (http://www.cdc.gov/nchs/nhis.htm)). Historical data covering the 1960-2016 years for all the OECD countries are also available (https://www.oecd.org/health/health-data.htm). Similar data are available through the International Smoking Statistic (Forey *et al.*, 2002).

Data are also available at a finer resolution, as is the case for the Publich Health service in UK (http://localhealth.org.uk/). From this source, we show in Figure 7 the smoking prevalence at the ward-level in the UK at the end of 2012 with data aggregated for all the population older than 15 years. In the inset of the figure we show a zoomed vision over the London area. As one can see, the situation is far from homogeneous, with non-urban areas featuring a generally higher prevalence level with respect to urban areas. The initial population status and the starting smoking prevalence of a Synthetic Population must then reproduce with accuracy this spatial distribution so as to generate valuable predictions.

At a more refined level, detailed surveys accounting for other traits of the population under investigation are available. These surveys are valuable not only for their level of detail but also because they usually provide a longitudinal resolution (in time) of the evolution of smoking habits in the selected cohort. One of the longest tracks found in our preliminary work is given by the German Socio-Economic Panel (SOEP, https://www.diw.de/en/soep), tracking the health behaviors and the socio-economical status of a selected cohort from 1984 to 2014. Similar studies are also present for other countries (for the US see for instance http://www.actuary.org/) and they allow for the estimation of the risk factors of initiation at different age ranges as we show in the next section.

Figure 7: The prevalence of smoking in the UK (London area zoomed in) at the end of 2012.

3.1.2.4 Relative risk and initiation rates

The quantification of the increment of the individuals' death rate due to smoking-related diseases is of central importance to correctly model the global system under investigation because it enables the correct estimation of the population demography (see for instance Thun and Myers, 1997). Other data covering this topic are provided by the U. S. Bureau of the Census, Population Division, reporting the middle series vital rate inputs, and by the Health promotion and disease prevention supplement of the U.S. Department of Health and Human Services (http://www.icpsr.umich.edu) and in research work (Ahmad, 2005).

To conclude, also the smoking initiation rates and the quitting and relapse ratios are under investigation, as they have to be explicitly set in the simulations. Regarding this point, age-dependent initiation rates are usually measured from longitudinal surveys covering a sample of the population in consecutive years (Levy, Cummings and Hyland, 2000);(Levy and Friend, 2000). These data are available from the national healt agencies (e.g. it is contained in the National Health Interview Survey NHIS in the US that reported a cessation rate of 0.21%, 2.15% and, 5.96% for the age groups of individuals having 18 to 30 years, 31 to 50, and 51 and older, respectively at the end of the year 2000).

3.2 Modeling

In this section we first present the more prominent models found in the literature that allow us to spot the features to implement in an HPC-compliant Agent Based Model. Then, in

Sections 3.2.3 and 3.2.4 we review in detail the mechanisms usually implemented so as to simulate the decision-making process of the individuals and the outcome of sensibilisation campaigns or restrictive policies as well. Though most of the so far proposed decision-making and behavioural mechanisms have been implemented in the literature, a modelling framework encompassing the relevant processes at once is still missing due to the complicated design of such models, especially when having in mind an HPC-compliant target.

3.2.1 System dynamics with compartments

The System Dynamics (SD) family includes models in which the population is divided in compartments, i.e. individuals are not represented independently one from the other but are rather grouped in sub-populations based on some of their traits (age, sex, income, health-habits etc.). The evolution of the systems is then implemented by means of differential equations governing the transition of each sub-population (stock) from one group to the other (e.g. smokers quitting the habit and moving to the former-smokers group etc.). These dynamical equations usually operate on the density of one stock of individuals with respect to the total population, so that the implementation of more than a few traits of the population gets nearly impossible due to the need to define a separate sub-population for each combination of these values. That is why, despite the advantages of such a formulation (e.g. analytical tractability), the Agent Based Modeling approach is usually preferred when a detailed description of the system and the possibility to aggregate results in a custom fashion are mandatory requirements.

3.2.1.1 Discrete Time Hazard Model

The adoption of smoking can be modeled as a Discrete Time Hazard Model. In this framework, the adoption of smoking is seen as a failure in the temporal evolution of the individual's behaviour (Göhlmann, 2007). This is modeled through the survival probability $S(t|x_i(t))$, i.e. the probability for an individual i whose features are encoded in the $x_i(t)$ vector not to have failed once (i.e. to have never adopted the negative health habit) up to time t. The survival probability is basically the Complementary Cumulative Distribution Function (CCDF) of the failure time distribution $f(t \mid x_i(t))$, i.e. $S(t|x_i(t)) = 1 - CDF[f(t|x_i(t))]$. The parameters of such a survival function are fit to empirical data describing smoking prevalence, where data has to be collected longitudinally, i.e. tracing the health habits of each individual in time so as to distinguish the causes of the first time adoption and the causes of relapse to smoking. The longitudinal nature of the dataset also allows for a dynamical characterization of the risk factors, as they may be encoded in a time-dependent vector $x_i(t)$ for each individual i. Furthermore, this approach can facilitate the individuation of the relevant environmental and individual's traits that lead to initiation or cessation through a detailed and thorough statistical analysis of the dataset, as in (Göhlmann, 2007). There, factors incrementing the initiation rate are reported: the smoking habits of parents during an individual's childhood and their marital status (divorced or not) together with the income seem to affect the smoking initiation rate of the individual, while some other features seem not to affect it (e.g. the labor market status or the parental

education). As we discuss later, other influencing factors on the initiation rates may come from price changes, availability of alternative products (such as electronic cigarettes), the introduction of restrictive policies on smoking permission and cigarette retail and public concern about health issues (Levy and Friend, 2000);(Lang, Abrams and Sterck, 2015);(Verzi *et al.*, 2012).

3.2.1.2 PRISM and Mendez-Warner models

Mendez and Warner initially modeled smoking prevalence starting with a population that could only cessate the smoking habit, and in which the initiation rate is accounted for by reproducing the percentage of smokers aged 18 years entering into the dynamics (Mendez, Warner and Courant, 1998). The model has later been extended using SD to account for both youth initiation rates and relative risks factors (see Section 3.2.3.1 for details) varying by sex, age, age at cessation, and years since quitting for smokers (Mendez, 2011). Though tracing different birth cohorts, the models do not include migration effects and smoking initiation occurs entirely at 18 years of age.

A more detailed model belonging to the SD family is the Prevention Impacts Simulation Model (PRISM). The latter allows for the evaluation of health care intervention strategies outcomes, accounting for health conditions related to cardiovascular disease (namely smoking habits), tracing the individuals' sex and age groups (introducing three of them). These traits of the population allow for a sex and age-dependent definition of health-related risk as in the previous case (Anderson *et al.*, 1991).

3.2.1.3 SimSmoke

SimSmoke is a project developed by David Levy and collaborators belonging to the SD family. Under the hood, it contains a demographic population dynamics model that regulates the flow of persons between population sub-groups (i.e. never, former, and current smokers), as well as demographics (birth and mortality). The rates by which individuals flow from one compartment to the other depend on age, sex, race/ethnicity as well as smoking status (Levy, Cummings and Hyland, 2000). The model has been gradually expanded to account for different policies such as restrictions on selling cigarettes to youths, fines increase, and clean indoor air laws. These effects are encoded by means of elasticity in the cigarette consumption and by a mechanism of perceived risk by smokers and retailers who do not comply with the applied regulations (see Section 3.2.3.3).

3.2.2 Agent based models for a SIS

Though relevant, SD models divide the population in compartments and they do not provide a detailed introspection of the system able to describe the situation or the outcome of a campaign in a customizable way. On the other hand, the modeling framework of Agent-Based-Models (ABMs) introduces some remarkable advantadges with respect to the System Dynamics (SD) approach. Indeed, in an ABM each agent is represented as a separate and autonomous entity, storing all of its traits and features at the agent-level. This allows to easily account for demographics, changes of behavior and, more importantly, it allows for a

detailed and highly customizable aggregation of the population a-posteriori. Indeed, as the demographic and behavioral information are stored at the agent level, results may be aggregated according to any relevant characteristic (e.g., age, geographical location of the agent, etc.) without changing the population structure and without the need to update it when switching to a different aggregation scheme. This can be done even in the presence of intricate demographics and overlapping characteristics of the population.

Of course, the tradeoff of such an approach resides in the heavy computational costs in terms of memory (a large number of agents has to be stored rather than information on few sub-populations), compute time (multiple simulations runs are needed to provide a confidence interval in the predicitions) and time to generate the underlying synthetic population that, in the presence of a large number of traits, can be extremely computationally demanding.

Notwithstanding these limitations, the benefits given by the detailed description of the population and by the possibility to easily implement both arbitrary interaction patterns between the agents and custom decision processes tailored on the agents' properties motivate the wide application of this framework in the literature. Here we present the more prominent work done in this direction, swiftly pointing out their feature and main mechanisms that will be presented in more detail in the next section.

3.2.2.1 Markov chains

In its simplest formulation an ABM can evolve according to a Markov Chain (MC) defined on the possible states of the agents. In other words, one has to define the possible states (or stages) that an agent may feature and the rates by which an agent may update its state from one stage to the other. A schematical representation of a MC is shown in Figure 8, where we indicated the three possible states, i.e. i) never smoker, ii) current smoker and, iii) former smoker. The agents enter the system either by birth or by immigration as non-smokers (unless a probability for an immigrant to be smoker is defined). The transition rates driving an agent from a state to another are shown as arrows and they may in principle depend both on the traits of the agent (e.g. age, sex or ethnicity) and her past/current behaviour (e.g. the relapse rate may depend on the years since quitting). Note that this framework can be applied not only to the behaviour of the agent but also to her health condition as we show in the lower part of Figure 8. Here, as in the above panel, transitions and additional states of the agent that may be implemented at a later stage are shown with dashed lines.

This approach has the advantage of being easily extendible with the addition of more possible states both for the behavioural status (e.g. by adding a competitive product such as electronic cigarettes alongside the traditional ones) and the health one (e.g. with more transitory steps between Good Health and dead) as we show with the dashed elements in Figure 8. This can be done at the relatively low cost of defining the rates leading from one agent state to the other (that may again be expressed as a function of different agent traits). The downside of this approach is that one has to encode the functioning of a potentially complicated transition mechanism in a simple rate, thus losing information on the described

dynamics. Moreover, the interactions between agents are difficult to define in the model formalism, so that other approaches are usually found.

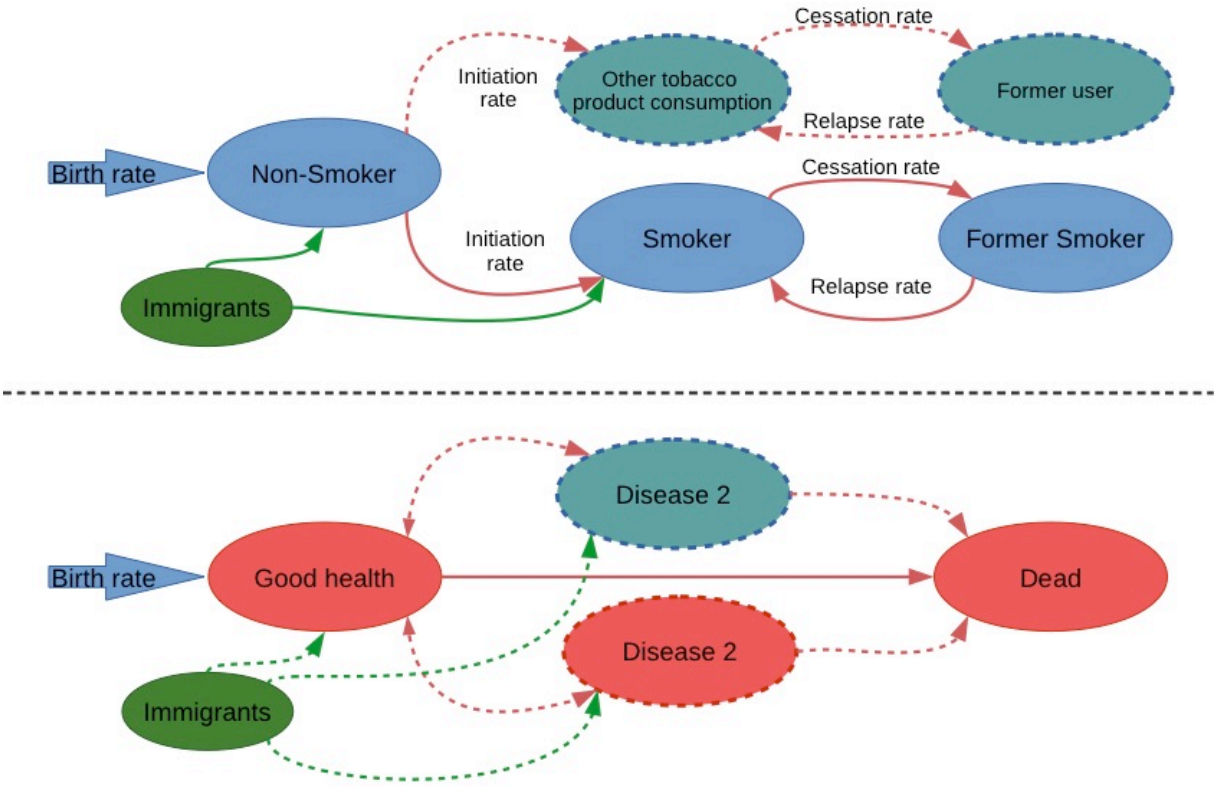

Figure 8: The representation of a Markov chain process for (top) behavioural states and (bottom) health sates.

3.2.2.2 BOID models and development of risk behaviours

In order to extend the MC approach, interaction between agents can be modelled by means of a network of contacts whose link (i.e. the edges connecting different agents) are the medium of the reciprocal influence of nodes. This influence can be thought of as a force driving the behaviour of one agent based on her observation of the system. For instance, an agent may update her behaviour after having observed the behaviour of the other nodes she has been in contact with. The task is now to define both the network of contacts (i.e. which nodes gets in contact with a particular node) and the updating rule of a node's behaviour.

In ABM disease epidemics, the network of contacts is usually encoded in the synthetic population structure (households, workplaces, schools, etc.) that specifies for the system individuals which agents they will contact while staying at home (i.e. the ones belonging to the individual's household), which are contacted in the working/school hours (depending on the age of the agent) and so on (Chao *et al.*, 2010); (Merler and Ajelli, 2010). When in contact, two agents update their health status depending on the respective health situation: an infected agent may infect a susceptible agent with a probability that depends on the agents' traits (Chao *et al.*, 2010); (Merler and Ajelli, 2010).

When dealing with complex contagions, the network of contacts (in its basic implementation) may still be put into practice by means of a synthetic population. On the other hand, the spreading of the disease is now complex and the reciprocal influence of agents in the system has to account for more details and different "infection" mechanisms with respect to the traditional epidemic processes (in which the infection depends only on the reciprocal health status of the agents, regardless of their willingness to get sick).

Though the ABM framework in developmental psychology is not widely used yet, some implementations have been proposed so far. The mechanisms usually included in such approaches are either the modelling of normative effects on individuals through the BOID architecture or the implementation of influence mechanisms in the evolution of both the contact network and the behaviour adopted by the users.

The first extension proposed deals with normative agents within the BOID architecture (Beheshti and Sukthankar, 2014). Here, BOID stands for Beliefs, Obligations, Intentions, and Desire, meaning that each agent decides her behaviour depending on a non-trivial interplay of both personal and social factors. Moreover, a *normative agent* is defined as an entity that demonstrates normative behaviour, i.e. that can recognize and reason about the norms it should comply with and, occasionally, violate them (Bicchieri, 2005);(Luck *et al.*, 2013). While we present the details of such architecture in the next section, let us note that the BOID architecture is the natural expansion of the classical BDI approach with the addition of the notion of obligations so as to account for social commitments, i.e. norms (Broersen *et al.*, 2001). Though this certainly is a relevant aspect of the modelling framework, we will present other ways to encode effects of norms to the agents in Section 3.2.3.3.

A different approach to the task can be done as in (Schuhmacher, Ballato and van Geert, 2014). In the latter the focus is put on the modelling of the implicit interdependence of contacts between agents (i.e. whom an agent is more likely to interact with) and their mutual influence in updating the behaviour (i.e. how much an agent is influenced by another when updating her behaviour). This is a relevant issue in complex contagion as the two mechanisms are expected to be deeply bounded one to the other. In other words, an agent is more likely to get in contact with agents having a similar behaviour (homophily). At the same time an agent may give a larger importance to the opinion (or to the behaviours to adopt) observed in people behaving in the same way (i.e. we get influenced by people similar to us rather than from different ones). The homophily and the influence depending on the relative difference of opinions are known to generate cliques (i.e. densely connected groups of nodes) in social networks. These are composed by individuals sharing very similar interests (segregation) and less likely to listen to ideas and opinions of other groups (echo-chambers) (Schuhmacher, Ballato and van Geert, 2014);(Castellano, Fortunato and Loreto, 2009).

3.2.3 Agent specification

3.2.3.1 Risk factors

The first risk factors to account for in an evolving ABM accounting for smoking related health habits and demographics are: i) the rate of initiation/cessation/relapse of smoking, ii) the birth/immigration and death rates. While the number n_b of birth individuals is usually computed as proportional to the number of women in fertility age times a reproductive factor, the number of immigrants has to be manually set based on census prediction. Moreover, while born agents are initialized as non-smokers, a pre-set fraction of immigrants may be initialized to smokers.

As the system evolves, time-dependent mortality rates have to be introduced so as to reproduce correct results (Verzi *et al.*, 2012). The overall, age-dependent mortality rate $Pr\{death\ at\ age\ a\ |never\ smoker\}$ has to be adjusted by multiplying the baseline value by a smoking relative risk so that, for example:

$$Pr\{death\ at\ age\ a\ |\ smoker\} = Pr\{death\ at\ age\ a\ |\ never\ smoker\} \cdot RR_{smoker},$$

where RR_{smoker} is the relative risk factor for smokers (Anderson *et al.*, 1991).

To conclude, the rates for smoking initiation, cessation and relapse also depend on the age and status of the smoker as we show in Figure 9 (where cohorts explain the highly discontinuous behaviour of the curves). These rates have to be either directly implemented in a MC realization of the model or accurately reproduced by the selected behavior evolution mechanism. Especially, let us note that the initiation process mainly takes place up to 25 years of age, thus denoting the importance of policies and sensibilization campaigns against smoking initiation targeted at youth people.

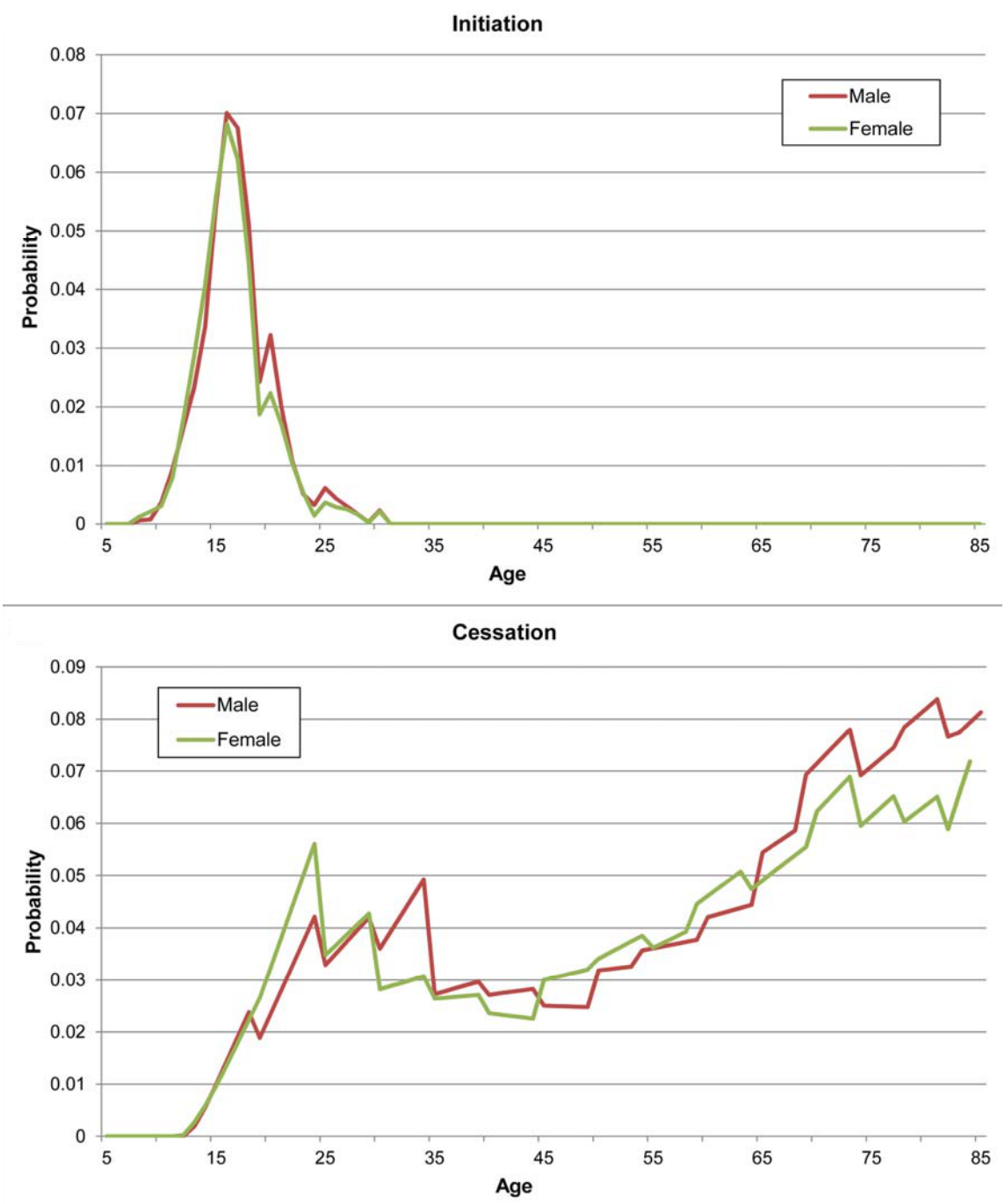

Figure 9: Initiation (top) and Cessation (bottom) rates by age and sex as found in different cohorts from NHIS data. Image from (Verzi et al., 2012).

3.2.3.2 Layers of interaction

As already stated, the agents in the system have to interact following a prescribed network of contacts. While the direct contacts of the agents with the individuals in their social circles are the most prominent source of influence for an individual, other layers of interactions (i.e. other kinds of influencing factors) have to be encoded in the model.

The first layer is the *Personal* one, including a set of personal values as introduced by Schwartz in his work on cultural value orientation (Schwartz, 2006). These are three bipolar

cultural dimensions measuring i) individualism, ii) achievement and, iii) equality, i.e. the propensity for an individual to i) get influenced by the others' opinion, ii) the priority of norm compliance (also enforcing it on the others) and, iii) consider others as equals or negligible. Alongside with these ingredients, in (Beheshti and Sukthankar, 2014) other *personal values* are added to the model, i.e. i) regret, ii) health concern and, iii) hedonism that describe the significance that an individual associates with i) being regretful to smoke, ii) being concerned with her health and, iii) seek for pleasure in the smoking act. Each personal value v_i is represented as a continuous variable belonging to the $[0, 1]$ interval.

Another layer of interaction is the *Social* layer. The latter helps in quantifying the effect of the local community (i.e. the individuals directly interacting with the agent) on the agent's future decisions. The definition of such network of contacts depends on the problem under consideration (it could be a network generated from a standard generative model for small systems or a synthetic population with household structure for simulations covering one or more nations). Another important decision regards the temporal nature of these connections, as they can be either static in time (Beheshti and Sukthankar, 2014), evolve with a dynamics generated by, for instance, daily commuting (Chao *et al.*, 2010);(Merler and Ajelli, 2010) or even varying at faster time scales, with interactions being continuously created and disrupted (Holme and Saramäki, 2012). We support the idea that given the long time scales of the process under examination, a first implementation of the model should feature a static representation of the synthetic population, thus neglecting dynamics on the contacts faster than the demographic characteristic time scales. In other words, we should account for households and population rearrangements (people moving out of home, immigration, deaths and so on) while ignoring daily commuting, travels etc. A further layer of non-local (in space) social interaction accounting for the contacts that an individual experiences in online social networks or with long-distance relatives/peers may then be introduced at a later stage of the model development.

To conclude on the social part, let us note that the intensity and duration of an interaction between two individuals may also be modulated in terms of their reciprocal behavior. For instance, in (Schuhmacher, Ballato and van Geert, 2014) the behavior B_{t+1}^i that the agent i adopts at time $t + 1$ can be expressed as a continuum variable falling in the $B_t^i \in [0, 1]$ interval, with higher values representing a more robust compliance to such a behavior (for instance, the attitude to smoke). The temporal evolution of the behavior can be expressed as

$$B_{t+1}^i = B_t^i + \sum_{j=1}^{n} \left[\alpha_t^{ij} \cdot \left(B_t^j - B_t^i\right)\right] \quad (1),$$

where the sum is over the n agents of the system and α_{ij} weights the influence that agent j exerts on agent i in trying to drift i's behavior toward its own. The weight of such interaction can be written as

$$\alpha_t^{ij} = c \cdot E_t^{ij} \quad (2),$$

where c is a parameter of the model and E_t^{ij} is defined as the *Evaluation* that agent i puts on the interaction with agent j (i.e. weighting how likely she is to update her behavior toward the one of j because of this interaction). We will introduce the definition of the evaluation E_t^{ij} in the next section within the framework of development risk behavior and social influence.

The last layer of interaction is the *Environmental* one. This can be thought of as a category of global and external factors that affect individuals' decision processes. These can be condensed as the following indicators: i) others, ii) advertisements, iii) cessation facilities and, iv) global awareness. These indicators globally account for i) the fraction of people adopting good habits in the system, ii) the frequency by which an agent is exposed to messages promoting good health behaviors, iii) the availability of cessation facilities such as nicotine replacement therapies and, iv) the global awareness about the health risks implied by tobacco consumption. These variables can be just a superimposed value measuring, for instance, the level of coverage of a sensibilization campaign or the amount of funds allocated to finance cessation facilities or they can be dynamically bounded to measurables extracted from real world data. One remarkable example of such a procedure is found in (Lang, Abrams and Sterck, 2015) where the smoking initiation rate at a given time t (and thus the smoking prevalence in a population) is determined considering not only the social pressure and the utility that the individuals find in smoking, but also accounting for the cumulative number of published research papers about tobacco-related health diseases up to time t.

3.2.3.3 Opinion formation and decision making

BOID

As pointed out before, a possible framework to work with in the modeling of behavior adoption is the one of BOID systems. In the latter, personal, social and environmental values are encoded in the system as countinuous variables whose value ranges from a lower to an upper bound (e.g. from 0 to 1). The personal values usually accounted for are individualism (*ind*), achievement (*ach*), regret (*rgt*), health concern (*htl*) and, hedonism (*hdn*). From the social point of view, a friendship (*frd*) indicator is defined based on the agent's network of contacts. This can be evaluated borrowing from game theory the notion of payoff matrices governing the diffusion process of the smoking behavior on the network. We report in Table 1 an example of such matrices, exhibiting the possible states of the node we are focusing on (rows) against the status of her neighbors (columns). There, α and β are two fixed positive parameters that enforce the preference for an agent to keep its current state (i.e. a smoker tends to continue smoking while a non-smoker is more likely to continue in the positive health habit). The value of the coefficients is based on the personal values of node i, e.g. $ss = ind' + ach' + hlt' + hdn$, and $sn = ind + ach + hlt + hdn'$ (where the primes denotes the complement of the measure $1 - value$). These values can be explained as follows: when two smokers meet, their payoff is positively correlated with their i) complement individualism ind' (as they are both observing someone acting like them and

the individualism measures the affection that the others' behavior has on the agent), ii) complement of achievement ach' (as the they do not want to stop smoking), iii) complement of health concern hlt' and, iv) hedonism hdn as they both find pleasure in smoking. One can then estimate the other coefficients by applying the same procedure.

Table 1 Payoff matrix of the diffusion process in the network

Nodes status	Smoker	Non-smoker
Smoker	$ss + \alpha$	sn
Non-Smoker	ns	$ss + \beta$

The overall frd friendship value for a node i is then evaluated using this payoff matrix by iterating over all the contacts of the node.

Regarding the interaction of the individual with the environment, this can be modeled with a Q-learning process (or equivalently using a neural network as presented later by the GCF pilot). In the latter, the state S of the agents (i.e. its personal and social traits) and the set of possible actions A that the agent can perform (namely whether to comply or not with a restriction rule and stop smoking) define a quantity per state-action combination $Q: S \times A \rightarrow \mathbb{R}$ that is updated every time the agent makes a choice (i.e. updates her state)

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \cdot [r_t(s_t) + \gamma \max_{S,A}(Q_{t+1}(s_{t+1}, a_{t+1})) - Q(s_t, a_t)],$$

where s_t and a_t are the state and the action of the agent at time t, $\alpha_t(s_t, a\_t)$ is a learning rate that in principle depends on time and action, r_t is the reward function and γ is a discount factor weighting the opportunism of the agent (the lower the more the agent will decide based on immediate rewards rather than long time rewards).

The reward function can be as complex as desired (and may also depend on the undertaken action a_t). However, in (Beheshti and Sukthankar, 2014) it is implemented as:

$$r_t(s_t) = \frac{(rgt + hlt - 2 \cdot hdn)}{2},$$

so that $0 \leq r_t(s_t) \leq 1$. This mechanism then regulates the interaction of the environment (identified by campaign, restriction policies etc.) with the agent (i.e. its *personal values*) and can drive both the individual sensibility to such campaigns or the individual propensity to smoke.

As pointed out before, the BOID modeling framework projects the current status of the agent (intended as the interplay of her personal, social and environmental values) into a smoking value sv defined as

$$sv = \frac{k_1 \cdot ind' + k_2 \cdot ach' + k_3 \cdot hlt + k_4 \cdot hdn' + k_5 \cdot rgt + k_6 \cdot env + k\_7 \cdot frd}{\sum_{i=1}^{n_{traits}} k_i},$$

where k_i is the weight of the i-th value and *env* accounts for all the global variables. As one can see, a higher sv value corresponds to a lower inclination toward smoking. One can then define two thresholds for sv, i.e. th_a and th_c, that set the adoption and compliance of the individual to good health habits, respectively. The behavior of an individual is then projected as: i) smoker aware of health and environmental restrictions if $sv < th_a$, ii) smoker who may temporarily quit smoking and adopt good health habits (e.g. that is contemplating the idea of quitting) if $th_a \leq sv < th_c$ and, iii) former-smoker who permanently complies with good health habits if $th_c \leq sv$.

Development of risk behaviors

Another modelling framework that can be applied is the one describing the development of risk behaviors. As already argumented in Section 3.2.3.2, the interactions between agents may be modulated by their respective behavior due to the homophily of human social networks. The task is then to encode in the model the mechanisms governing both the agents' selection of the alters to interact with (based on their perceived homophily) and their decision-making process in about their behavioral changes based on the former interactions.

A possible modeling framework is proposed in (Schuhmacher, Ballato and van Geert, 2014), where the behavior of the agents, how they perceive the others' behavior and their personal values all determine both the network of contacts and the spreading of the habits (may they be conventional or risk). While the behavior-updating rule has been shown above, here we present the details leading to the formation of the network of contacts. In particular, besides the actual k-th behavior $B_t^{k,j}$ of the agent i at time t (where the different behaviours may refer to attitude toward study/work, propensity to smoke/drink etc.) one introduces the agent j's k-th behavior as perceived by agent i as $B_t^{*k,ij} = B_t^{k,j} + \delta_t^{k,ij}$, where $\delta_t^{k,ij}$ is a distortion factor that accounts for the fact that agent i may perceive a different behavior of j with respect to her actual one. This distortion factor usually exponentially decays with time from the first interaction of the two agents, as the repeated interaction will eventually lead i to know the real behavior of j.

The next ingredient is the *similarity* S_t^{ij} (as well as *perceived similarity* $S_t^{*\,ij}$) computed as

$$S_t^{ij} = 1 - \frac{\sum_{k=1}^{n_t} \left| B_t^{k,i} - B_t^{k,j} \right|}{n_t},$$

where n_t is the number of traits (behaviors) under consideration and so that $S_t^{ij} \in [0,1]$. This quantity measures the similarity between the two nodes, by weighting their different propensity toward each of the n_t possible behaviors. The perceived similarity of i toward *j,* S_t^{*ij}, is computed by just substituting $B_t^{k,i}$ with the perceived j's behavior by i, i.e. $B_t^{*k,ij}$ and viceversa for S_t^{*ji}. The perceived similarity between two nodes influences the *preference* that a node has for another. The preference is evaluated as a logistic function with the

growth rate $g_t^{ij} = r \cdot (S_t^{*ij} - \theta)$, with θ being a superimposed threshold. The preference that i has toward j then evolves accordingly to:

$$P_{t+1}^{ij} = P_t^{ij} + g_t^{ij} P_t^{ij} \cdot \left(1 - P_t^{ij}\right).$$

The preference that i has toward j may not be symmetric, that is why the mutual preference is encoded in the *mutuality* $M_t^{ij} = \min\left(P_t^{ij}, P_t^{ji}\right)$. The mutuality weights the interaction frequency, i.e. the intensity of the interaction, between the two agents. For instance, in a dynamical representation of the network of contacts, the ij link is set as active at time t if a uniformly distributed random number $r \leq M_t^{ij}$. Moreover, the preference also sets the *popularity* Pop_t^i of a node that is expressed as

$$Pop_t^i = \frac{\sum_{j=1, j\neq i}^{N} P_t^{ji}}{N-1},$$

where the sum runs over all the N nodes of the system and so that the more an individual attracts the preferences of the others, the more he is popular.

Finally, agent i may consider the interaction with agent j either positively, negatively or neutrally. This is encoded in the *evaluation* E_t^{ij} that accounts for both the similarity and the influence that a node exherts on the other. Specifically,

$$E_t^{ij} = I_t^{ij} \cdot \left(w_v \cdot V_t^{ij} + w_{Pop} \cdot \mathcal{F}\left(Pop_t^j\right) + w_p \mathcal{M}\left(P_t^{ij}\right)\right),$$

where $I_t^{ij} = 1$ if the ij connection is active (0 otherwise), V_t^{ij} is a random variable that accounts for fluctuations in the evaluation of the contact by agent i, $\mathcal{F}(x), \mathcal{M}(x)$ are functions that measure the relative importance of j's popularity and the preference that i puts on her and w_i-s are parameters weighting the contributions of all these terms. In this way it is possible to weight the drift that each alter j puts on the i's behavior as reported in Equations 1-2 that we report here for clarity:

$$B_{t+1}^i = B_t^i + \sum_{j=1}^{n} \left[\alpha_t^{ij} \cdot \left(B_t^j - B_t^i\right)\right], \quad where \quad \alpha_t^{ij} = c \cdot E_t^{ij}.$$

Retail Compliance and price elasticity

Apart from the agent, other entities may influence the outcome of the tobacco epidemics. Amongst others, we found in the literature that retailer compliance to restrictive laws on selling tobacco to minors and price variations of tobacco are the ones usually under investigation.

The first mechanism can be thought of as a quantification of the easiness of access for youth to tobacco (which is a central point as the initiation usually happens before being 25 years old). When a restrictive law on tobacco retail gets into action, the compliance of clerks and retailers can be described within the Cobb-Douglas framework (also known as cost-utility function) (Levy and Friend, 2000). From this perspective, the *availability* of tobacco supplies

is expressed as a sum of different indicators measuring the easiness of access for each source

$$availability = a = OTC + theft + VM + OS,$$

where OTC stands for Over The Counter shops, $theft$ is a measure of the provisioning of tobacco from stolen stocks, VM is the availability of tobacco from vending machines and OS represents the contribution of other sources. This is done so that $availability \in [0,1]$, with 1 representing the total absence of policies and restrictions on the sources. The introduction of a policy restricting the retail of cigarettes to minors results in a perceived risk by the retailer who contemplates the idea of continuing to do so (thus violating the law). The objective risk is defined as the product between the probability of sanction and the level of punishment, measuring the number of outlets and retails checked by the authorities and the amount of fines introduced by the policy itself:

$$(obj.risk) = (prob\ of\ sanction) \cdot (level\ of\ punishment),$$

where all the variables fall in the [0, 1] interval. However, merchants may have a different concern about the new policies based on their experience or environmental variables. The objective risk can be further modulated by their concern so that

$$perc.risk = (merch.concern) \cdot (obj.concern),$$

with $(merch.concern) \in [0,2]$ as the merchants may ignore the restriction (limit of concern to 0) or be very sensitive to it due to social pressure and observation of other merchants respecting the rule (concern ~ 2). From these ingredients the $retail\ compliance$ can be defined as

$$retail\ compliance = (merch.concern) \cdot (checks\ per\ outlet)^{\alpha} \cdot (punishment)^{\beta},$$

where $\alpha, \beta < 1$ are two exponents account for the non-linearity between the number of compliance checks on the stores carried out or the amount of fines and the resulting increase in retail compliance.

Besides the compliance of resellers, also the availability drop for youth when new policies are applied can be modeled. For instance, the provision of tobacco from shops selling Behind The Counter (BTC, with $0 < BTC < 1$) and from Self Service machines ($0 < SS < 1$) may be modulated by a Self Service Ban variable ($0 < SSBan < 1$) accounting for the entity of such bans. The net effect of the policy and the merchant concern on the availability and easiness of acess for youths of tobacco from Over The Counter stores (OTCPol) reads

$$OTCPol = [BTC + SS \cdot (1 - SSBan)] \cdot (1 - retail\ compliance)^{\gamma},$$

with $\gamma > 1$ (Levy and Friend, 2000).

Alongside this approach, others have been proposed so as to model the impact of price changes and restriction policies on tobacco prevalence. For instance in (Ahmad, 2005) an age-dependent price elasticity is introduced, while in (Levy, Bauer and Lee, 2006) different kinds of policies have been implemented. As a final note, the outcome and performance of

campaigns are usually measured as years of life saved (or just lifes saved) by comparing the evolution of the system with and without the activation of polices. Other useful indicators are the Quality Adjusted Life Years (QALY, i.e. years of life saved weighted by the quality of life, where a weight of 0 corresponds to a complete disability and 1 indicating optimal quality) and the estimated balance between the savings in private and public health expenses and the nation's income variation due to drop of tobacco consumption and increment of its price (Ahmad, 2005).

3.2.4 Synthetic Population

The core element in the definition and deployment of an ABM is the definition of the synthetic population (SP). The latter is a virtual object representing with good accuracy the general traits of a population. In other words, when inquired about a particular distribution of a given feature or when evaluating the average value of an observable, the synthetic population should return a result in good agreement with empirical data.

In practice, SPs may be either unstructured or structured (or ungrouped and grouped). While the former consist of a system composed by a number of nodes that do not have particular relations bonding one to the other (and thus a network of contacts may be arbitrarily defined), the latter generally present the typical social structure of nations, i.e. individuals (agents) are divided into households, communitites/cells, neighborhoods, cities etc in a hierarchical fashion. This grouping is usually done accordingly to either population count (i.e. composing cells having more or less a constant number of agents) or covered area (with a cell being defined as all the agents lying in a given area of a given extension, usually some tens of square kilometers). In this project we will make use of structured SPs.

Within a SP, each agent or entity of the hierarchical representation may store information on personal traits, spatial location, the number of agents per building block and so on. This structured view of the SP must closely reproduce empirical data on household size distribution, average age of households etc. Moreover, other entities accounting for places where people usually meet and interact can be introduced, for example schools and workplaces. Also in this case, data on the schools size and their spatial distribution, as well as workplaces size and working-age distributions per kind of business are needed.

Once the structural organization of the population is set, the agents' mobility, interactions as well as the intensity of such interaction have to be defined. Mobility is usually included either by mimicking real-world data on commuting or by gravity models (Merler and Ajelli, 2010);(Chao *et al.*, 2010). Moreover, since the SP is usually implemented to model traditional epidemics processes, social contacts depending on the agent working/studying status and the community/cell of pertinence are defined. In particular, the spreading rates by wich an agent infects another depends on their belonging household (people in the same household may infect one to the other more easily that people living in the same geographical cell as they experience more frequent and lasting contacts), the relative age and the meeting place (spreading in schools from children to adult teachers usually features

higher rate than from an adult to the other in a workplace, etc.). Whithin the health habits pilot we will initially module the SP as composed by household and communities/cells, thus neglecting commuting, travelling and other fast dynamics that can be introduced later when suitable GSS problems will be tackled. Indeed, the typical time scales of such population dynamics is much smaller than the average time required for smoking behavioral changes to take place.

In order to develop such SIS, high quality, up to date data are obviously needed. Sources of such data are the Eurostat and National Statistical Offices. From their databases it is possible to retrieve most of the relevant information and distribution for the building process of SP. More data on schools and workplaces can be found from the Census, PIRLS and PISA projects.

The generative mechanism of a Synthetic Population (SP) can be found in (Merler and Ajelli, 2010);(Chao *et al.*, 2010) and usually comprehends a computationally expensive multidimensional optimization or an iterative procedure to sample empirical distributions and recreate their traits in the virtual replica of the population. Let us stress that the definition of a correct and significant SP is a fundamental ingredient for mainly two reasons: i) it provides a realistic ground for the modeling of the epidemics/behavioral change process by accurately reproducing real world data, and ii) it is the entity storing the relevant traits of agents as well as their status and it then is the one and only object from which to retrieve data, projections and relevant distibutions once the simulations has been done.

3.3 Exploratory work on HPC-compliant SIS

The pilot's preliminary activities included an exploration of different methods to implement parallel computing in GSS problems tackled by using SIS. To this end, an already existing code for the stochastic modelling of worldwide pandemics (the GLEAM simulator) has been used so as to get an insight about the potential advantages and drawbacks of each parallelization approach.

A traditional approach to code parallelization in the HPC environments is represented by the Message Passing Interface (MPI) standard. Indeed, this is a portable message-passing system already efficiently implemented in many parallel computer architectures. However, this approach requires special attention to minimize the overall amount of inter-process communication. Another important point is the need to ensure a proper synchronization of the different processes (e.g., make sure that the sequences obtained using pseudo-random number generators are statistically independent). Moreover, a critical performance bottleneck is represented by I/O operations on parallel-distributed file systems (such as Lustre) usually found on HPC architectures: to take full advantage of them the I/O subsystem of a program must be re-designed in order to reduce metadata requests and possibly use only collective MPI-IO procedures to access data on disks.

Another possible approach to parallel computing, especially suited for shared-memory architectures, is given by threads. As threads share process state (in particular, memory and files), they provide a simpler and faster way of communication without the overhead related to message passing from one process to the other. However, these advantages come at a price as threads also require to pay great attention to data synchronization to avoid deadlocks or race conditions. On the other hand, threads usually allow considerably reducing the memory footprint with respect to multi-process parallelism. This, in turn, can also lower the need for costly I/O operations on disks by making more main memory available to the multi-threaded process.

From a practical point of view, an interesting way to overcome (to some extent) the difficulties related to directly dealing with threads is represented by the concept of futures, for which many computer languages provide support.

3.4 Outlook

In the presented ABM modeling framework we define an agent as a collection of diverse features describing different traits of an individual, such as age, income and education level. The agent also stores information regarding her attitude toward the health habit under investigation, e.g. its opinion regarding topics that may influence its choice of behavior. At this stage one could possibly set as personal values the individualism, health concern and the achievement insicators (i.e. the propensity for the agent to achieve a good health habit and/or enforce it on the others). These traits may be mainly represented as floating point numbers that can then be read by the evolution model to evaluate the agent's status. For instance, all of these indicators may be put together to compute the smoking value and then decide whether or not the agents is a smoker by comparing it to a certain threshold. Moreover, given the intrinsic dynamical nature of individuals' practices, all of these properties generally evolve in time accordingly to both the underlying behavior evolution model and the demographics.

Once the agents are defined, their arrangement in a SP has to be defined. As the SP models the social structure of the agents in the system as well as its demographics, the spatial-proximity network of people and their hierarchical interaction have to be correctly tuned to the problem under investigation. For instance, in the smoking epidemics it is important to catch the correct households' size and age structure (as parental smoking is known to affect the initiation rate of youth) as well as the schools size and classes arrangements (as smoking initiation usually happens before the 25^{th} year of life). Agents may then interact at the household, cell/community, school and working place levels. Once this first level of social structure representation will be implemented, the SP may be extended by allowing for non-local, long distance contacts between agents interacting on on-line social networks or similar sources of long-range social proximity. However, this will require the collection of specific datasets that are so far publicly unavailable. Specifically, these datasets should provide the

topological properties of the social graph to be implemented in the system, in the same way as household data are needed to generate a realistic social structure.

Regarding the system's dynamics, at a first stage human mobility at fast time scales can be neglected, as the only relevant mechanism of agents' redistribution is household recombination (due to young agents moving out of home and creating a new household).

A required step in the aim of facilitating the creation of suitable SPs for each studied GSS problem is also the creation of an optimized pipeline leading from raw data to the computer representation of the model. To this end, besides the technicalities involved in the synthetic population generative process, data need to be pre-processed and cast together in a single and common format that can be passed as an input to the synthetic population generator. The latter, when provided with the distributions of the features that determine the adoption or cessation of a certain habit, builds the synthetic population with a tunable level of abstraction and details.

Finally, the rules governing the evolution of the agents have to be defined. From a technical point of view the implementation of the model should be done outside of the agent's scope, so as to keep a higher level of code generalization and to easily avoid race conditions and synchronization issues (i.e. avoid methods like Agent.evolve() and delegate the task either to the synthetic population or to a specific module acting as the engine driving the system evolution). In this way the dependency tree of the different modules composing the system is neater, as every part of the system does its task without interfering with the others. For instance, the agents store their updated status and pass it to the updating component.

As for the modeling framework to adopt for the description of the behavior update and evolution, an internal discussion is ongoing. At a first stage an implementation of a hybrid model borrowing ideas and concepts both from the BOID architecture and game-theory may be implemented, leaving for future work the encoding of different mechanisms as they get available in the literature.

This leaves room for a fruitful discussion within the CoeGSS consortium so as to find the optimal tradeoff between the theoretical model details, the complexity of the social interactions and an HPC compliant implementation of the developed architecture.

4 Status of the Green Growth pilot

As described in Deliverable D4.1, the global car population has been chosen as the global system under study for the Green Growth pilot. The mobility sector is one of the few economic sectors that does not show a reduction of CO_2-emissions in comparison to 1990.

A SIS for studying how the global car population, which currently counts over a billion members (Voelcker, 2014), may evolve and what this means in terms of CO_2 emissions is under construction. It shall become a tool to help structured thinking about options and strategies available to different players in the field – from policy makers, who might regulate CO_2 emissions, pass rules for self-driving vehicles, or make investments into (charging) infrastructures, via car manufacturers who may devise different business models based on marginal improvements to internal combustion engines or the shift to hybrids and electric engines, to households deciding between ownership of and access to a car, to name some examples.

Starting this work, an explicit aim was to first develop the simplest possible version of an HPC-based SIS, in order to quickly start learning how to use supercomputers for tackling GSS challenges. Therefore, a preliminary “green car diffusion” model was first specified and implemented for testing in the HPC environments in the HPC centers in Stuttgart and Poznan. This very first SIS version, based on this model, and what was learned by building and testing it, is described in Section 4.1.

The diffusion process for green cars represented in the preliminary model is to be understood as resulting from an underlying diffusion process of beliefs, values and infrastructural conditions. For some problems, such implicit modelling can be quite useful. However, we have designed the first model as a stepping-stone towards explicit modelling of optimising agents interacting in networks and making choice under limited information. We are currently working on a SIS based on an agent-based model. Section 4.2 presents the current status of the work together with challenges identified on the way. Section 4.3 then gives an outlook on the next steps of work to follow.

4.1 Preliminary SIS

4.1.1 Conceptual Model

A basic conceptual model has been developed beginning with just two classes of cars: “brown” (e.g. internal combustion engine) and “green" (e.g., hybrid or electric) cars that produce more or less emissions, respectively. Cars are distributed on a gridded global map, that is, in the first instance “agents” are simply cells in the grid, which automatically specifies a neighbourhood network between them, and their characteristics are a number of brown and one of green cars. Taking numbers of total cars as given by data, we first consider a simple diffusion dynamics for green cars. Each cell at each time step (currently a month) computes the number of new cars to be added from total numbers and a percentage of cars

being scrapped. The decision how many of the new cars are green is taken (in deterministic or stochastic manner) based on an innovation component, influenced by GDP data for the cell's country, and an imitation component, that is, the fraction of green cars in the neigbourhood (a weighted average of the cell itself and surrounding cells).

4.1.2 Input data and "synthetic population"

Data from several sources have been collected, pre-processed, and integrated (creating a rudimentary synthetic population) into a gridded global map.

4.1.2.1 UN-Adjusted Population Count

The UN-adjusted population count grids consist of estimates of the number of persons per 30 arc-second (~ 1 km) grid cell and adjusted to match United Nations country totals (Socioeconomic Data and Applications Center (SEDAC), 2016). This data gives the general spatial grid on which all other data sources are mapped. Thus, the maximal resolution, for which data is available on a global scale, consists of 43200 x 17400 = 751 Mio. cells. For the initial simulations, a reduced resolution of 8640 x 3480 is used.

The data set provides population estimates till 2020. For simulations beyond 2020, scenarios for the population growth with yearly growth rates are also available. These projections can be applied to the spatial structure to imply the global future trend, yet conserve the local structure.

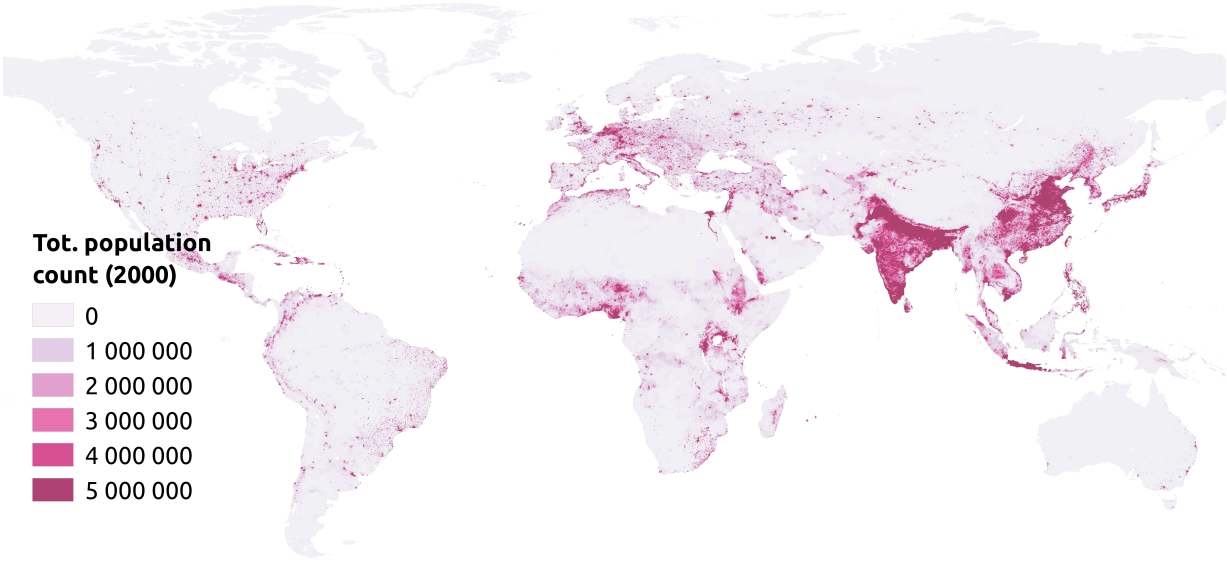

Figure 10: Population count of the world in the year 2000

4.1.2.2 Cars per 1000 inhabitants

The indicator 'cars per 1000 people' (OICA, 2015) is the most important one to evaluate a country's contribution to the global fleet of cars. Country specific data can be mapped to the spatial grid via country indentifiers. In combination with the inhabitants per cell, the car data allows to estimate the total number of cars in a spatial cell, see Figure 11. The procedure has been applied to data for total car numbers for the years 2005 to 2020 and provides the

dynamics in the total car population. These dynamics in combination with a certain replacement rate serves as an external input to the dynamic model. Figure 11 visualizes the fact that mainly North America, Europe and Asia contribute to the global car fleet.

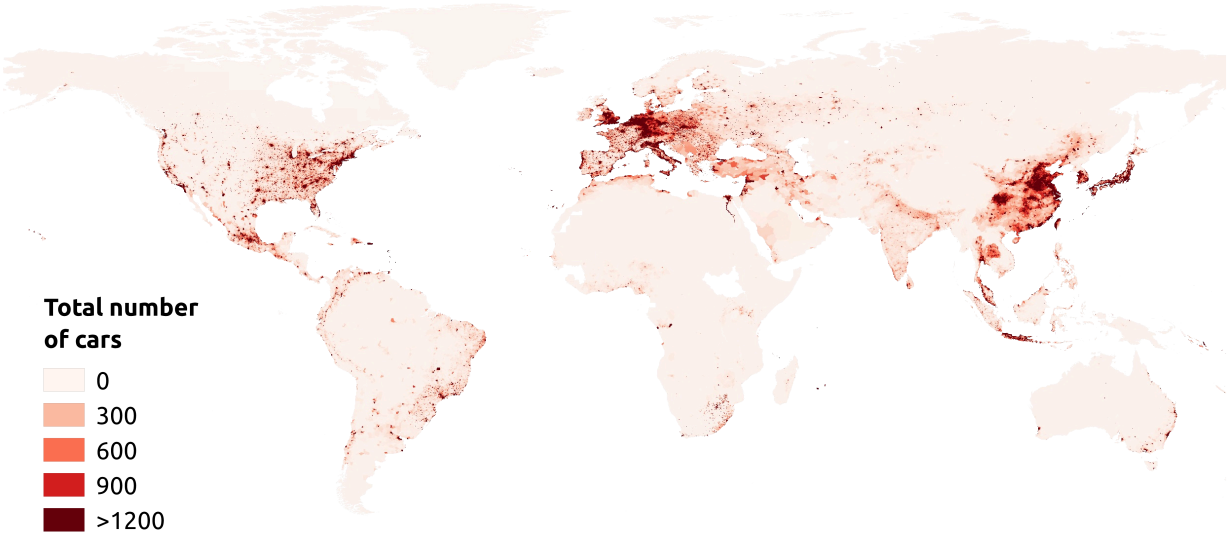

Figure 11: Distribution of cars in the year 2000

4.1.2.3 GDP per capita

The GDP data per country is provided by the World Bank (2015) and is available for most countries in recent years. Thus, the data can be mapped to the simulation grid via the country identifyer, see Figure 12. In the preliminary model, the GDP and its difference between regions serve as an indicator for the purchasing power of the population.

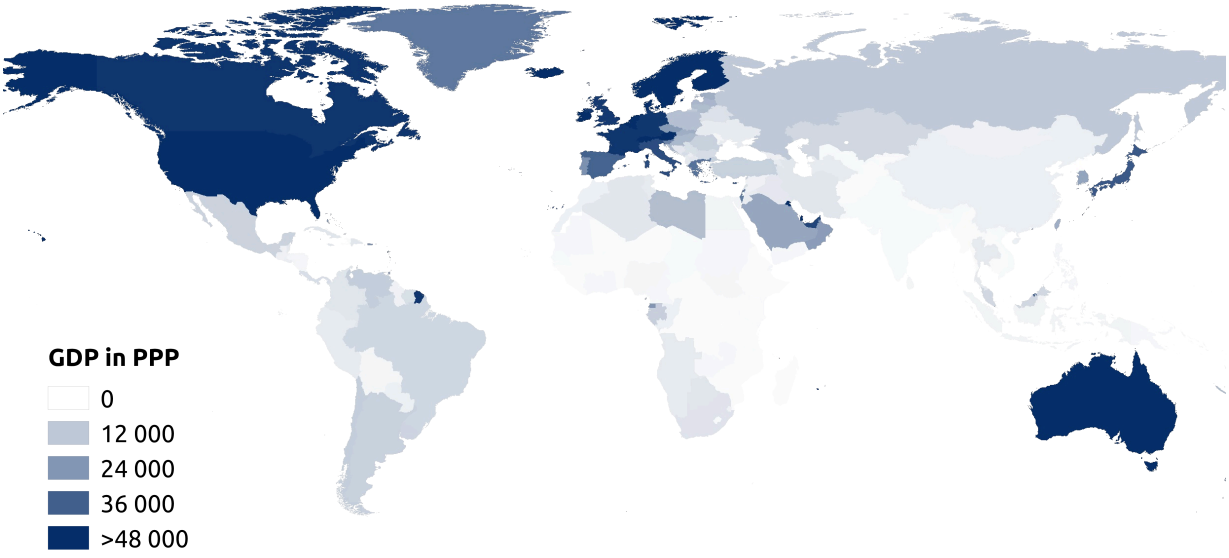

Figure 12: GDP per capita in Purchasing power parity (PPP)

4.1.3 Dynamic model

The basic dynamic HPC model has been implemented using the HPC-ABM framework Pandora (Rubio-Campillo, 2014) for the reasons given in Section 2.4. There is a deterministic and a stoachstic implementation of the preliminary SIS. Both calculate for each time step for

every cell the share of newly bought green cars $s_{x,y}^{\text{new}}$, with $0 \leq s_{x,y}^{\text{new}} \leq 1$. The difference is that in the deterministic version the absolute number of green cars bought in the cell is calculated by multiplying this share with the total number of bought cars, while in the stochastic version each car bought in the cell is a green car with the probability $s_{x,y}^{\text{new}}$.

Apart from the spatial input data, GDP per capita, population count and cars per 1000 people, the models require two additional parameters:

- η determines the share of "innovators" that buy green cars independently from the observed neighbourhood. To reflect the purchasing power of different countries and the higher costs of green cars, the share of innovators is modified based on the country's GDP per capita.
- κ determines the resistance to change in "imitators". For $\kappa = 1$ the imitators' $s_{x,y}^{\text{new}}$ would be exactly the share of observed green cars in the neighbourhood. For $\kappa > 1$ the imitators are sceptical about the new technology and $s_{x,y}^{\text{new}}$ will be lower than the share of observed green cars.

The calculation for the number of new green cars at a single time step depends on the share of existing green cars s^{exist} in an extended Moore-Neighbourhood with a Chebyshev distance of r. The influence $w_{x,y}$ of a cell in the neighbourhood of the cell at location $\hat{x}, \hat{y}$ is reduced by it's Euclidean distance, with $w_{\hat{x},\hat{y}} = 1$ and $w_{x,y} = 0$ for cells outside the Moore-Neighbourhood. Using those influence weights the "visible" share of green cars $s_{\hat{x},\hat{y}}^{\text{vis}}$ and the value $s_{\hat{x},\hat{y}}^{\text{new}}$ is calculated by:

$$
\begin{aligned}
s_{\hat{x},\hat{y}}^{\text{vis}} &= \frac{\sum s_{x,y}^{\text{exist}} \cdot w_{x,y}}{\sum w_{x,y}} \\
s_{\hat{x},\hat{y}}^{\text{new}} &= \eta G + (1 - \eta G)(s_{\hat{x},\hat{y}}^{\text{vis}})^{\kappa} \\
\text{with: } G &= \frac{\text{GDP per Capita}_{\text{country of location } \hat{x},\hat{y}}}{\text{GDP per Capita}_{\text{world average}}}
\end{aligned}
\tag{1}
$$

The total number of new cars per cell $n_{x,y}$ is exogenously given and derived from the data described in Section 4.1.2. The stochastic version draws $n_{x,y}$ random numbers between 0 and 1, and adds a green car for each number smaller than $s_{x,y}^{\text{new}}$.

In the deterministic implementation with the number of new green cars per cell we have the problem that the calculated number of new cars is a real number but the number of green cars an integer. The solution used was to add the fractional portion $r_{x,y}$ to the value calculated in the next step:

$$
n_{x,y}^{\text{green}} = \left\lfloor n_{x,y} \cdot \min(s_{x,y}^{\text{new}}, 1) + r_{x,y} \right\rfloor \tag{2}
$$

4.1.4 Simulations

First simulation experience has been gathered running the model on local computers as well as on supercomputers at the HPC centres involved in the project: HLRS Stuttgart and PSNC

Poznan. First scalability tests have been carried out. Some tests used the examples that are part of Pandora, additional tests used the preliminary Green Growth pilot model. Also first adaptations of Pandora to the needs of CoeGSS have been made.

All simulations discussed in this section use the input data described in Section 4.1.2 with the resolution of 8640 x 3480 cells for the years 2005-2015. About 8 Mio. cells contain land mass, for each of those cells an agent is constructed. The cells are distributed to the processes via a rectangular space partition with partitions of equal size. Due to the landmass distribution, this approach implies a huge variation in the number of agents per process.

Beside the MPI parallelization via the spacial domain decomposition, Pandora allows has also a second parallelization strategies, using OpenMP. This strategy parallelizes the iteration over the agents to call their updateKnowledge and selectAction methods. It is also possible two mix both strategies. In the current version of the dynamic model both methods are lightweight, and some tests confirmed, that currently the best strategy is to disable the OpenMP parallelization.

The following tables show the absolute computation time and the efficiency of the MPI parallelization strategy for different number of processes. To check the influence of the file I/O, we also created a version where the file output was reduced by writing a reduced set of raster files and no agent files at all:[9]

Table 2 Computation time for different MPI configurations

(in seconds)	4 processes	16 processes	36 processes	144 processes
Eagle	1366.3	789.1	645.1	165.2
Hazelhen	832.8	372.6	213.4	90.7
Hazelhen (minimal output)	50.3	29.5	18.0	10.1

Table 3 Efficiency of current decomposition strategy

Efficiency	16 processes	36 processes	144 processes
Eagle (compared to MPI/4 procs)	0.43	0.24	0.23
Hazelhen (compared to MPI/4 procs)	0.56	0.43	0.26
Hazelhen (minimal output, vs. 4 procs)	0.43	0.31	0.14

[9] See (Rubio-Campillo, 2014) for details about the scheduler and file output.

As can be seen by the difference of the two cases with results from Hazelhen, most time is spent for the file output, which is not that surprising considering the simple calculations that are done at each timestep for each cell.

Additionally, we compared the calculation time for the first step of the different processes, to check the imbalance caused by the spatial distribution where we can have some processes that contain only water and have nothing to do. The following numbers are for the Hazelhen/full output case:

Table 4 Imbalances caused by spatial domain decomposition

Processes	Min (sec)	Max (sec)	Average (sec)	Max/Average	Max/Average * Efficiency
4	81.8	234.3	125.7	1.9	
16	1.8	107	35.6	3.0	1.68
36	1.0	60.2	15.1	4.0	1.72
144	0.9	21.6	4.7	4.6	1.2

The Average column shows the averaged calculation time of the processes; the fraction Max/Average therefore shows the theoretical upper limit of improvement in case the distributed load could be evenly between the processes. Without taking additional changes into account, Max/Average * Efficiency shows the upper limit for the parallel efficiency.

4.1.5 Preliminary results

In this section we present some figures generated with the preliminary model. We start with figures where the resulting data from the different cells are aggregated into time-series for the whole world.

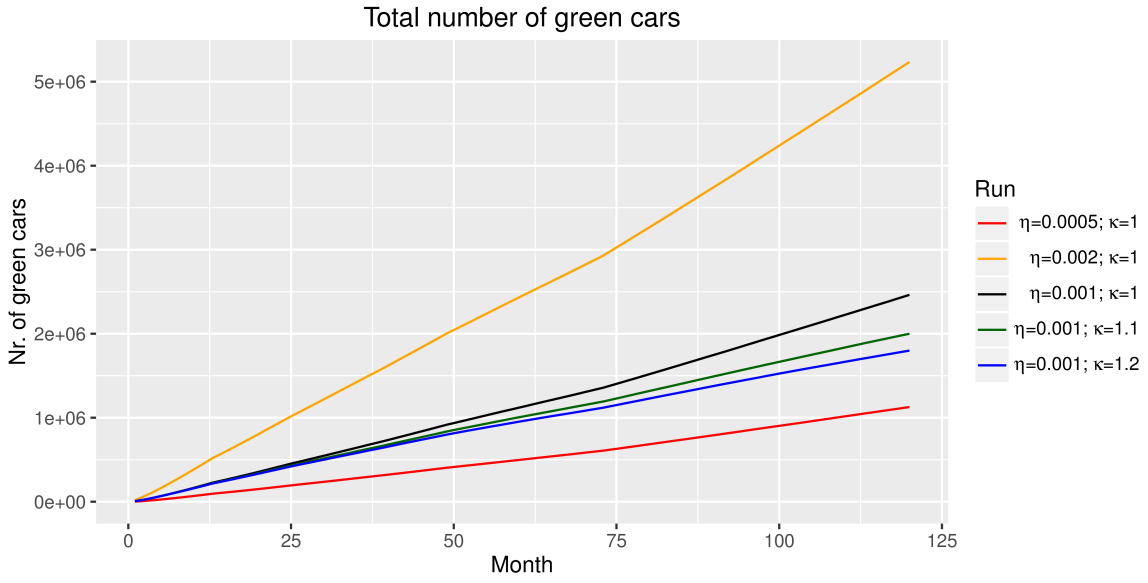

Figure 13: Total number of green cars for different model parameters.

As described in Section 4.1.3 we have two parameters: η which determines the share of innovators, and κ that describes the resistance to change. In Figure 13 and 14 trajectories of the total number of green cars in the world produced using the deterministic model with varying eta and kappa are shown. To produce the other figures in this section the parameters $\eta = 0.0005$ and $\kappa = 1$ were used.

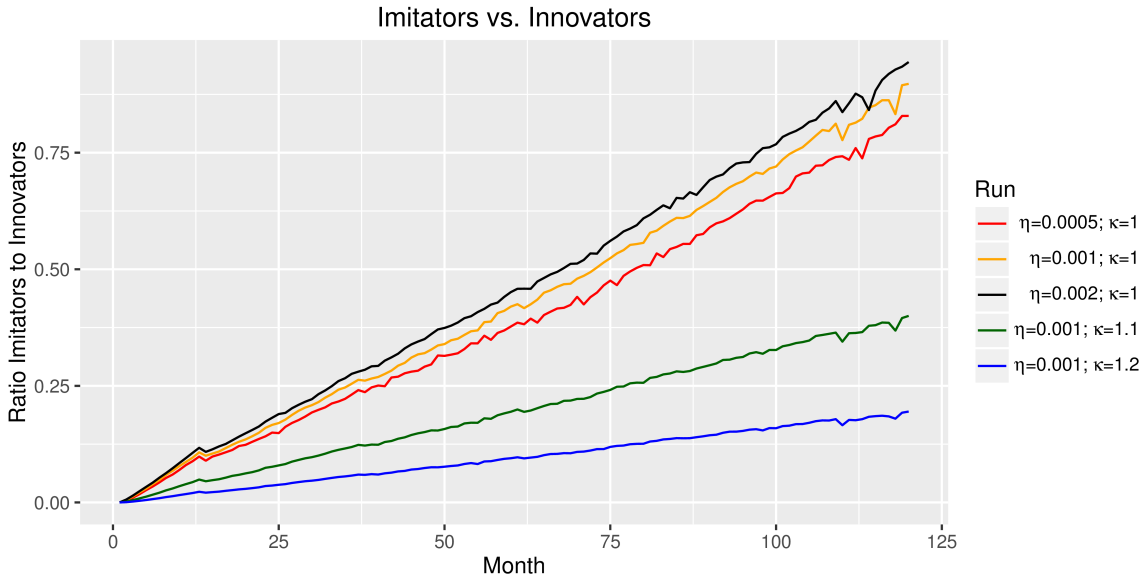

Figure 14: Ratio between number of green cars bought by imitators and innovators.

To see how many green cars innovators vs. imitators bought, additional counters were added to the model implementation:

$$n^{\text{inno}}_{\hat{x},\hat{y}} = \frac{\eta G}{s^{\text{new}}_{\hat{x},\hat{y}}} \cdot n^{\text{green}}_{\hat{x},\hat{y}}$$
$$n^{\text{imit}}_{\hat{x},\hat{y}} = n^{\text{green}}_{\hat{x},\hat{y}} - n^{\text{inno}}_{\hat{x},\hat{y}} \qquad (3)$$

Figure 14 shows how the ratio between n^{imit} and n^{inno} changes over time. As expectable, for higher κ the ratio increases more slowly. This holds also for higher η, in this case in the imitators' neighbourhood green cars appear in higher numbers, so the imitation process can start earlier.

Figure 15 shows different runs with $\eta = 0.0005$ and $\kappa = 1$. Two observations are conspicuous: The different random seeds do not have a real impact on the results,[10] but we get a much lower trajectory for the deterministic run. As shown in (2) we use the floor function to convert the calculated real number of new green cars to an integer and transfer the fraction portion to the next step. While in the random cases for the same $s^{\text{new}}_{x,y}$ the same number of green cars are bought on average in all time steps, this is not the case in the

[10] The variation of the total number of green cars after 120 month is smaller than 0.4%.

determinstic version, where in the first period $r_{x,y}$ is zero, and therefore a new green car is only bought when $s_{x,y}^{new} > \frac{1}{n_{x,y}}$.

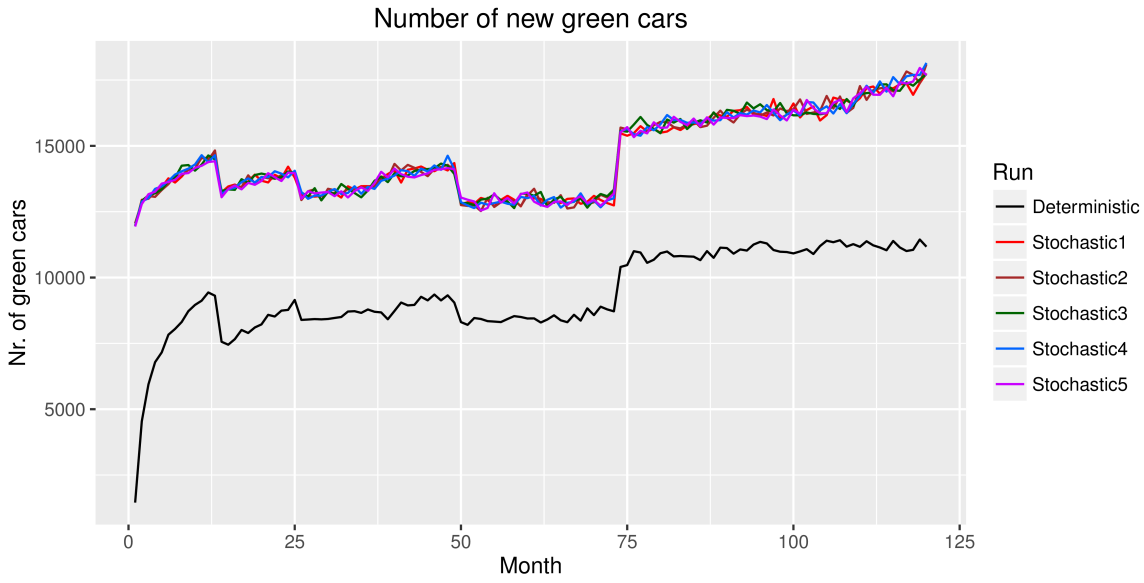

Figure 15: Green cars bought per month for a deterministic and five stochastic runs

If we aggregate the data and run the simulation with a single cell that represents the whole world, $\frac{1}{n_{1,1}}$ is so small that both versions of the model should produce roughly the same numbers.

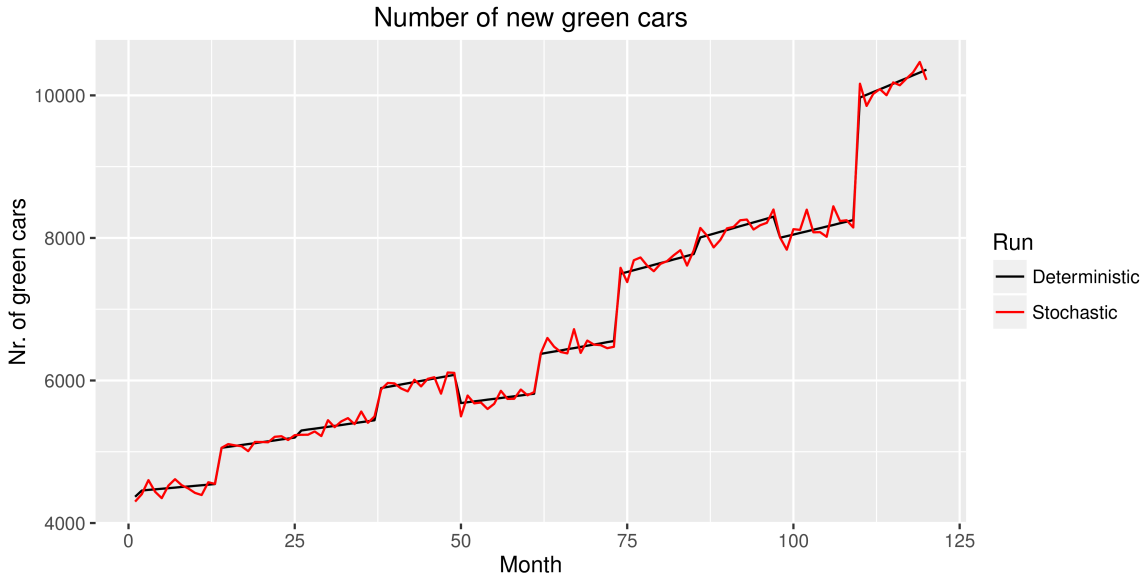

Figure 16: Green cars bought per month for a deterministic and a stochastic run with a 1x1 cell resolution

That this is really the case is shown in Figure 16. This example shows how carefully implementation decisions must be taken, because e.g. rounding of a variable is not executed one time per period, but one time per cell and period.

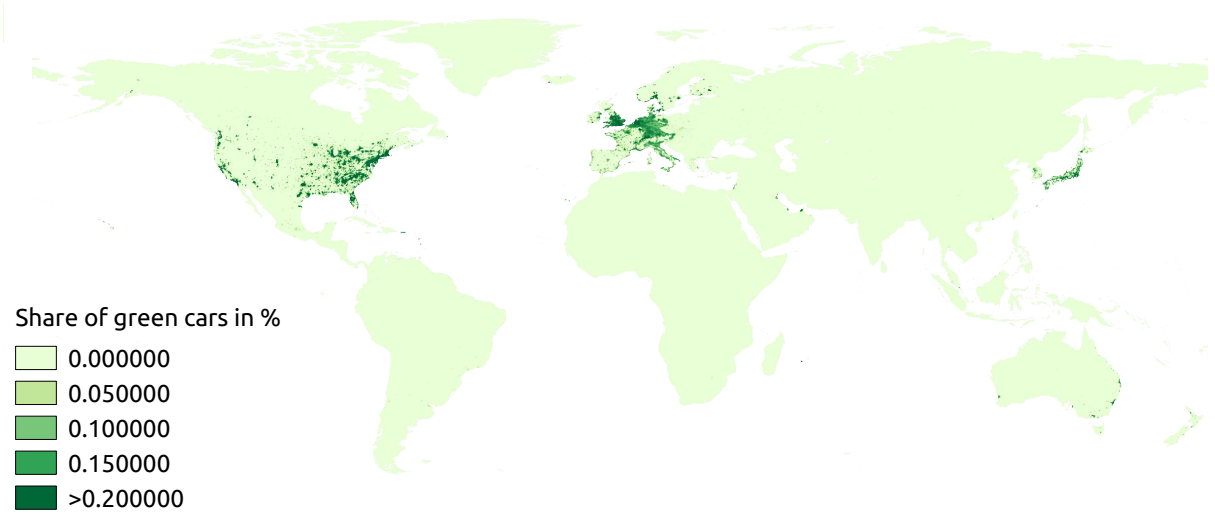

Figure 17: Green cars share after 120 months

The last figure in this section shows a world map of the share of green cars compared to the total number of cars after 120 months for a deterministic run with $\eta = 0.0005$ and $\kappa = 1$. It is easy to comprehend that the share correlates to the GDP per capita. But it is also recognizable that in rural areas of the countries the share is lower than in the urban areas. This can be explained by the lower number of new cars per cell in the rural areas, implying that it takes longer till the innovators buy a significant number of green cars, so that the imitation process is delayed compared to the urban areas.

Overall, even these first results with the preliminary model, which was mainly developed to get familiar with the HPC world and does not claim that it produces reliable results, shows that going from an aggregated to a fine grained view can evoke additional insights, but must also be done with great caution.

4.2 Agent-based SIS

In moving to an agent-based model as the heart of the next version of the GG pilot SIS, we keep some of the previous model's features for simplicity, while pursuing the objective to incorporate realistic decision processes for individual agents. An interesting approach for the basic agent features can be found in (Epstein, 2014), which presents a minimalistic model of agent decision making, yet capable to create complex patterns of interaction.

We rely on standard economic theory in specifying the agents, with the idea that the dynamics obtained when parametrizing agents to conform to standard economic models with optimizing representative agents should reproduce standard model dynamics, from which we can then move away step-wise by adding agent heterogeneity, networks, contagion, and social learning.

4.2.1 Conceptual model specification

The following list gives an overview over the conceptual model. After a brief sketch of expected utility theory (Section 4.2.1.1), details of the model specification are added where necessary. The most important model components comprise expected utility theory, the concept how agents learn and change their beliefs and how the environment affects the agents.

- The model still iterates by time steps; in each step, all agents may gather information, and some agents buy a car. The level of detail at which the agents' activities are modelled will induce the necessary time discretization.
- As we consider the decision of buying a car a household level decision, we chose **household** as the agent level. Yet, members of the household are nevertheless included to derive properties of the household.
- Agents are based on standard economic theory: in particular, they are characterised by some **features** (such as their income, the number and ages of people living in the household, etc. – see Section 4.2.3), and they have **preferences**, represented by a utility function, as well as **beliefs**, represented by subjective probability distributions.
- One feature of agents that is often neglected in standard economic models is the **spatial location**. Agents belong to a cell in the gridded map described for the previous model. This location has various impacts on its interaction with other agents, its information exchange and its observations. All this will enter in the generation of the social network and the interaction between agents.
- The feature under study of agents in this model is whether they own a **car** and in particular what type of car they choose when buying a new one. Cars also correspond to a set of properties and features that they provide.
- While agent features will generally change over time, we will add these changes one by one once the model is up and running. For now, only obvious features shall change (e.g., age increases every year) and of course the feature under study changes according to the agents' decisions taken.
- **Preferences** of agents concern three needs relating to cars: safety, ecology, and convenience. Different agents consider these three felicities of different importance (e.g., based on their features), but for all agents, utility is increasing in all these features (see Section 4.2.1.3 below).
- Cars have properties that relate to these needs: e.g. the number of airbags says something about safety, or the engine type and fuel consumption relate to ecology (see Section 4.2.1.4).

- However, agents do not know the levels of safety, ecology and convenience they will obtain from a given car. Rather, there is uncertainty for the agents as to what level of each of these needs a certain car yields for them. They therefore have **beliefs** about the levels of safety, ecology, and convenience they will obtain from each type of car they might buy (see Section 4.2.1.5).

- To remain close to the previous model, for now we stick to the fact that at each time step, "cells buy a number of cars" given from data for the past (see Section 4.1.2 above), and from scenarios about the evolution of the global car population (produced separately, see Section 4.2.2.2) for the future. The corresponding agents are then triggered to take the decision which car to buy. This can be done deterministically or randomly, based on some properties of households, e.g., income, current utility, car age.

- When chosen to buy a car, agents take the decision which type of car to buy by **maximizing expected utility given some constraints**, e.g. their budget. That is, given their preferences and beliefs, they compute which car provides them with the maximal expected utility under the constraints they face (see Section 4.2.1.6).

- Once a household has bought a car, it obtains the "true" levels of safety, ecology and convenience this car provides.

- While preferences are kept fixed for now, agents **learn** by adapting their beliefs (see Section 4.2.1.7). This happens due to observations (e.g., levels obtained from the car bought, or the number of cars of a certain type seen in the neighbourhood), and due to information exchange with other agents. In particular, agents are linked to others via **networks**. The exchange of beliefs in this network leads to **social contagion** and **social learning** (see Section 4.2.1.8).

4.2.1.1 Expected utility theory

Decision-making under uncertainty means that the choice of an action, by itself, does not determine a unique outcome or consequence. Several consequences are generally possible, and at the point in time when the decision is made, it is not clear to the person taking the decision, which one out of the possible outcomes will be obtained. Standard economic theory has agents maximize expected utility when faced with such decisions under uncertainty. Expectation presupposes a probability measure, according to which an expected value is computed. Several models are available in the literature, differing whether these probabilities are objectively given (von Neumann-Morgenstern) or subjective for agents (Savage), or a bit of both (Anscombe-Aumann) (see Kreps, 1988).

Without going into any detail here, we consider the assumption of objective probabilities not fit to analyse the evolution of the global car population, and therefore start out using a Savage approach with subjective probabilities. This allows for a broader class of preferences

to be considered as "rational" because different agents may consider different probabilities in weighing the utilities of uncertain outcomes.

Savage poses 7 postulates on preferences over uncertain outcomes (Karni, 2005), under which these preferences are considered to be rational. Given the postulates, a representation theorem holds, which says that the preferences can be expressed as expected utilities with a utility function (unique up to monotonic transformations) and a unique (subjective) probability measure. In formulae, there is

- a set of possible outcomes, X (often also referred to as consequences),
- a utility function over outcomes $u : X \longrightarrow \mathbb{R}$, which attaches a value $u(x)$ to each consequence (and fulfils the usual conditions for utility functions), and
- a set of possible actions, A.
- An action leads to an outcome, however, there is uncertainty as to which one of generally several possible outcomes. In order to compute the expected utility of an action, our decision maker has to weigh the utilities of the action's possible consequences with the probabilities of the action leading to each of these possible consequences. Assuming for simplicity the set of possible consequences discrete

$$\mathbb{E}U(a) = \sum_{x \in X} u(x) P(\text{“}a \text{ leads to } x\text{”}) \tag{4}$$

- To make "a leads to x" a little bit more formal, consider a sample space S, with the usual interpretation that its elements $s \in S$ represent all possibilities, and one of them is (going to be) the "true" state.[11] Events E can be represented as subsets of S, and an event is said to take place, if its subset contains the true s. Usually these are referred to as "states of the world", and interpreted as "a description of the world so complete that, if true and known, the consequences of every action would be known" (Arrow, 1971). Consider the simple example of the actions "take an umbrella" or "not take one" and the consequences "getting wet" or "staying dry". Intuitive states of the world here are "it rains" or "it does not rain". Given an action, a state of the world s determines a consequence x, so actions are sometimes formalized as functions from S to X. In the above example, "take an umbrella" maps both states to the consequence "stay dry", "not take an umbrella" maps "it rains" to "get wet" and "it does not rain" to "stay dry". Formally, the functions $u : X \to \mathbb{R}$

[11] Note that this sample space is not to be confused with the state space of the agent-based model, which would record, for each time step, all the relevant information that is needed to restart the simulation at that time-step. There will be other sources of uncertainty within this state space (e.g., due to random components in which agent interacts with which other agent), which, if to be explicitly specified, would require a much larger sample space.

and $a : S \to X$ can be composed to yield a random variable $u \circ a : S \to \mathbb{R}$, denoted $U(a) : S \to \mathbb{R}$ in the following, the expectation of which corresponds to the expected utility of the action. Assuming a discrete sample space for simplicity, we write

$$\mathbb{E}U(a) = \sum_{s \in S} u(a(s))P(s) \tag{5}$$

However, the agents do not need to distinguish all states of the world, and it would actually be very difficult if not impossible to specify a complete space of possible states of the world for each agent, let alone for all agents, at each time step in our GG pilot model. When evaluating an action, an agent only needs to have subjective probabilities for all possible events of the form "action a leads to consequence x": $E_{a,x} = \{s|a(s) = x\}$. In the above example, the agent would not need to distinguish the states "it rains" and "it does not rain" when considering the action "take an umbrella". However, the expected utility of taking an umbrella is useful only in comparison with the expected utility of not taking one, where the subjective probability of rain and no rain does make a difference. To compare expected utilities of taking or not taking an umbrella, three events are of interest:

$$E_{\text{take umbrella, stay dry}} = S$$
$$E_{\text{do not take umbrella, get wet}} = \{\text{rain}\}$$
$$E_{\text{do not take umbrella, stay dry}} = \{\text{no rain}\}$$

The expected utility can thus be considered at this "event level of granularity"

$$\mathbb{E}U(a) = \sum_{x \in X} u(x)P(E_{a,x}) = \sum_{x \in X} u(x)P(\{s|a(s) = x\}). \tag{6}$$

4.2.1.2 Households

In this model version, agents are households since the decision of buying a car can be related to that level of aggregation. Therefore, the properties of the household aggregate the properties of the people living in the household that include p_{agent} = {location, number of people, age(s), income(s), education level(s)}. In addition, the social network of the agent will be described such that its social interaction with other agents and the contagion of ideas in the network can be modeled. The network description will comprise $p_{network}$ = {location(s) of work or school, friends}.

Each action or decision of an agent is subject to uncertainty and subjectivity. Only later stages of the model will reveal, whether it will be necessary to switch to the household members as agents.

4.2.1.3 Utility

The input for the utility function $u : X \to \mathbb{R}$ are the consequences X. These consequences are levels of basic needs or felicities resulting from car usage. For this version of the model,

we assume that buying a car contributes to the three felicity dimensions safety, ecology, and convenience: an $x \in X$ is therefore a tuple (x_s, x_e, x_c) where x_s indicates a level of safety and so on.

The utility function u describes the agents' individual preferences over such tuples. We use Cobb-Douglas utility functions from the field of economics which have the following form:

$$u(x) = u((x_s, x_e, x_c)) = x_s^{\alpha_s} \cdot x_e^{\alpha_e} \cdot x_c^{\alpha_c} \text{ with } \alpha_s + \alpha_e + \alpha_c = 1 \quad (7)$$

Different agents may consider the importance of each felicity differently, represented by different parameters α_s, α_e and α_c. In this version, we assume the preferences (and thus the α-parameters) are related to the agents' properties p_{agent} and a random component. The preferences will be determined before initialisation of the ABM, when the agents have been created, based on their individual properties (p_{agents}). At the moment, preferences are derived by simple rules, like households with children prefer higher values of safety. The preference will furthermore not change during the simulation; this may be a task for further model versions.

4.2.1.4 Cars

At the current conceptual stage cars are the only available commodity for mobility. As the decision under consideration is that of which car to buy, possible actions in the set of actions A correspond to the different available car types on the market. At later stages, alternative mobility concepts like car sharing, public transport or the decision against having a car need to be included. Different types of cars provide sets of properties that can be observed by an agent.

At the moment, we assume the "true" level of safety, ecology, and convenience to be exogenously given and fixed for each car. These have been computed based on data about car properties as explained below. This ensures consistency in that levels for similar cars (in terms of car properties like size, number of airbags, etc.) are similar. For a start, we consider graded levels of the felicity dimensions for simplicty, e.g., with 5 levels, so that the set of consequences X has 125 possible elements.

Thus, an initial model was created to derive the levels of safety, ecology and convenience. The model is very general and will be constantly adapted to the available information, data and conceptual model. To improve the intuition of the model, each felicity consists of subclasses. Safety can be split into a physical component, relating to physical properties like weight, safety extensions, like brake assistant, and handling factors like acceleration. Ecology is similarly separated into an emission component, important for densely populated areas, a consumption component related to the fuel consumption, and a resource component that includes the materials for the construction of the cars. Finally, convenience is the most compound felicity that embraces a sportive user experience, the aim for comforts and the social status provided by a car. Such a separation into subclasses will also allow, at a later stage, to refine the importance of the subcomponents depending on the agents' properties. For example, younger agents will have a larger tendency to be interested in a car with sporty

features. Felicities and their subclasses are illustrated in the figure below, where green indicates a positive contribution to the felicity, red a negative one and black stands for an undefined relation.

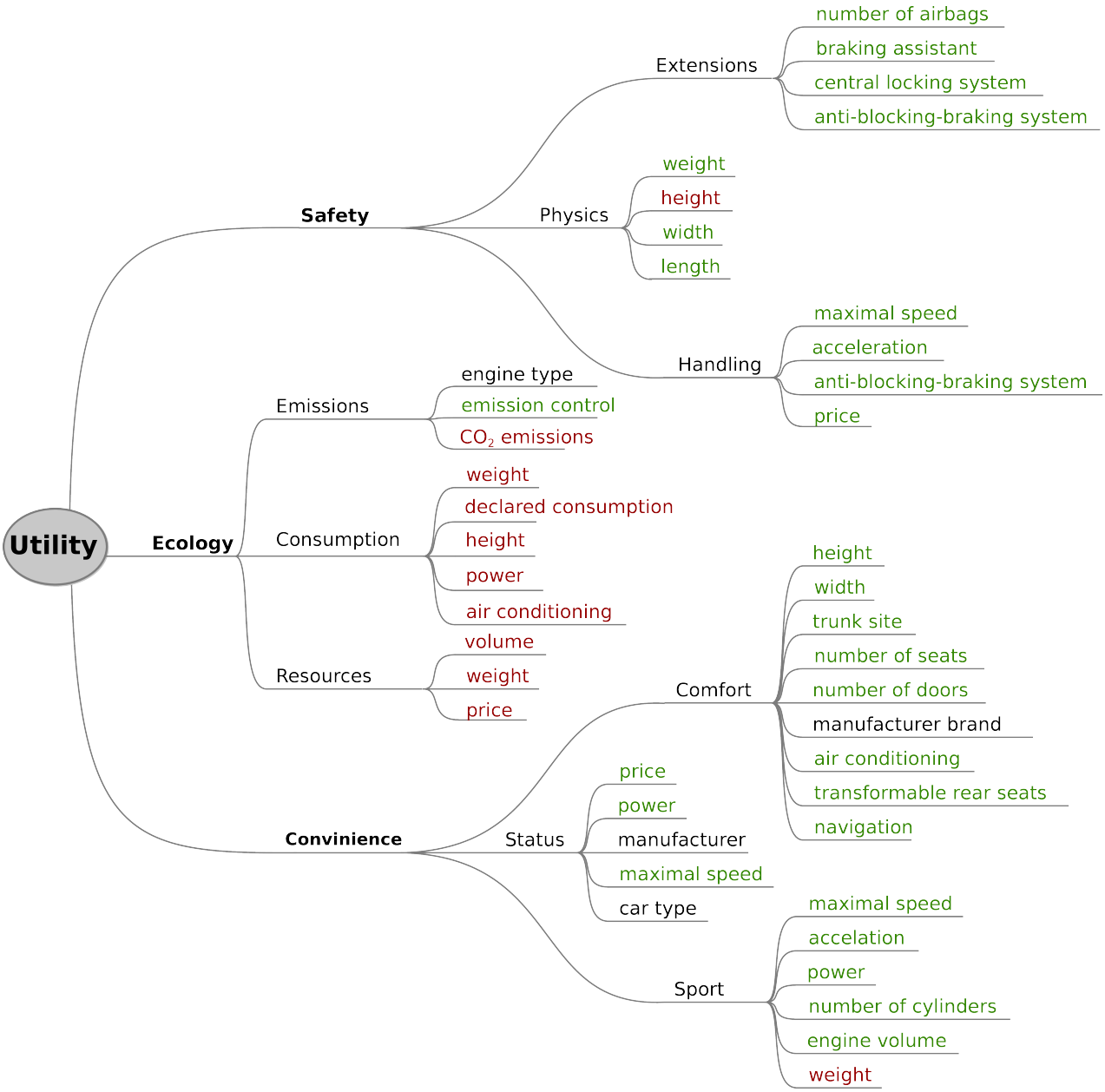

Figure 18: Summary of the links of the technical properties of a car to the individual felicities and their subclasses.

For the calculation of the subcomponents, the distribution of all properties in the current car market pool is normalized. The normalization is done relative to the percentiles of the current distribution, thus, e.g. the 80th percentile relates to a certain value in the normalized values. Such a value for the 80th percentile can be seen as the current technological state of the art, thus, mapping it to a defined value seems reasonable. Similarly, we consider the 20th percentile as the border to out-dated technology. Applying the $tanh$ allows to project any real value into the range between 0 and 1, where 1 represents a high felicity. Thus, the normalized values can be derived as:

$$x_{norm} = 0.5 + \tanh\left(\frac{x - p_{50}^x}{p_{80}^x - p_{50}^x}\right) \quad (8)$$

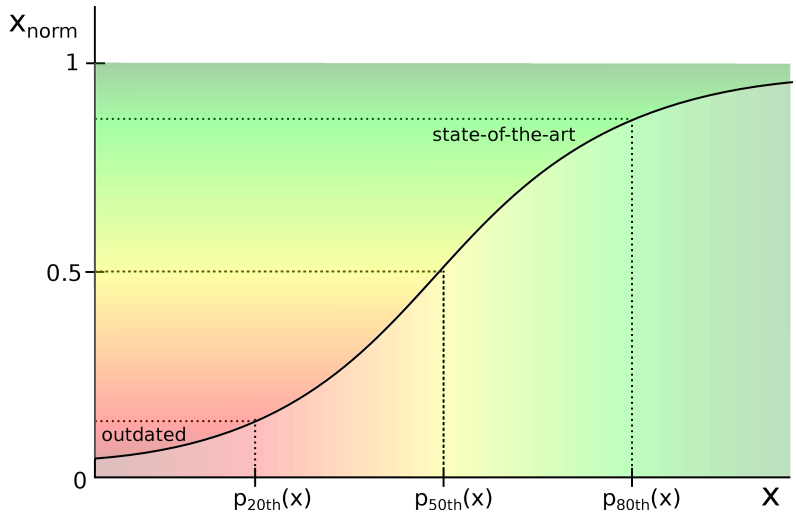

Figure 19: Normalization of the properties relative to the available state-of-the-art

This normalization has the important benefit that, if new cars become available on the market, this will change the percentiles and thus the rating of the existing cars. This map one of the important features of human personal experience that, when better products become available, the evaluation of the older products changes. The subcomponent of the felicity is the weighted mean of all normalized properties that contribute to this subcomponent. The final felicities are computed as the geometric mean of the subcomponents.

At later stages, the model could be refined: Levels might vary for different types of agents, for example, a big car may be more convenient for a big family living in the suburbs (e.g., to easily transport kids and things), but a small car may be more convenient for a single-person household in the city (e.g., to easily find a parking spot). This can be accounted for by different weights for the subcomponents of the felicities. Also, the levels may vary stochastically, e.g., the levels might be given by a function of some car properties and a random error term.

Cars owned by agents further may have features like age, etc.

4.2.1.5 Subjective probability distribution

Agents have beliefs about the levels of safety, ecology, and convenience each car provides, i.e. about the consequences an action comes with, represented by a subjective probability distribution P over the set of events of the form E_{a,x_h} where $x_h \in X$. While fixed at the point of decision-making, these beliefs can change dynamically in a simulation.

A subjective probability distribution for a car is then given by a vector of probabilities (i.e., non-negative real numbers summing up to 1) for the 125 possible elements of X, i.e., possible combinations of the levels of the three features safety, ecology, and convenience. In a first step, we assume these probabilities given at initialization for all agents for all cars available.

As cars have properties relating to safety, ecology, and convenience, an agent may in principle look up all the properties for each car, so that one can also consider the choice between cars as a choice between sets of properties of cars. The agents' beliefs can then be translated into beliefs about properties of cars yielding levels of safety, ecology, and convenience. For example, a common belief or meme is that heavier cars provide more safety, yet are not ecologic in their consumption. In later model versions, agents may start out never having thought about the probabilities they attach to a given car, and may "approximate" beliefs from their beliefs about other cars that are similar in terms of properties. This will also be relevant once new cars enter the market: they may be assessed via comparison of their properties with the properties of existing cars.

4.2.1.6 Decision

The decision the agent has to take by optimizing his/her utility is to choose the most fitting car. Using the notation introduced above, the agents aim to make the optimal decision a_{opt} to maximize the resulting expected utility $\mathbb{E}U(a_{opt})$. Indexing with i, j, and k the possible levels of safety, ecology, and convenience, and denoting the subjective probability distribution of the agent by P, the decision is taken by finding:

$$\begin{aligned} a_{opt} &= \arg\max_{a \in A} \mathbb{E}U(a) = \arg\max_{a \in A} \sum_{x_h \in X} u(x_h) P(E_{a,x_h}) \\ &= \arg\max_{a \in A} \sum_i \sum_j \sum_k x_i^{\alpha_s} \cdot x_j^{\alpha_e} \cdot x_k^{\alpha_c} \cdot P(E_{a,(x_i,x_j,x_k)}) \\ &\text{s.t. the action } a_{opt} \text{ is feasible for the agent} \end{aligned} \quad (9)$$

where feasibility still has to be specified in terms of constraints the agent faces. In this model version, we abstract from agents buying other things than cars, having bank accounts, being able to take out a loan from a bank, to buy a car paying by installments etc. Instead, we use as an upper bound to be spent when buying a car a percentage of the household's yearly income, specified according to the literature (see Section 4.2.2.1).

4.2.1.7 Updating beliefs

In a way, human beliefs are heuristic surrogate models for many complex processes in the real world. Mostly, they may not be correct, but have been proven useful and have been confirmed by observations. Thus, beliefs may change with new observations. For our model, agents observe actions and consequences. These pairs are either observed from themselves or from other agents, where observing others may provide less detailed information.

This means creating a new belief from an old belief and some new information. In statistics this is known as subjective beliefs and the mechanism of Bayesian updating can be used to update these.

In our case of beliefs about actions leading to outcomes, many observations of actions leading to some outcomes can induce learning of the relationships between actions and outcomes; knowledge of the sample space is not required for this. In the umbrella example

above, a person who does not know anything about the weather or rain would be able to learn that using an umbrella always leads to the consequence of being dry by observing many agents using an umbrella or not in different states.

In the green growth pilot model, once an agent buys a car, he will directly observe the consequences of this action. Theoretically, values of safety, ecology, and convenience of cars could be learned by buying very many cars and adapting the belief to the observed frequencies of certain levels. However, buying a car is not an everyday activity, and thus does not lend itself to experimental consumption.

Thus, there are other sources of new information that are important for the agent. These could be observations of the environment (e.g., numbers of types of cars in the neighbourhood, or properties of a new car being announced by advertisement) or direct information exchange with other agents. An agent can share the consequences experienced after buying a car with other agents. For example, agents may observe other agents' pairs $(a, x + \epsilon_c)$ (and here only those experienced by the other agents themselves, or also those heard about from third agents) or other agents' probability distributions, for the car these agents have, or for all cars. The error term ϵ_c is introduced to include an observation or communication error.

In this way, the community of agents gathers a level of information about the consequences of different actions, which a single agent could never have. Of course, the quality of this social knowledge depends on how reliably the information is shared, on the subjective framing of individuals and on possible sources of mis-information. This path of mutual learning or information sharing for example may lead to the assumption that bigger cars provide more safety, unless only very few people experience such consequences.

Depending on the form the new information has, the mechanism for updating the agent's beliefs has to be chosen accordingly. Here, in information obtained from other agents, also features of this other agent might play a role, e.g. if an agent weights information obtained from agents with similar features as more important than information obtained from dissimilar agents.

Computationally, learning of heuristics for the relations between car properties and levels of safety, ecology, and convenience can be modeled using various concepts to represent the partial information agents possess. In this context, high performance computing alleviates the limitation to use only low-complexity heuristics for a large number of agents. Possible choices for implementation are calibration of parametric functional dependencies, neural networks or fuzzy sets of memes.

4.2.1.8 Social contagion and learning

There are two relevant theoretical bodies that are relevant in the context of social spread of technologies.

The first is the theory of social or behavioural contagion (Wheeler, 1966) which describes the spreading of new ideas. The theory transfers mechanisms and processes from epidemiology

and pandemics by treating ideas like viruses that infect agents. These models use classes of agents like infected, susceptible and ignorant (the latter corresponding to the “recovered” in epidemiology) individuals. Such models only work for discrete ideas or memes, which can be followed through the network like an infection. Awareness of specific subjects e.g. electric cars or the incremental change of individual preferences due to news, events and media could be modelled using this approach.

The other relevant theory is the social learning theory (Bandura, 1977), which deals with the learning of individuals in groups. This concept will relate to the learning about the relation between actions and consequences. The theory, also known as observable learning, describes mainly four steps. The first is the attention process that determines whose and which characteristics raise the attention of the agent. Second, the retention and reproduction process generates or memorizes the observed behaviour. The third step is the reinforcement process that possibly overlays the original observation with some positive or negative reinforcement, depending on whether the observed consequences are positive or negative. Finally, in the learning theory, the adapted knowledge will not be actively applied unless there is a motivation for it.

4.2.2 Input and data

4.2.2.1 Car Purchases and Income Levels

This section collects some information about (the money spent on) car purchases in relation to income for Germany and the US, similar information for other countries may be added whenever available. The information serves to determine a budget constraint for an agent’s decision which car to buy.

At an aggregate level, over the last twenty years Germans spent a nearly constant percentage of their income when buying a car: 60% of their yearly net household income for a new car and 30% for a used car (Deutsche Automobil Treuhand GmbH, 2016). This means that the price people paid for a car constantly increased with the household net income (HHNI).

In 2014, the previous car was owned on average 6.3 years if bought used and 5.8 years if bought new, in 2013 it was 6.9 years or 6.3 years respectively. In comparison, in the end of the 70s and beginning of 80s, people kept their cars for less than 3 years (Deutsche Automobil Treuhand GmbH, 2015).

Only 22% of new car buyers and 14% of used car buyers even considered alternative engine types as a buying option and less than 10% of the car buyers intensively looked into that topic (Deutsche Automobil Treuhand GmbH, 2016).

At a more detailed level, some differences appear for different income levels. In 2014, with a monthly HHNI of less than 1,500 EUR car buyers spent an average of 14,670 EUR on a new car or 5,530 EUR respectively on a used car. This was less than in 2013 with 15,370 EUR for a new car and 6,110 EUR for a used car. With a monthly HHNI of 3,000 EUR or more car buyers

spent 32,550 EUR on a new car and 12,920 EUR when buying a used car. On this income level numbers increased compared to 2013 when these were 31,250 EUR and 11,500 EUR.

Income levels and car prices also differ depending on the status of the car buyers. First-time buyers that chose a new car had an average monthly HHNI of 3,114 EUR and paid 15,470 EUR whereas those buying a used car had an average HHNI of 2,507 EUR and paid 5,310 EUR. Buyers replacing a previously owned car with a new car had an average monthly HHNI of 3,701 EUR and spent 30,200 EUR while those replacing the previous car by a used one had a HHNI of 2,924 EUR and spent 11,410 EUR. (Deutsche Automobil Treuhand GmbH, 2015)

The time car buyers in Germany had owned their previous car changed a little over the last five years from 81 months in 2010 via 74, 79, and 83 months to 76 months in 2014 for used cars and from 73 months in 2010 via 68, 70 and 76 months to 70 months in 2014 when the previous car was bought new (Deutsche Automobil Treuhand GmbH, 2015).

A total number of more than 10 million cars was licensed in Germany in 2014 of which 3.04 million were new registrations and 7.07 million changes of ownership, corresponding to 2.3 used cars changing owner per car bought new dat2015.

For the US, data of the Bureau of Transportation Statistics of the United States Department of Transportation indicates that car buyers in the U.S. spent an average of 13,105 USD on a car purchase in 2010 with 26,850 USD for a new car and 8,786 USD for a used car (U.S. Department of Transportation, Bureau of Transportation Statistics, 2011). In that year the median and mean household income in the U.S. were 51,915 USD and 70,883 USD (U.S. Department of Commerce, Census Bureau, 2014). However, seperate data on the average income of car buyers in the U.S. seems unavailble.

In the US, car sales in 2010 totalled 51.434 million, of which 14.55 million were new cars and 36.884 million used cars. This corresponds to 2.5 used cars per new bought car (U.S. Department of Transportation, Bureau of Transportation Statistics, 2011).

4.2.2.2 Evolution scenarios for the total global car population

For simulations, future data on numbers of new cars per cell are needed. Therefore, the aim of this section is to project the global car population per country from 2015 to 2025. Projections are based on the model presented by (Dargay, Gately and Sommer, 2007).

The model provides numbers of cars per 1000 people per country depending on the country's GDP per capita. It is influenced by the country's population density and level of urbanisation and is bounded by a country specific saturation level of car ownership:

$$V_{it} = (\gamma_{MAX} + \lambda\, \overline{D}_{it} + \varphi\, \overline{U}_{it})(\theta_R R_{it} + \theta_F F_{it}) e^{\alpha e^{\beta_i GDP_{it}}} + (1 - \theta_R R_{it} - \theta_F F_{it})\, V_{it-1} + \varepsilon_{it}, \quad (10)$$

where

- V_{it} is the number of cars per thousand people in country i at time t,

- γ_{MAX} is the maximum saturation level specified as the saturation level of the U.S. γ_{USA},
- $\overline{D}_{it} = \begin{cases} D_{it} - D_{USA,t} & \text{if } D_{it} > D_{USA,t} \\ 0 & \text{otherwise.} \end{cases}$ and D_{it} the population density of country i at time t,
- $\overline{U}_{it} = \begin{cases} U_{it} - U_{USA,t} & \text{if } U_{it} > U_{USA,t} \\ 0 & \text{otherwise.} \end{cases}$ and U_{it} the urbanisation of country i at time t, D and U are normalised by taking deviations from their means over all countries and years in the data,
- GDP_{it} is the per capita income in thousands of 1995 $ in country i at time t,
- R_{it} and F_{it} are dummy variables to indicate rise and fall in a country's GDP to allow for the adjustment coefficients θ_R and θ_F to be different,
- and ε_{it} is an error term.

The model estimation by Gately and colleagues with pooled cross-section time-series data on vehicle ownership for 45 countries that comprise about three fourths of world population provided the following coefficients:

$$V_{it} = (852 - 0.000388\,\overline{D}_{it} - 0.007765\,\overline{U}_{it})(0.095R_{it} + 0.084F_{it})e^{-5.897e^{\beta_i GDP_{it}}} + (1 - 0.095R_{it} - 0.084F_{it})\,V_{it-1} \quad (11)$$

and a specific β_i for each of the 45 countries. For the rest of the world they estimated the β coefficient as $\beta_i = 0.21$

There is data available for total car numbers and numbers per 1000 people up to 2013 (see Section 4.1.2 above). Projections of data needed for the model were found as follows:

- GDP per capita for 2013-2021 based on purchasing-power-parity (PPP) in current international dollars provided by the International Monetary Fund (IMF) (International Monetary Fund, 2016). For the years 2022-2025 data on GDP in the same units as above was calculated according to growth rates provided by the U.S. Energy Information Administration (U.S. Energy Information Administration, 2016). Values for R_{it} and F_{it} were computed as follows:

$$R_{it} = \begin{cases} 1 & \text{if } GDP_{it} - GDP_{it-1} > 0 \\ 0 & \text{otherwise.} \end{cases} \qquad F_{it} = \begin{cases} 1 & \text{if } GDP_{it} - GDP_{it-1} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Population density from data on population prospects 2013-2025 and land area, provided by the UN and World Bank (United Nations Department of Economic and Social Affairs, 2015); (World Bank, 2015a)

- Urbanisation data from 1960 and prospects until 2050 as percentage of population in urban areas, provided by the UN (United Nations Department of Economic and Social Affairs, 2014)

The actual projection of vehicle numbers was therefore the conflation of the pooled data and then calculation according to the model. For the 45 countries of the paper's sample the specific β coefficients were used. These countries account for more than 80% of the world's vehicles. For each of the other countries, the coefficient regressed for the whole group was used. Figure 20 sketches the model output of car numbers per 1000 people; as details are not readable in such a format, further pictures show aggregated results.

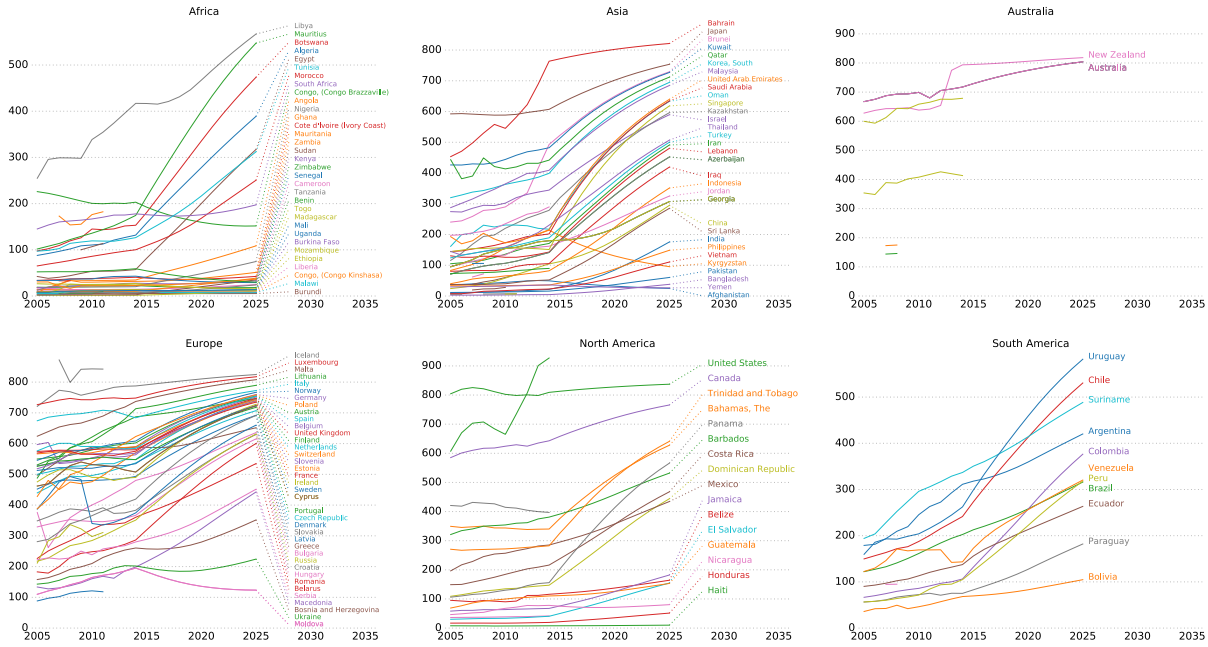

Figure 20: Numbers of vehicles per 1000 people by country per continent - Dargay et al. (2007) projection

Results per continent as well as per country with most cars on its continent were considered. For comparison, a second projection is based on the linear trend of numbers from the past using the mean growth rate for calculating future numbers. Output from both models is displayed in Figure 21.

Figure 21: Total numbers of vehicles by continent (left) and chosen country (right), projections by Dargay et al. (2007) (solid) and linear trend (dashed) with the numbers' projected increment by 2025 in percent of numbers 2014

A big difference depending on the projection method can be seen for countries like China and India. Therefore it is important to consider models like the one by Dargay et al. and not only rely on current trends.

Since the Dargay et al. model is based on projections of GDP per capita, population and urabanisation, for sensitivity analysis, we ran the model also with half and one and a half times the GDP growth rates provided by th U.S. EIA as well as high and low projection cases of world population by the UN.

Figure 22: Total numbers of vehicles by continent, Dargay et al. (2007) projection with high, medium and low population prospects (UN), left, and 1.5, 1 and 0.5 times GDP per capita rates (EIA), right

In Figure 22 one can see, that results are more sensitive to variations in GDP growth than to the variation of population numbers.

The thus obtained numbers of cars per 1000 people per country and year can be used as external input for the dynamic model in the GG pilot SIS as scenarios about the evolution of the global car population for the future.

Data about what kind of cars are preferred and how this changes in the individual countries will be an important source of calibration data for the preference system. Thus, one major challenge will be to gain access to such a data set, although this might only be possible in cooperation with a car manufacturer at later stages in the project.

4.2.2.3 Infrastructure data

The project "Open street maps" provides a global data set about infrastructure and transportation network that is freely available and mainly used for navigation purposes. By processing this data set of about 50 GB we were able to calculate total kilometer of streets per cell, for a given grid (see Figure 23). This indicator can serve as a spatial measure for the density of the infrastructure or urbanization. Other derived indicators are the total rail km per cell and the number of fuel stations per cell.

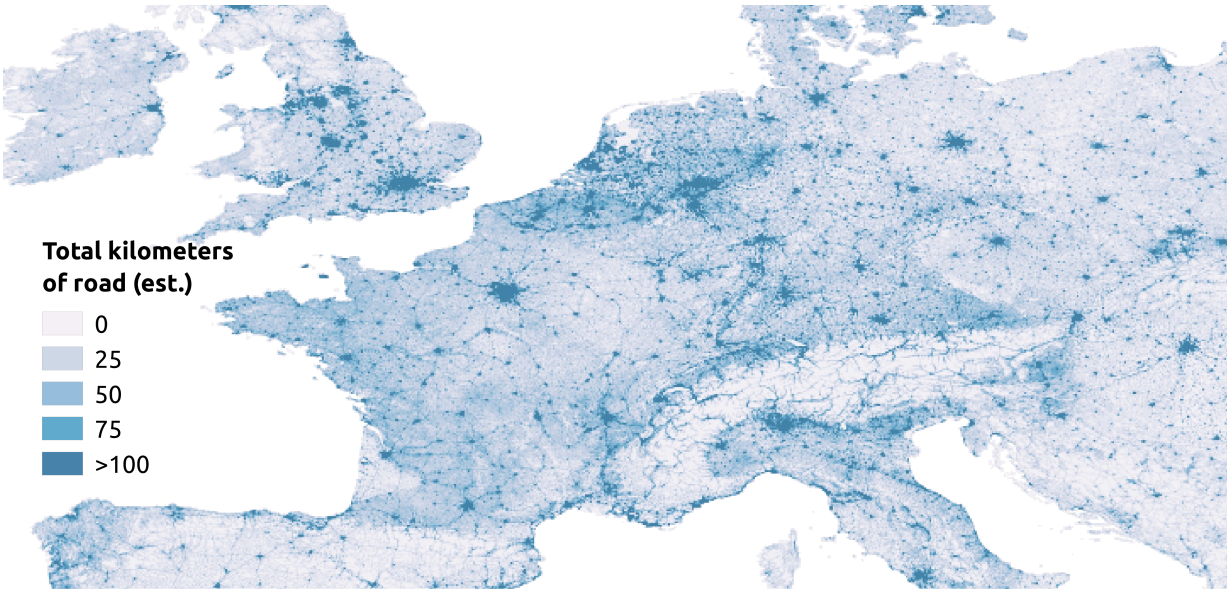

Figure 23: Approximated total km of road for Europe in 2015.

The provider openchargemap (http://www.openchargemap.org) offers an extensive database of electric charging stations for electric cars. This allows similarly calculating the number of electric charging stations in a cell. Although this is a very important type of environmental data to the market of electric cars, the license of the data is currently unclear. Thus, the data remains unused until the situation is solved.

4.2.2.4 Technical data of the available car market

We currently have data of about 50,000 different car types, downloaded from the ADAC Germany. The data is freely available, however, the license of the data is as yet unclear, thus no final statement about usage of the data can be made. The data consists of 64 different manufacturers, 791 car series and 153 technical properties per car. The data is available for models starting in the year 2004 to 2015. The contributions of the individual properties to

the three basic felicities are depicted in Figure 18 in Section 4.2.1. Figure 24 shows relations between car properties: such visualizations can provide hints in the later simulations stages, on how well the learning concept of the agents allows to map these complex dependencies.

In future, safety could be related to the insurance class for each car, however this data is not yet available. Such insurance classes are provided by car insurance for every car type, as yet the availability and license of that data is unclear.

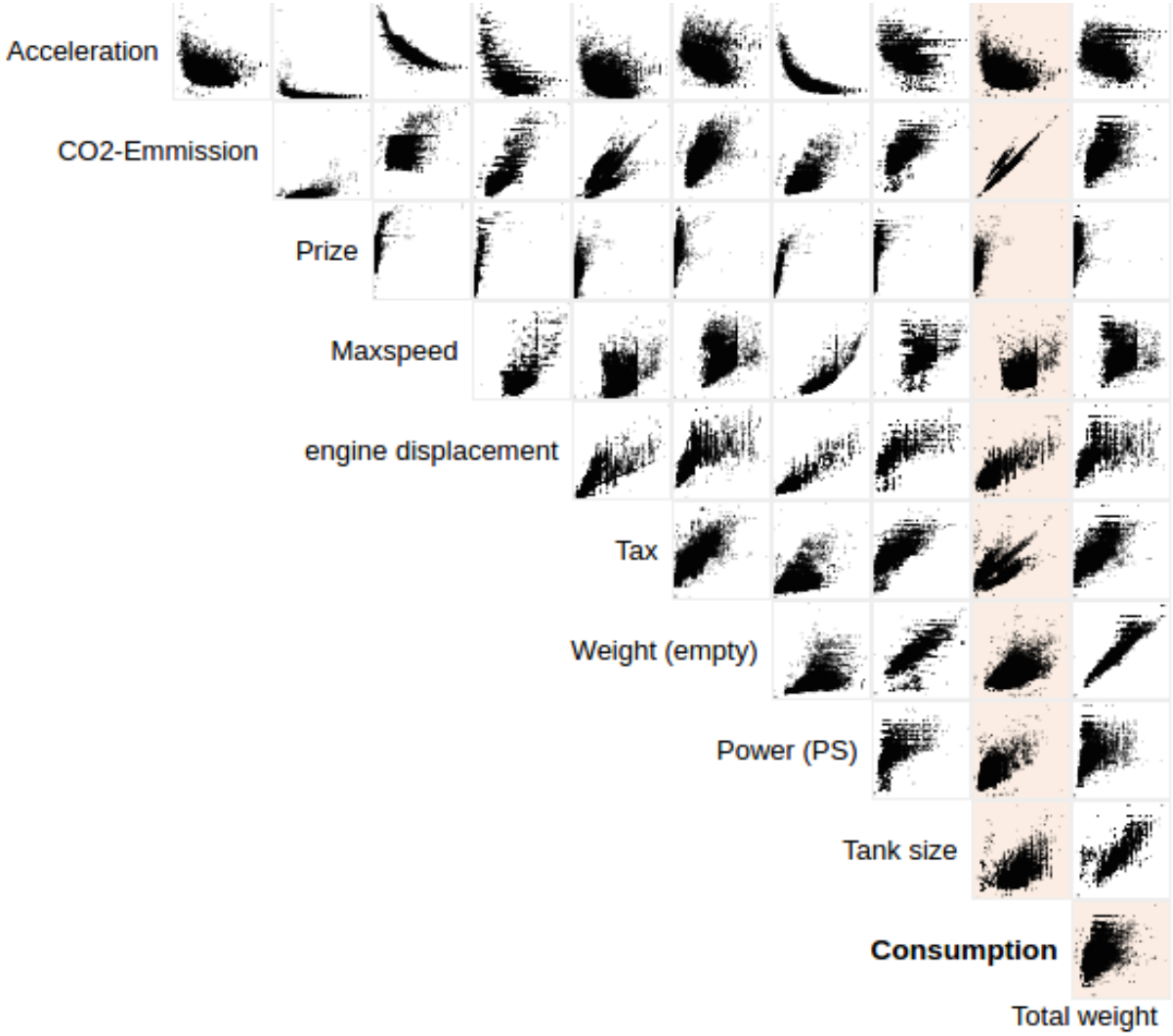

Figure 24: Relation between selected car properties that illustrate the dependence between different technical properties.

4.2.2.5 Microdata

Micro data is the core data source for the generation of synthetic populations (see Section 4.2.3). Micro data mainly consist of survey data that is collected from a small group of people. For this group of people, as a representation of the entire population, a large amount of data is available. This includes the relations between different data, e.g., the income, age, education and living condition. For the entire population, only limited data is available, mostly the marginal distributions of the single indicators. Micro data allows

relating statistical information of the entire population, which are only available as individual distributions. Thus, dependence between different variables can be reconstructed, e.g., a person with a high income gets assigned a reasonable age and so on. In this way, a synthetic population that matches the global marginal statistics and the joint relations form the micro data can be generated.

Up to the moment, we have identified several promising sources for micro data:

- EuroStat: http://ec.europa.eu/eurostat/web/microdata/overview
- European Social Survey: http://www.europeansocialsurvey.org
- Worldbank: http://microdata.worldbank.org/index.php/catalog
- IPUMS: https://international.ipums.org/international/

However, mostly the data is only available for non-commercial, scientific or educational purposes. This applies to the three pilot studies, yet for a commercial usage in CoeGSS, new alternatives to gather the required micro data need to be evaluated.

4.2.3 Synthetic population

A synthetic population is the major input for the modelling of social behaviour using an agent-based simulation model. The GG pilot study will need a synthetic population, ultimately for all countries. Only a realistic population of agents, which are equipped with statistically correct properties, allows simulating realistic dynamics based on these agents and their properties.

Thus, the custom generation of such populations is a core topic in the entire project, tackled by different work packages. At the current development stage, we seek for access to a synthetic population for a start.

The MIDAS-project (Models of Infectious Disease Agent Study) (http://www.epimodels.org/drupal-new/) offers synthetic populations for over 80 countries. The data is freely available and we are in contract with the developer about the possibility to generate custom populations. The amount of data ranges from 1 to 256 GB per country, containing data about synthetic persons and their households. The populations were tailored for the use in agent based models in the area of pandemic contagion models. They are applied for models that predict the spread of the flu in the US.

Since the MIDAS-project is centred in the US, currently important countries in the EU are missing. Thus, simultaneously, we look into the tools and libraries to tailor our own custom made synthetic populations for later stages. The data-base Eurostat (http://ec.europa.eu/eurostat/data/database) offers detailed information about households in the EU. This comprises e.g., the distribution of different household types, the number of children in the household, the average number of persons living in it and the number of working persons. Furthermore, data about the age distribution for the different countries is provided by the UN (https://esa.un.org/unpd/wpp/Download/Standard/Population/), and

The World Bank provides the income distributions for most countries in the world, including all European states (http://iresearch.worldbank.org/PovcalNet/).

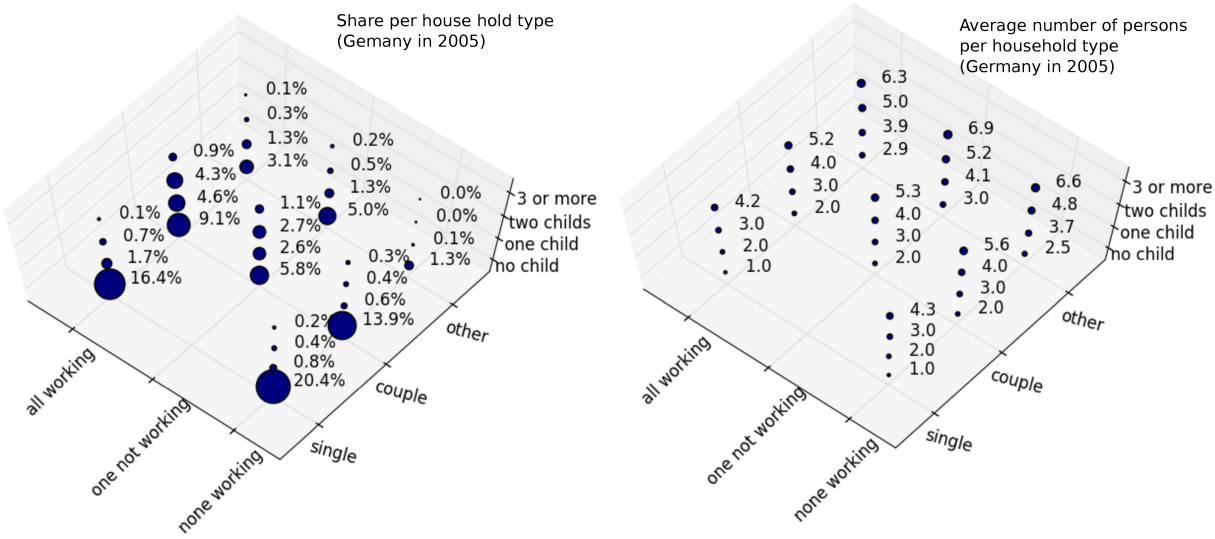

Figure 25: Example of the household data for Germany in 2005, provided by EuroStat.

Altogether, these data sets can be used to supplement the already available synthetic populations, however on a different level of detail. However, within the simulation process, different synthetic populations can be used within the simulations and thus, their influence on the simulation results can be evaluated. Together with WP3 and WP5, we also evaluate the possibility to derive custom synthetic populations by extending already computed synthetic populations stepwise.

4.2.4 Possible implementation strategies

The conceptual model specified above shall be implemented using (as before) the HPC-ABM framework Pandora. This section describes some elements that have already been implemented for testing.

4.2.4.1 Agent networks

The interaction with other agents presupposes social networks between agents to be specified. Here, interaction will first of all be information exchange that serves to update beliefs. To model such interaction, we will first focus on the interaction between agents that are spatially close. Thus, an agent network of friendship or acquaintanceship is generated to mimic the interaction within the agent social network. For the beginning this means the real live social network is not related to the Internet.

At the current model stage, we stick to the simple neighbourhood relations from the preliminary model for simplicity. For a start, a stochastic generation mechanism of agent networks has been implemented. It constructs an agent network based on urbanization and population data. This means, that connections between agents are assigned in proportion to a certain travel resistance that is assumed to be inversely proportional to the infrastructure

density or urbanization. This rule of thumb is very rough and only an intermediate solution, yet, covers many realistic features. It implies that people travel along main traffic lines more often. Agents in heterogeneous regions tend to have more interactions with agents in more densely populated areas. Last, with less urbanization, agents tend to travel longer distances to meet friends and to pursue their activities.

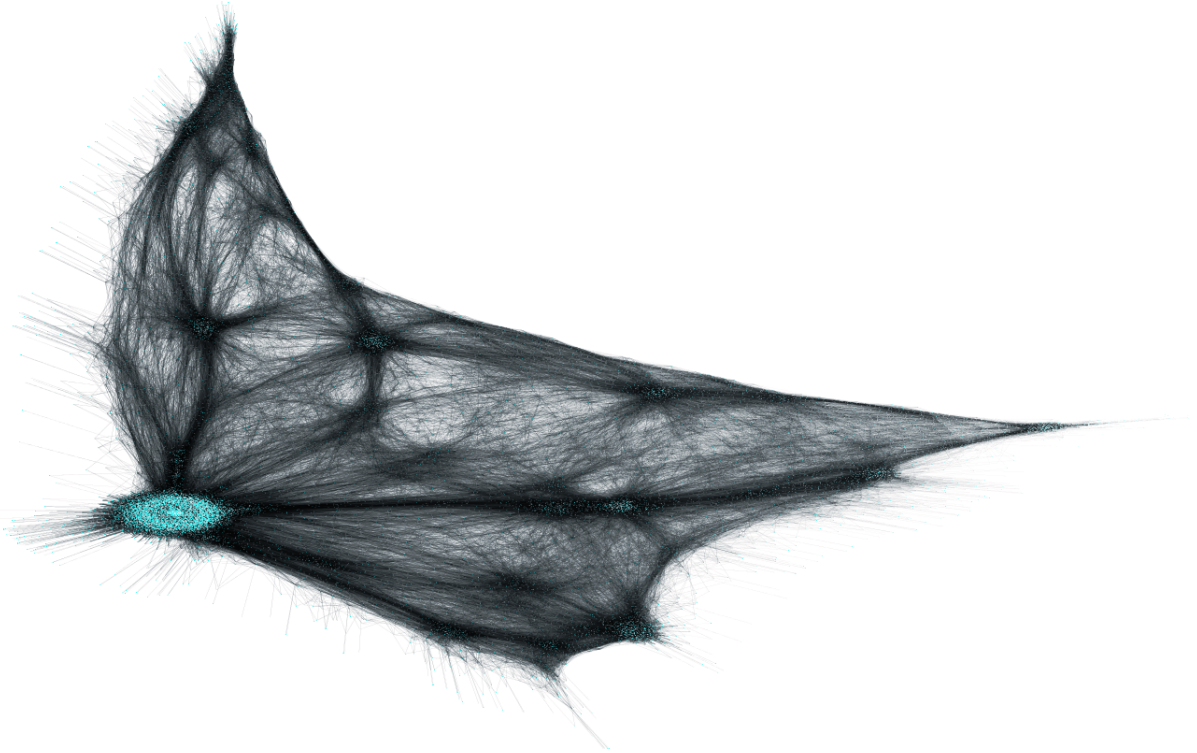

Figure 26: Visualization of an approximated social network graph for a reduced population of north France

Figure 26 shows a preliminary agent graph of Northern France. The graphical representation is arbitrary since the spatial location is ignored for the visualization, yet such visualizations are important for identifying important characteristics of the networks. All 3D layouts produced by algorithms are arranged in a more or less 2-dimensional structure. This indicates for example the underlying spatial structure that is implied by the travel resistance in the generation. Furthermore, main connecting strings are apparent that follow the connection of large urban cities. At later stages, Internet social networks will connect nodes in the graph, regardless of the spatial distance and thus, will heavily increase the connectivity in the graph.

This visualization already provides the modeller an impression, how the contagion of ideas and behaviour might follow these main paths that connect urban centres. In contrast to epidemic contagion modelling, the extension of internet-based social networks will add short cuts to these main connection paths.

More realistic social networks between agents will be added in later model versions, based on common work between the pilots and the network experts involved in the project, see Section 2.5.

4.2.4.2 Self-learning using neural networks

Artificial Neural Networks (ANN) or Bayesian neural networks are one possible implementation of an agent's surrogate heuristic for the complex relations between actions, states and consequences. Neural networks are similar to functional relationships, but are constructed by many atomic relations and fit well in the HPC context. Within an ANN, several layers of neurons are connected by a simple functional dependence. For a multi-agent model, very small neural networks can be used for the individual agents. For example, a complex two-dimensional function can approximate a network of 8 neurons. An example how this network is constructed is shown in Figure 27. The first two layers are connected by a transfer function $tanh$, the influence between connected nodes is controlled by the weights w_{ij} and each node has a bias b_{ij}. The state of each neuron is evaluated as the sum of all connections. The output layer is calculated by simple summation without transfer function.

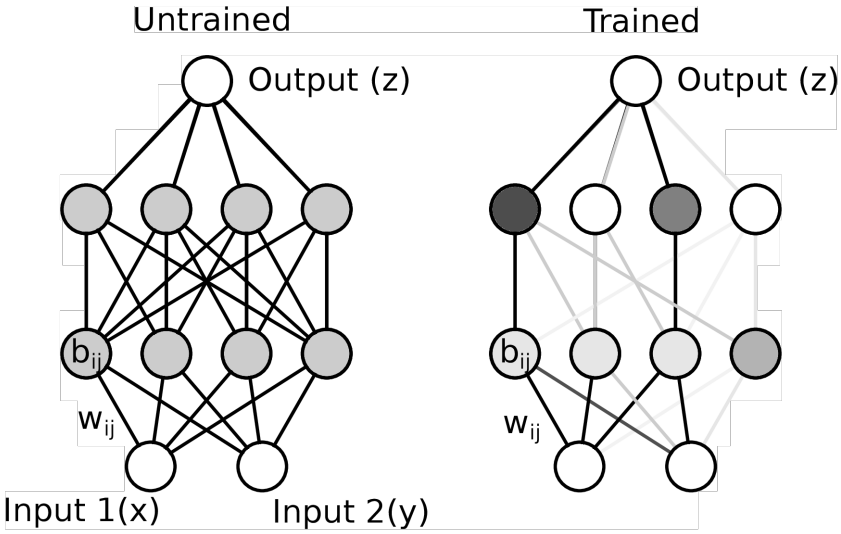

Figure 27: Simple example of a two-layer neural network with 8 neurons, connecting two inputs to one output, untrained (left) and trained (right).

Nevertheless, this simple network allows to model arbitrary two-dimensional functional relations. One example of a stepwise training with new observations is shown in Figure 28. It can be observed how the initially badly approximated function improves to a very nice surrogate in the final stage.

On the other hand, the modeller loses the ability to select the class and shape of the function. This may lead to very unexpected function shape, similar to the one in the first plot of Figure 28. Several steps are shown for an arbitrarily chosen example relation, given by $z(x, y) = 1/(x/4 + .05) + 5 * (\sin(y * 7) + (y + 1)^2)$. Each plot shows the original function in gray and the dots indicate the available observations. At each step, ten more observations become available and are used to train the NN. The colored plane illustrates the function approximated by the NN. Such shape will heavily influence the latter optimization and decisions of the agent. However, this might model the unexpected and

irrational behavior that can actually be observed and will eventually be corrected by later observations. Thus, the benefits of using ANNs for the agents' learning need to be evaluated.

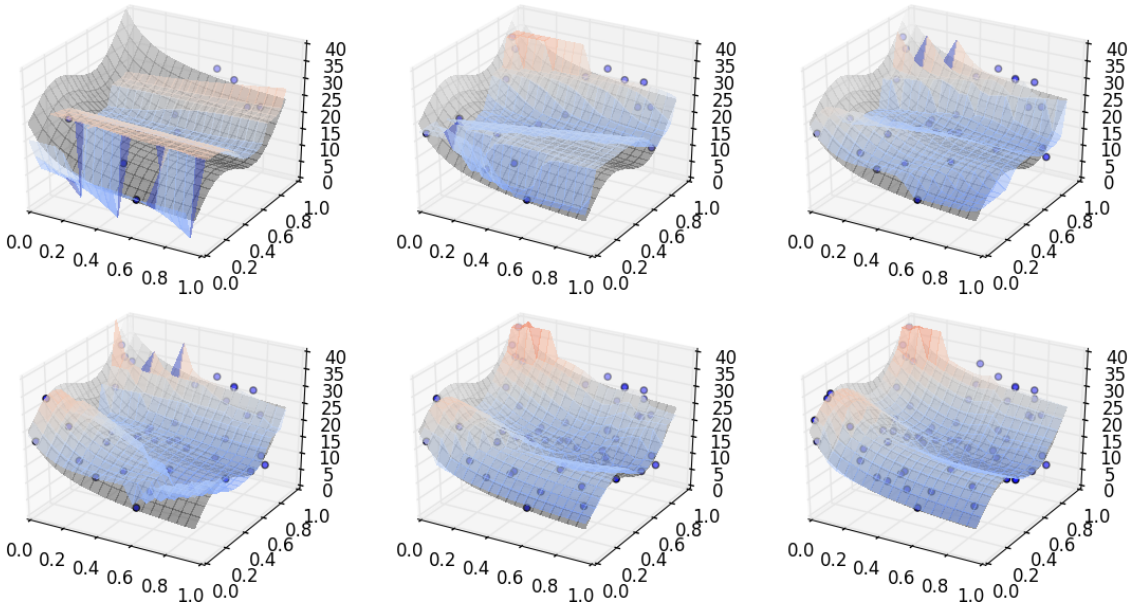

Figure 28: Example of the learning of a neural network.

Potential open source libraries are: The Fast Artificial Neural Network Library (FANN) (http://leenissen.dk/fann/wp/), which is an older, but popular library for C++ and multiple other bindings. Tensor Flow, developed by Google (https://www.tensorflow.org/) is a highly sophisticated and new machine-learning library. Yet, a lot of the functionalities will not be necessary for the basic ANNs in their current form. It will rather serve as a source of implementation solutions, especially since Tensor Flow heavily relies on parallelization.

4.3 Outlook

The further work on the Green Growth pilot will of course be geared to producing a more complete HPC-based SIS for exploring potential evolutions of the global car population with a truly agent-based model at its centre. A blueprint for such an ABM has been presented here. In the actual implementation, the aim of making small steps in adding complexity will be further pursued. At each step, model analysis and visualisation will be important elements for coming up with a sound model.

Further, stakeholder involvement, as one of the characteristic traits of GSS, will become important in further developing the SIS. As soon as possible, we plan to approach, for example, important players in the car industry for feedback on questions like what type of options the SIS should be able to help to explore.

5 Status of the Global Urbanisation pilot

5.1 Identifying thematic and scientific problematic and questions

5.1.1 Global urbanization pilot thematic specific stakes and questions

Optimizing city development choices is an essential challenge for the future, driven by the growth of world population increasingly living in cities and the opportunities of more overall and complete approaches assisted by always more global and intelligent technological innovations.

Cities are complexly defined by the interaction of processes as different as real estate, transportation, economy, society and politics. Insightful decisions prove therefore essential to determine city development from qualitative and quantitative points of view and its harmonious integration in its environment. Indeed too quick quantitative development can endanger qualitative development, on which quantitative development depends. Furthermore, quality determines attractiveness for inhabitants and businesses, which depends on many different elements going from, at the lowest scale, traffic fluidity and pollution, to city accessibility and economic policies and advantages at the highest one. This multi-fold evolution depends principally on the city council's decisions but also on the population and on every interdependent citizen's behaviours, which influence the everyday life and spirit of a city.

Therefore, while being an essential stake for the world of tomorrow, finding insightful decisions for the qualitative and quantitative complex city requires specific modelling and simulation tools.

Modelling city and population dynamics rests on individuals' multi-fold characteristics, particularly here geographic (concerning housing or transportation), which will influence their evolution. Therefore the possibility of creating sets of synthetic populations with realistically statistically distributed characteristics' values will prove precious for any simulation.

Modelling cities and linked population dynamics concerning very different features going from opinion propagation to housing preferences or transportation behaviours, requires a global system science approach, coupling sub-models from very different fields describing very different processes: real estate, transportation, economy, ecology… These sub-models can further correspond to different modelling approaches and formalisms (deterministic, stochastic, equation or agent-based …) and correspond to very different time and space scales, going from hourly traffic congestion to long term city development. These processes can prove non-linear and difficult to predict. Furthermore not only are these sub-models

essential, but also their two-way, heterogeneous, multi-fold and evolving interactions require precise modelling.

Consequently a complex systems approach proves essential to conceptualize in a clearer way and predict more realistically the complexity of city evolution, based on synthetic populations. High performance computation will prove precious to help optimize their simulation time.

Modelling city complexity allows improving the understanding of the city's evolution and attaining more realistic predictions. It can permit furthermore to clarify influences between very different elements of the city and point to possible or more efficient levers to improve city's everyday life. For instance, it will allow exploring the impact of development choices, the precise two-way relation between price mechanisms and infrastructure decisions, but it might also open the way to study the effect of overall opinion and behaviour dynamics and assess the benefit of information campaigns or other public incentives.

Finally, it might provide a more global and complete vision to assess not only the immediate but the far and long term consequences of city-linked decisions to help truly optimize them.

The global urbanization pilot aims at studying the two-way relation between transport infrastructure decisions and price mechanisms, particularly concerning real estate. This pilot is also particularly concerned with the link to geographical information systems.

5.1.2 Global urbanization pilot scientific specific stakes and questions: how to put into light the benefit of HPC for GSS

A purpose of the pilot is to put into light the benefits of high performance computing to tackle global system science challenges. From a general point of view such benefits could be imagined at the two extremes where global systems science tries to push further the frontiers of knowledge.

Firstly, to help focus on precise evolutions of processes by explicating low scale evolution (specifically concerning heterogeneous network mediated interactions), as opposed to aggregate dynamics based on homogeneous assumptions necessary to achieve analytical solvability.

Secondly, by allowing to explore more thoroughly a model, testing its limit behavior, assessing risks (and allowing for scenario testing) but also calculated high level key indicators evaluating resiliency for instance, which might require a high number of simulations.

5.2 Gathering and pre-processing data

We chose Paris as first use case for the global urbanization pilot.

These first simulations are based on data collected for Paris, particularly based on

- http://opendata.paris.fr/page/home/
- http://www.insee.fr/en/

5.2.1 General data

These data sources provide a set of interesting sets.

5.2.1.1 Administrative data

The administrative spatial description is provided at different scales, going from districts (“arrondissements”) (Figure 29) to municipalities (“communes”) (Figure 31) passing by intermediate areas (“quartiers”) (Figure 30), but also green areas (Figure 32).

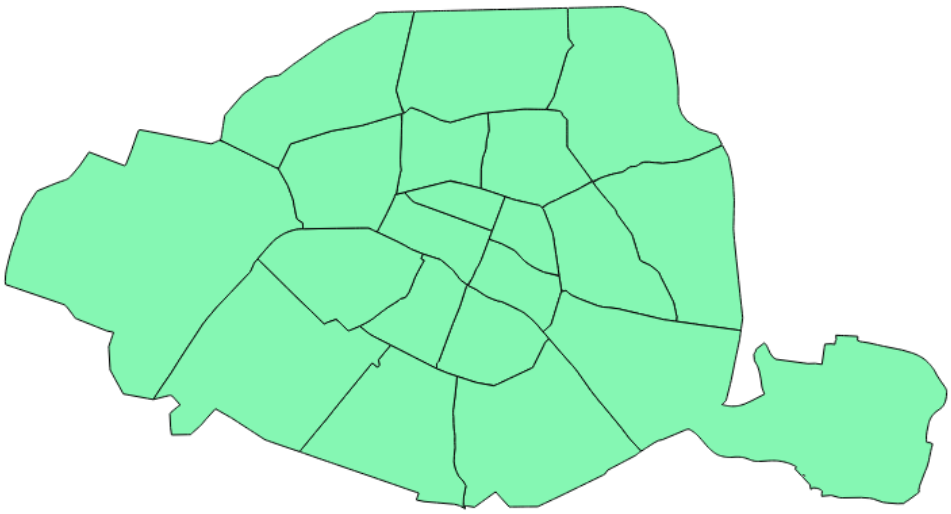

Figure 29: Paris intra-muros districts (arrondissements)

Figure 30: Paris intra-muros intermediate areas

Figure 31: Paris municipalities

We can also retrieve data linked to living proximity amenities, such as parks and gardens.

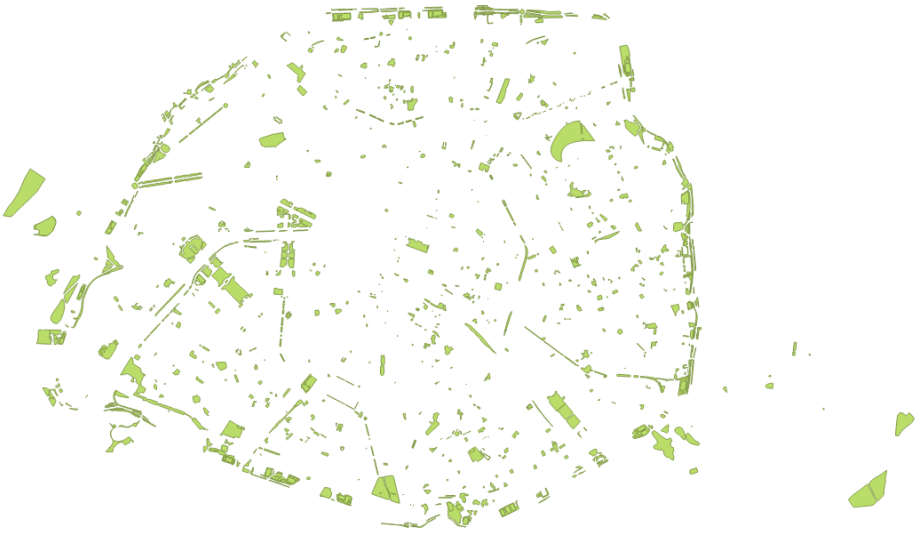

Figure 32: Parks and garden in Paris

5.2.1.2 Transport

Transport concerns roads (Figure 33), but also railways (Figure 34, Figure 35) and cycling paths.

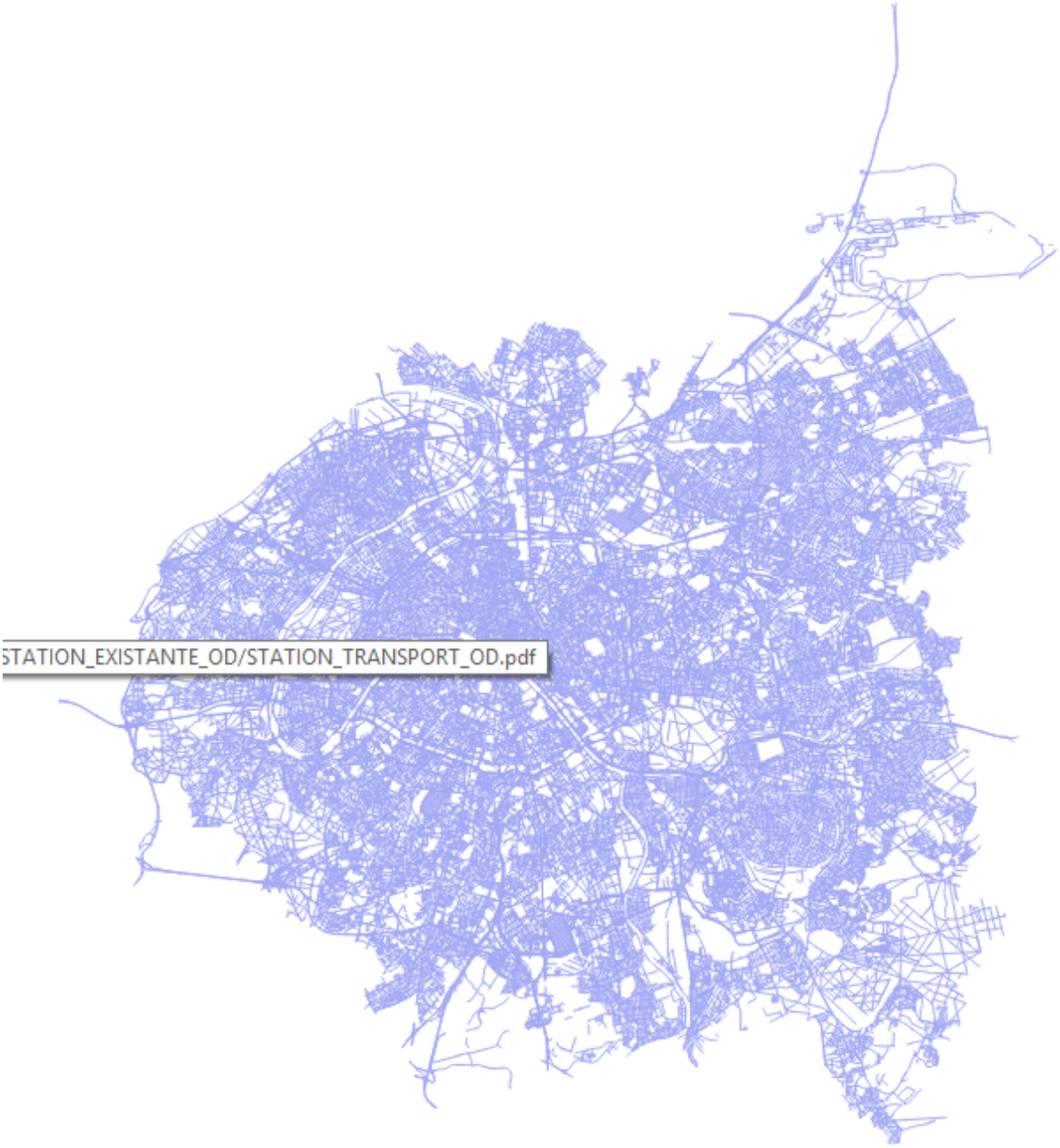

Figure 33: Paris streets and roads.

Figure 34: Paris railway lines

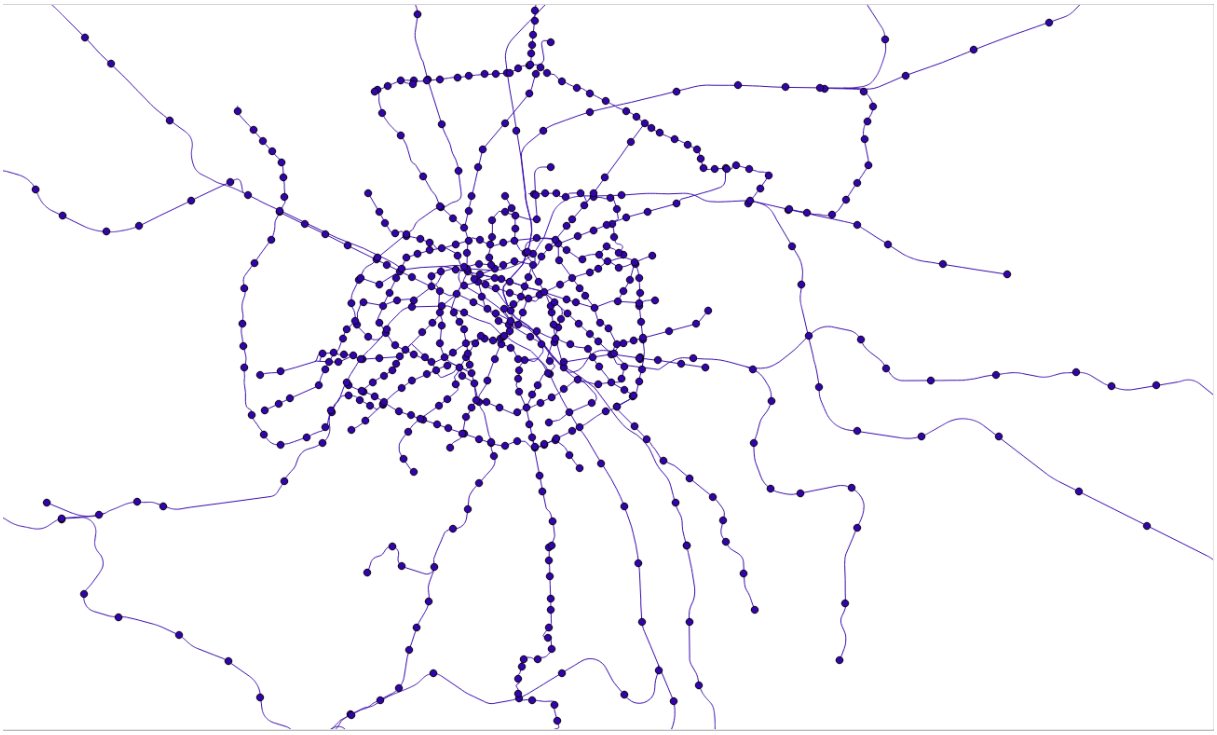

Figure 35: Paris railway lines and stations.

5.2.2 Real-estate pricing

We based our first simulations on the average real-estate pricing per district in Paris (Figure 36). Pricing goes from 6472 to 12688 Euro per square meter, with an average of 8619 Euro per square meter. It appears in the figure hereafter.

We can see that the real estate is quite typical of European cities, with high prices in the center and prices diminishing when moving to the periphery. However this gradient is not strict and following history but also current living conditions, some districts at a similar distance from the center appear to have different levels of pricing, accounting for natural heterogeneity in city pricing.

Consequently this first city can give us some first interesting insights on cities mixing a certain regularity of price gradient following the distance to the center (i.e. spatial a priori regularity) with a spatial heterogeneity, possibly arising also from city infrastructures.

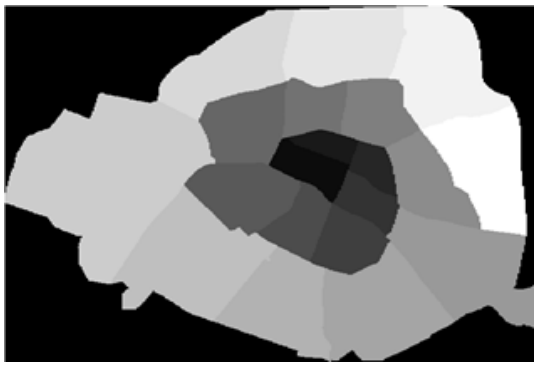

Figure 36: Real-estate pricing per district in Paris

5.2.3 Intra-muros commuting population (per district)

For these first simulations focusing on the influence of Paris intra-muros commuting on pollution and pricing, we considered only the intra-muros population living in one district and working in another one (see following paragraph).

Numbers of commuters per district (over the 20 districts) go from 1612 (in the very center) to 27910 (periphery), for a total number of 260465 intra muros commuters.

5.2.4 Intra-muros commuting between districts

We retrieved the commuting population from recent INSEE (French statistical institute) surveys (see Figure 37). The number of commuters between two districts goes from 101 to 4340 with an average of 749.

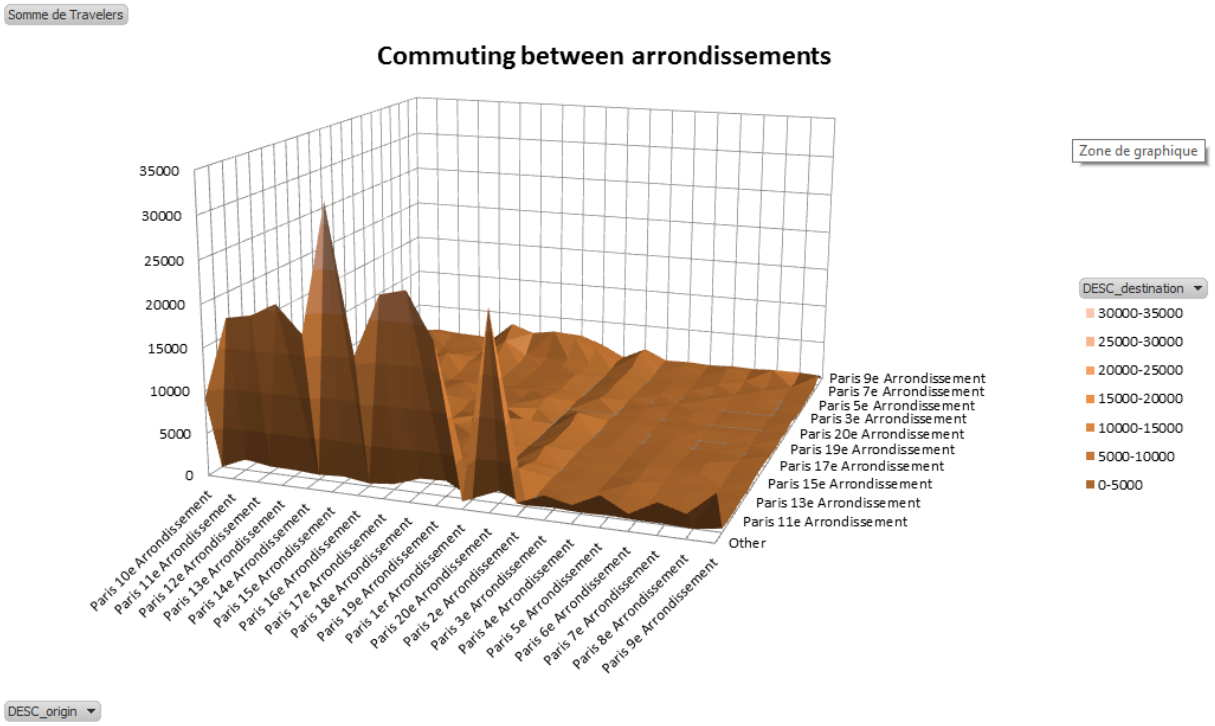

Figure 37: Commuting flows between the districts

5.2.5 Income

Income is also based on INSEE district (net) income data. Its median varies from 18334 Euro per year in the outskirts to 42250 Euro per year nearer to the center.

5.2.6 Car ownership in Paris

Recent surveys show that about half of inhabitants own a car in Paris. For this first model we have not detailed different kinds of vehicles.

5.2.7 Pollution

In this very simple first model, we have consequently not differentiated either the kind of car and taken average CO2 emissions (150 g of CO_2 per kilometer).

5.2.8 Technical and conceptual challenges

The global urbanization pilot, which is to study the two-way relationship between transport infrastructure and real estate pricing, requires data from different sources, potentially with different granularities and spatial subdivisions, if not topologies.

- Pollution is monitored at various measurement points.
- Transportation occurs along rail or road lines, linked into a transportation network
- Real estate pricing is mostly aggregated following administrative subdivisions.

Data can be available at different scales, following data collection facilities / opportunities and/or aggregation choices (for instance INSEE gathers population data following administrative rather than purely geographical subdivisions).

In this very simple first model we have simplified some of these difficulties as described below.

5.2.9 Data pre-processing

The real-estate data we found was averaged per district, which does not account for its spatial heterogeneity, so we used GIS tools to interpolate it over the city (into a 800 units grid) (Figure 38). Even if this interpolation is an approximation, it allows us to evaluate potential influence of data granularity on the model's results.

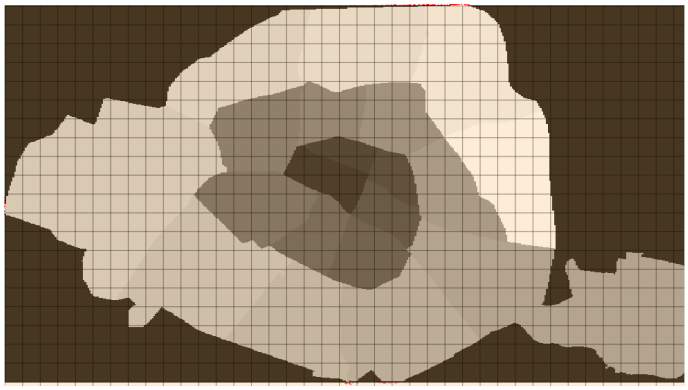

Figure 38: Refining spatial information onto a grid

We have particularly interpolated the value of real estate pricing and the income over the city (Figure 39, Figure 40).

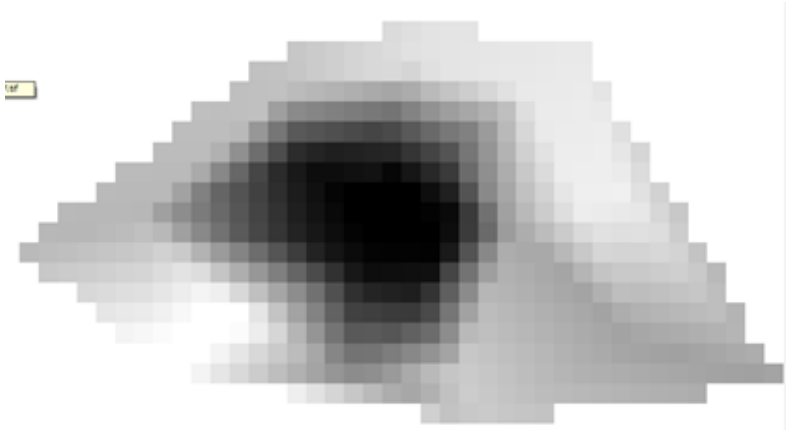

Figure 39: Real estate pricing interpolated to a grid (dark is high).

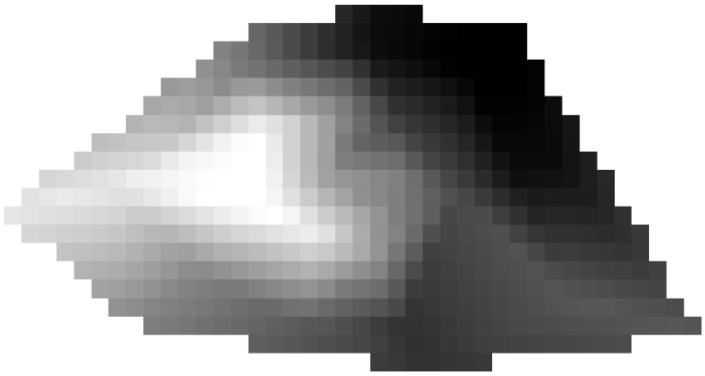

Figure 40: Income interpolated over a grid (light is high)

5.3 Conceptual model

5.3.1 Overview

Our conceptual model holds various elements corresponding to various themes (as foreseen following the specification document) (Figure 41).

- Transport
- Real estate
- Society
- Economy

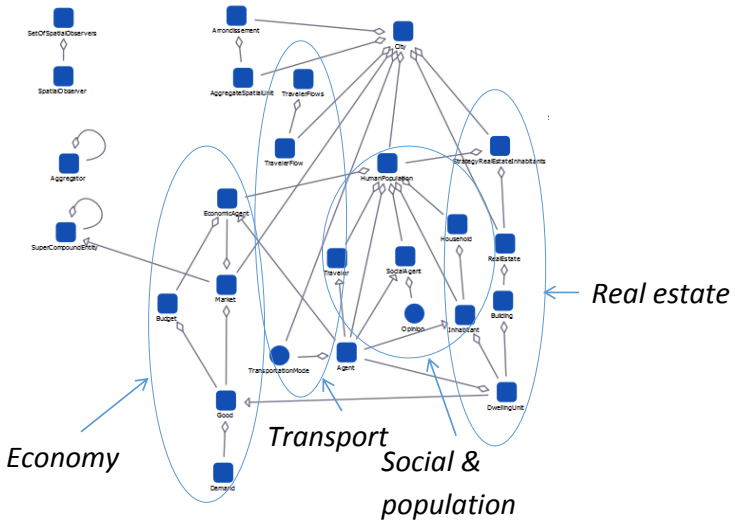

Figure 41: Global urbanization model holds elements corresponding to various themes.

Its scale goes from the city (at the highest level) to the agent (Figure 42).

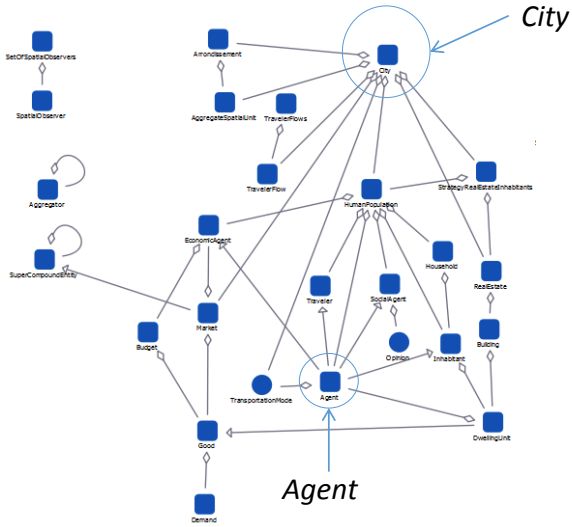

Figure 42: Global urbanization model has elements at different scales.

Between these two scales, it rests on various forms of aggregation (Figure 43), particularly :

- Spatial composition (an administrative unit aggregates a set of spatial units)
- Conceptual composition (for instance a population is composed of agents)

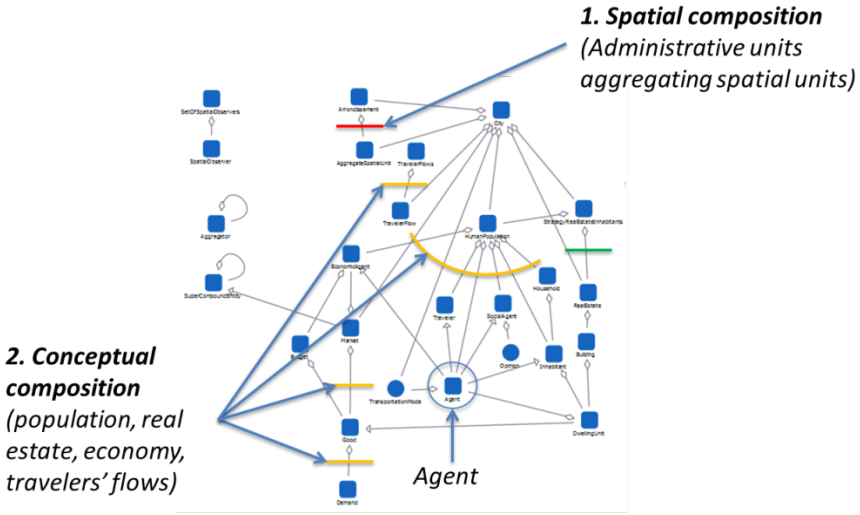

Figure 43: Urbanization model proposes spatial and conceptual forms of composition.

5.3.2 Conceptual model focus on studied dynamics

For the following study, we focus on specific dynamics of the model.

We observe agents who, by commuting between a dwelling unit and a workplace, generate more or less pollution following their use of their car or a green mode.

$$pollution_k = d_{residence,working\ place} \cdot \alpha_{mode}$$

Where α_{mode} is the emission linked to the transport mode. In this first simple model we have taken two possible values

$$\alpha_{car} = 150\ \mathrm{g\ CO_2\ /\ km}$$
$$\alpha_{green} = 0\ \mathrm{g\ CO_2\ /\ km}$$

The generated pollution influences real estate pricing. For every real estate unit *i* we define the following variables.

$$MinimumPrice_i = InitialPrice_i \cdot (1 - malus)$$
$$MaxmumPrice_i = InitialPrice_i \cdot (1 + bonus)$$

Where *InitialPrice*$_i$ is the initial real estate pricing of the real estate unit, and *bonus* and *malus* are parameters of the model. The pollution nuisance level is calculated as

$$NuisanceLevel_i = \frac{emissions_i - minimum\ (emissions)}{maximum(emissions) - minimum\ (emissions)}$$

The new price is then simply calculated as

$$NewPrice_i = MinimumPrice_i + (MaximumPrice_i - MinimumPrice_i) \cdot (1 - NuisanceLevel_i)$$

This leads to reallocate inhabitants following the new prices and their income.

The synthesized dynamics appear in Figure 44 hereafter.

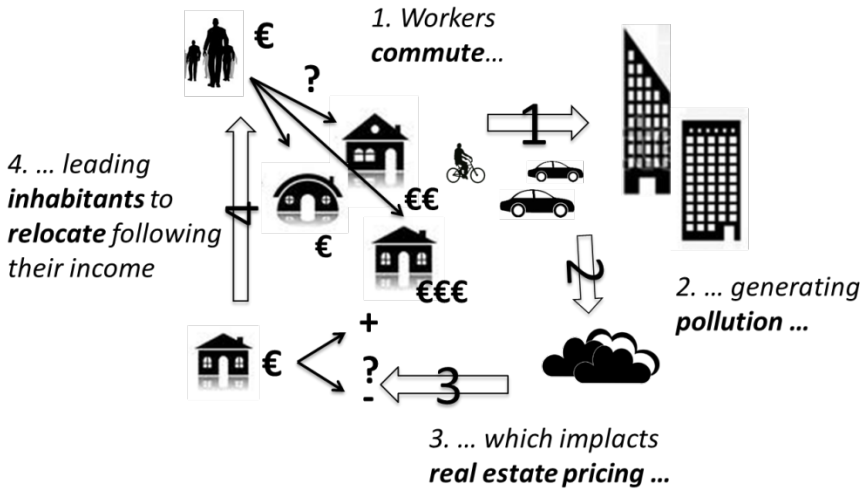

Figure 44: Synthetizing the principle dynamics in the first version of the model

5.4 Simulations

5.4.1 Scientific modeling and simulation questions

In these first simulations we aim at investigating various questions.

5.4.1.1 Model parameter values

When investigating a model, it is interesting to evaluate the influence of various key parameters on the dynamics of the model, echoing possible influence factors upon the evolution. These can be of various kinds, let's cite different examples.

- The value of objects' characteristics can allow studying a possible set of leverage (e.g. car pollution emissions which can be reduced by technical innovation or real-estate pricing limits which can be institutionally set).
- A level higher, behaviors (transport mode preferences and choices) can be influenced by pricing policies, incentives or enhancing awareness over information campaigns.
- Still a level higher, decisions can be influenced by an evolution of an overall perception, awareness and reflection (e.g. preferring residence in less polluted areas, due to long term health concerns).

These ways to influence the evolution can be more or less accessible

- Possibility of institutional (laws) action or not (opinions, free choice)
- Cost of information or incentives (in relation to evaluated benefit)
- Public / private sphere
- Objective information / Personal choices

5.4.1.2 Modelling choices

If modeling choices depend on the purpose of the model they can also be influenced by exterior limitations. We consider here more specifically data and model grains.

The data grain depends strongly on data available, particularly when observing the interdependent value of various characteristics for which surveys do not always exist (e.g. ecological involvement following working place location). There is therefore a major stake in assessing the possible influence of incomplete, inappropriately coarse, or partly uncertain data on a simulation.

The modeling grain is not only the modeler's choice. It can depend on theoretical limitations (finest possible modeling grain), but also on calculation time or available data. Our purpose is to evaluate how the choice of this modeling grain can influence the precision of the simulation describing the evolution of the model. Indeed showing how increased modeling grain can improve simulation precision might help put into light benefits of HPC / HPDA to tackle GSS modeling questions.

Finally these two questions can be crossed: is there any benefit in refining the modeling grain and evolution if initialization data are only coarse? How does a coarse grained evolution model take into account fine grained initialization data?

5.4.2 Purpose of the study

This first study aims at observing possible influence of

- Model parameters:
  - ***Travel behaviors*** (possibly depending on available public transport infrastructure): we study the influence of the *percentage of commuters* to take their car on the global dynamics (all, ~50%, ~25%)
  - ***Influence of pollution*** (and its perception) on prices (10% or 40%)
- Modeling choices:
  - ***Data granularity***, by comparing the results for real estate pricing initialized either per district or interpolated over a grid.
  - ***Agent granularity***: an agent standing for 1 or more (5) commuters.

5.4.3 Synthesis of results

Coming with some insights on the influence of various parameters on the model, we can observe the influence of data and agent granularity on the evolution of the simulation.

- Essential ***interactions*** appearing in the model results
  - Travel behaviors impact pollution
  - Pollution impacts more or less real estate pricing
  - Real estate pricing can influence in return locations and pollution
- ***Sensitivity of prices to pollution*** is essential in the dynamics
- Parameter values can lead or not to ***possible complex*** evolutions (retroaction and self-regulation)
- Depending on scenarios following ***modelling*** features have ***different levels of influence***

- Initial data coarseness (and / or uncertainty)
- Agent grain

5.4.3.1 *Influence of the public transport infrastructure / travel behaviours*

Our first set of results concerns the sensitivity of the evolution of the model to a 'simple' parameter, the percentage of commuters to prefer their car to green modes.

1. As expected, the higher the number of cars, the higher the pollution (Figure 45). (Detailed results show however that the level of pollution reached can depend both on data and model grain.)

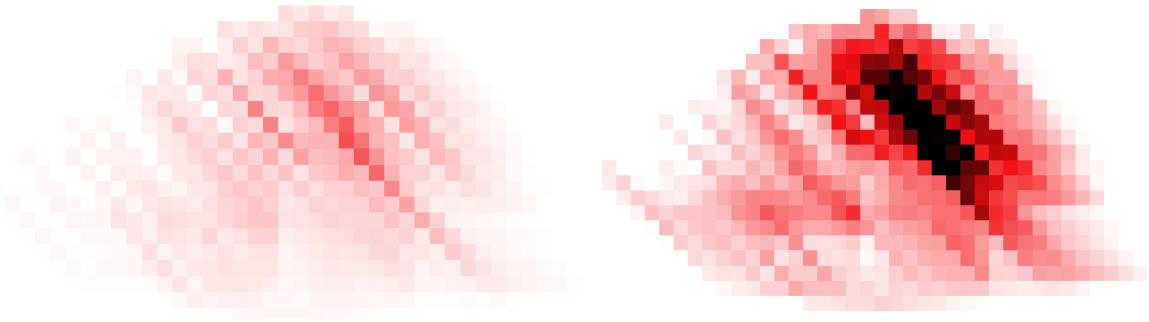

Figure 45: Pollution in scenarios: green versus all cars

2. However the observed impact on the real estate pricing depends on the value of the parameter defining the level of influence of pollution on real estate pricing.
   Indeed when pollution influences only weakly the real estate pricing, it seems to have little or even no impact.
   We show hereafter two compared sets of results: the first one (Figure 46), with a weak influence of pollution on prices, shows very similar outcomes of real estate pricing whether commuters choose green modes or not. The second one (Figure 47), on the contrary, shows how heavier car traffic, by generating more traffic and pollution, impacts negatively the prices.

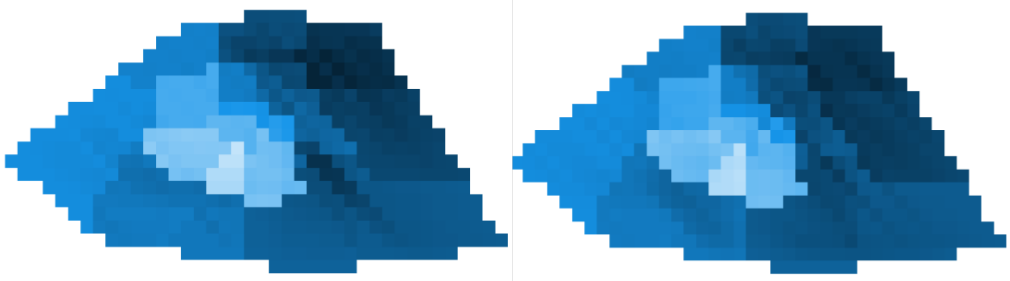

Figure 46: Real estate prices in scenarios : green versus all cars, lower influence of pollution on prices (10%)

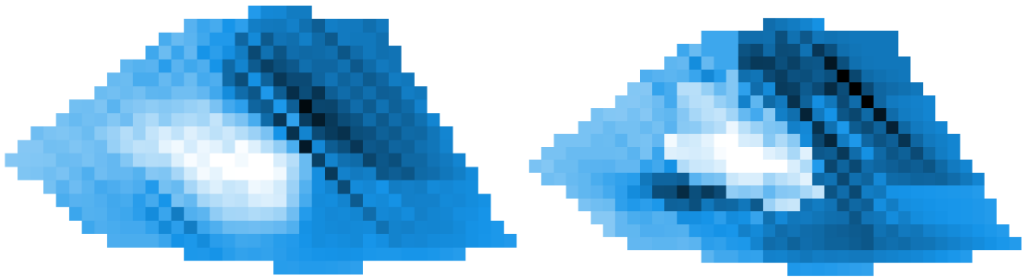

Figure 47: Real estate prices in scenarios: green versus all cars, higher influence of pollution on prices (40%)

5.4.3.2 Level of influence of pollution on prices

Let's now observe the impact of the value of a '*coupling*' parameter: the influence of pollution on real estate pricing. Its value influences not only the observed outcome concerning a single indicator (Figure 48), but more fundamentally the dynamics of the model (Figure 51), and its sensitivity to certain modeling choices (data grain) (Figure 49 and Figure 50).

1. *Higher sensitivity to pollution leads* ***prices to evolve more*** as it appeared in the previous section, and as appears clearly in the outcomes of the two scenarios hereafter.

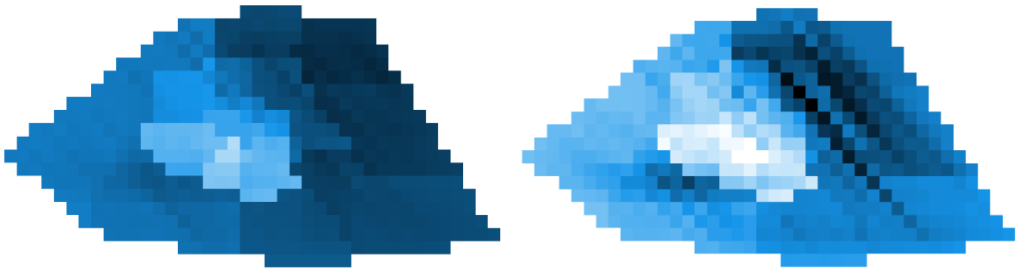

Figure 48: Real estate prices in scenarios: lower versus higher sensitivity of prices to pollution (same travel behaviors)

2. *Higher influence of pollution leads to* ***less sensitivity to initial values*** as a possible consequence.

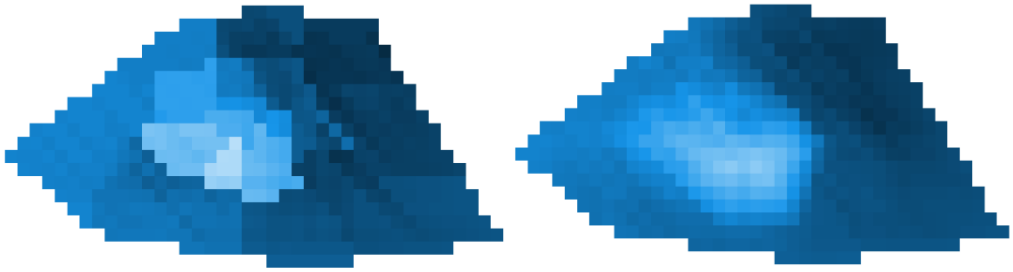

Figure 49: Real estate prices in scenarios : initialization per district vs interpolated, lower influence of pollution

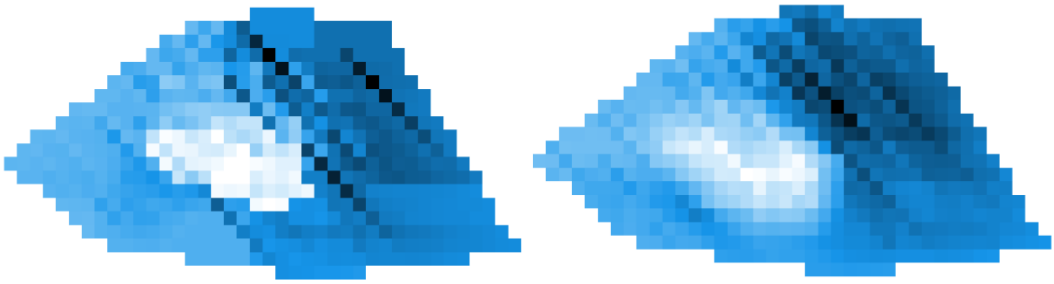

Figure 50: Real estate prices in scenarios: initialization per district vs interpolated, higher influence of pollution

3. A ***higher influence of the pollution on real estate pricing*** can help to ***reduce overall pollution***.

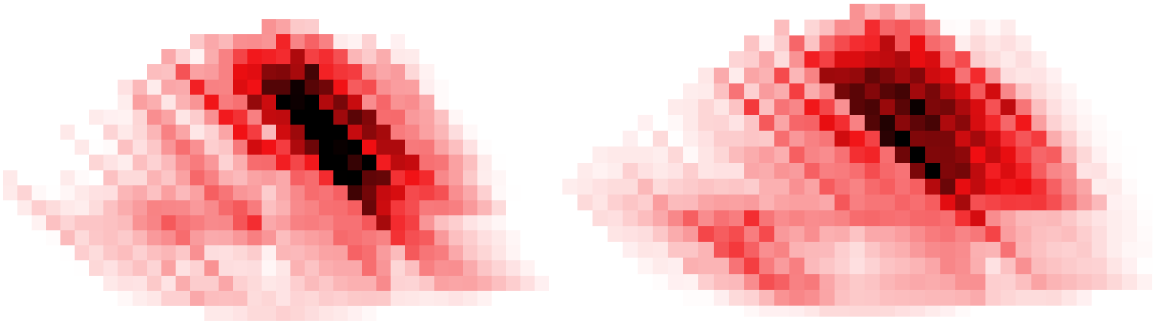

Figure 51: Pollution in scenarios: lower versus higher influence of pollution on real estate pricing

5.4.3.3 Initial values per district versus interpolated

If now we question the impact of modeling choices and more particularly how the initialization data grain influences the model evolution, we can observe different points. If a finer grain allows predictably to simulate (Figure 54) and observe (Figure 53) the evolution in a finer way, the importance of the grain of the initialization data varies following the scenario and the value of other parameters (a less clear influence appears in Figure 52). As previously stated, the sensitivity of the final state to the initial one is higher when pollution influences less real estate pricing.

1. Sometimes the final state ***varies little*** following initial state

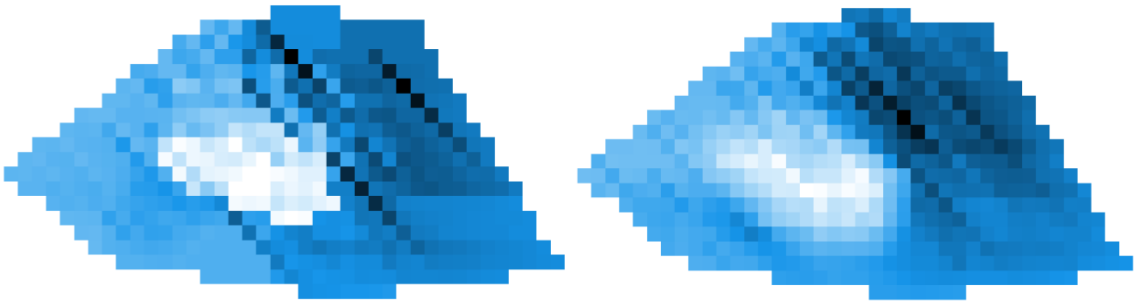

Figure 52: Real estate pricing, higher influence of pollution on prices, green behaviors, district vs interpolated initialization.

2. ***Observe finer*** evolution

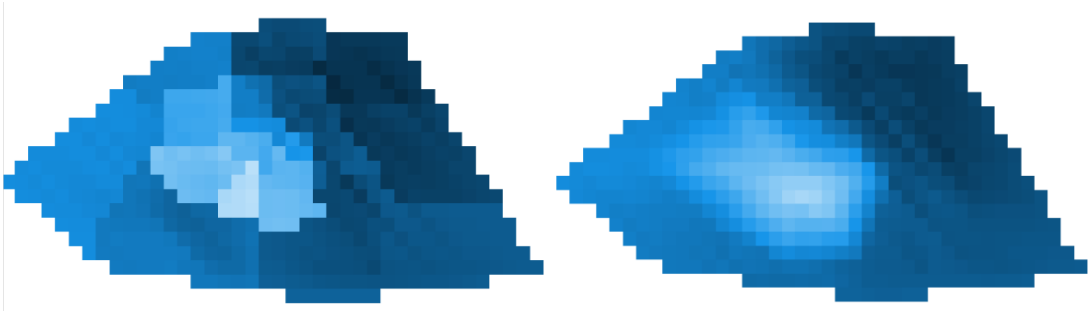

Figure 53: Real estate pricing: district vs interpolated initialization.

3. ***Model and simulate evolution in a finer way***

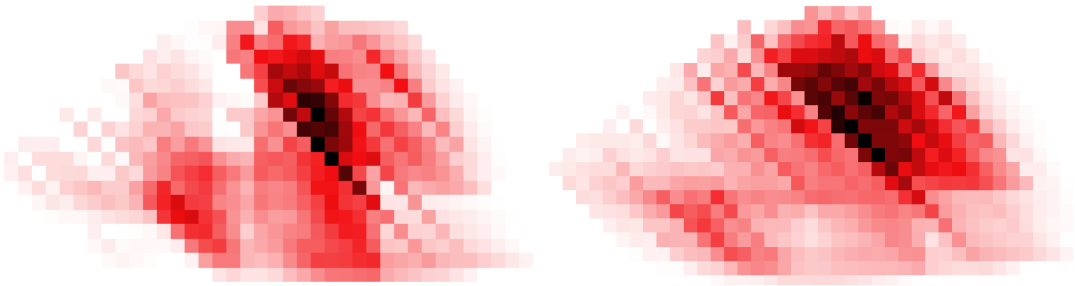

Figure 54: Pollution: district vs interpolated initialization.

5.4.3.4 Precision of agents

If we now investigate the impact of agent grain on the evolution, we can see here also that it varies following the scenarios: in some cases, simulating only aggregate agents can lead to similar results as with individual agents (Figure 55).

1. Sometimes no great difference appears

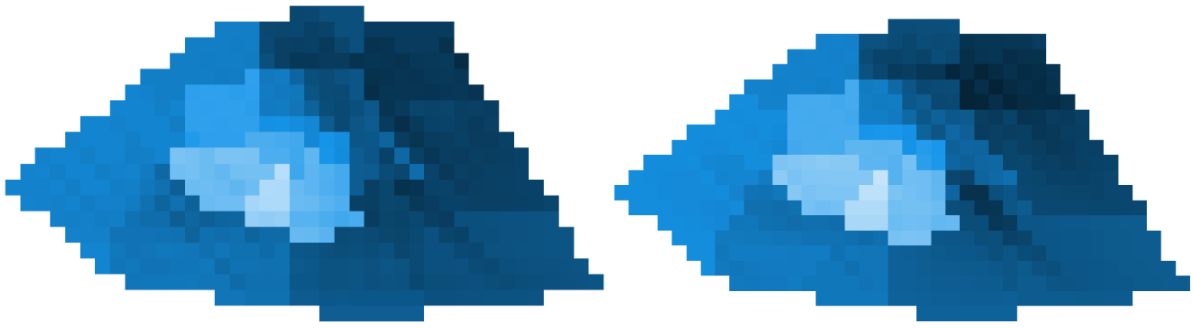

Figure 55: Real estate pricing: individual vs aggregate agents, green behaviours, lower influence of pollution, district initialization

However, in many cases refining the agent grain leads to potentially more precise results (Figure 56) and the evolution appears finer.

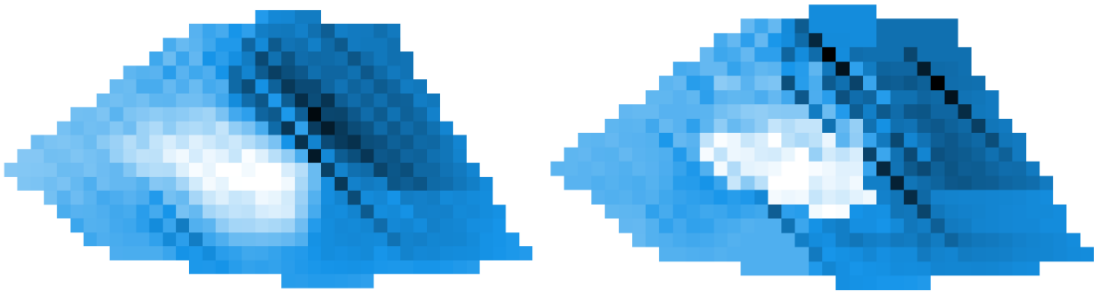

Figure 56: Real estate pricing: individual vs aggregate agents, green behaviours, higher influence of pollution, district initialization

6 Future Applications

The aim of this task is to identify needs and opportunities for future HPC applications in view of global systems. This section presents results from first scoping exercises (Section 6.1) and sketches two potential future applications that have been identified (Sections 6.2 and 6.3).

6.1 Scoping

This task is to play a scoping role for the European commission and the HPC community at large, using delphi methods both to partners of CoeGSS and to stakeholders and practitioners with whom CoeGSS interacts. In this line of work, a questionnaire was prepared and handed out both at the CoeGSS Kick-off meeting and at the Conference on Global Systems Science "Everything is Connected - Equilibrium and Disequilibrium in Social, Economic and Political Systems", Genoa, 28–30 October 2015. Unfortunately, the return rate of questionnaires was rather low (10 internal, 1 external), however, some interesting applications were pointed to. The results are summarised in the following:

The question "what are global challenges that can be addressed using HPC/HPDA?" provided a large variety of answers. Almost all-important global challenges were mentioned, which reveals an ongoing trend to provide answers to complex questions via computational simulations. Challenges about health care, climate, energy, migration and urbanization, food, water management, demographic participation, prediction of financial crises, pandemics, and autonomous systems basically cover all aspects of human life.

The question "why and how can HPC/HPDA be usefull in those contexts?" revealed a consistent pattern of linking different formerly separated challenges. Examples would be the connection between micro and macroeconomics by an agent-based model or urban development models that comprises a multitude of interacting simulations. Generally there is a trend that only the simulation of the technical or physical components (e.g., the simulation of the global warming) is not enough, but only the interconnection to the arising societal challenges completes the picture. Thus, HPC/HDPA is expected to provide more sophisticated analysis tools for simulations, to allow spatially explicit modelling, to handle computational-demanding statistical analysis even for large-scale modes and to provide a scalable, transparent and flexible middleware.

The question what is needed to apply HPC for the above mentioned challenges were straightforward. The demand for appropriate model and access to a larger amount of data was voiced most often. But also algorithms that can leverage the HPC power more efficiently and more comprehensive parallel libraries remain an issue to the community. Examples could be the prediction of the success of green growth initiatives. At last, the question about the connection between domain specific languages and compiler technologies raises an important issue of hiding the complex HPC infrastructure from the user.

Finally, the question to sketch potential business cases basically introduced global companies and policy makers and governmental agencies as customers. Three business scenarios were most prominent: providing such sophisticated simulation for policy making, providing software and tools to enable such simulations and to provide services and consulting for users.

Within the consortium, further scoping exercises will be carried out once a year. Once interaction with external stakeholders is set up and running well, scoping activities will be extended outside CoeGSS as well.

6.2 Financial contagion as a potential future application

As mentioned in deliverable D4.1, financial contagion presents a potential future application in that a synthetic information system could be a helpful tool for analysing systemic risks in financial markets and between finance and the real economy. However, fruitful applications of HPC to global challenges around the financial sector are not limited to constructing and using a full SIS for the financial system.

CoeGSS is in contact with a group of researchers around Prof. Stefano Battiston (ETH Zurich) in relation to their work on “A climate stress-test of the financial system” (Battiston *et al.*, 2016). This paper estimates the impact of climate policy risks on the financial system using a network analysis of the exposures of financial actors to climate-relevant sectors. Network analysis can benefit from HPC to deal with large and dense matrices representing these networks.

In estimating the impact of climate policy risks on the financial system, traditional risk analysis may lead to largely inaccurate calculations of expected losses or gains, due to intrinsic uncertainty of model estimates of climate policy effects. Through the network analysis approach, indirect effects through key economic sectors (e.g., energy-intensive sectors and housing) are taken into account. The methodology can be applied both bottom-up using micro-economic data and top-down using aggregate macro-economic sectoral data.

The above mentioned work uses empirical data of the Euro Area to show that while direct exposures to the fossil fuel sector are small (3-12%), the combined exposures to climate-policy relevant sectors are large (40-54%), heterogeneous, and possibly amplified by indirect exposures via financial counterparties (30-40%). Consequences suggested by these results, such as the fact that climate policies could result in potential winners and losers across financial actors and would not have adverse systemic impact as long as they are implemented early on and within a stable framework, constitute valuable information for actors in the climate policy arena.

Moving to micro-data, if these are available, would increase matrix sizes in the network analysis conducted, and make the use of HPC an interesting tool. Also, adding other types of links between nodes in the network, to represent further indirect channels of financial

exposure in addition to the equity holdings and lending between banks considered so far, will make the model more complex and thus the use of HPC beneficial. In the coming second project year, CoeGSS intends to explore this potential future application together with the researchers involved in the above-described work.

6.3 The blockchain technology in GSS as a potential future application

Today, the Internet provides totally new environments where people can meet and exchange ideas, feelings, money and services. Various collective (bottom up) initiatives, such as Open Source Software, Wikipedia, and other collaborative systems that do not require decisions passing from centralized entities, suggest a new model of organization for solving pressing global problems in a distributed fashion between citizens. Such problems include sustainability of energy production and use and sustainability of financial institutions (pension funds, credit to companies) to protection of personal data.

In such decentralised environments, one challenge is to assess the reputation of the counterpart in a transaction and also to make your own reputation available whenever necessary without certification from a central authority. The blockchain technology, which creates a distributed structure made secure by a shared cryptographic protocol, can play an important role here to help tackle GSS challenges and provide an improved possibility for people to exercise their rights as citizens and investors. Since the blockchain technology creates a distributed and decentralized database formed by a continuously growing list of records protected against fraudulent revision “as long as the computing power of honest people is larger than that of the hackers” (nakamoto2008) HPC has an obvious role to play here as well. CoeGSS intends to explore the potential of the blockchain technology as a future HPC-GSS application.

7 References

Ahmad, S. (2005) 'Increasing excise taxes on cigarettes in California: a dynamic simulation of health and economic impacts', *Preventive Medicine*, 41(1), pp. 276–283. doi: http://dx.doi.org/10.1016/j.ypmed.2004.10.024.

Anderson, K. M., Odell, P. M., Wilson, P. W. and Kannel, W. B. (1991) 'Cardiovascular disease risk profiles.', *Am Heart J*. National Heart, Lung, and Blood Institute, Framingham, MA., 121(1 Pt 2), pp. 293–298.

Armitage, C. J. and Arden, M. A. (2002) 'Exploring discontinuity patterns in the transtheoretical model: An application of the theory of planned behaviour', *British Journal of Health Psychology*. Blackwell Publishing Ltd, 7(1), pp. 89–103.

Arrow, K. (1971) *Essays in the Theory of Risk Bearing*. Markham Publishing Co.

Balk D., Yetman, G. (2004) 'The Global Distribution of Population: Evaluating the gains in resolution refinement', *Center for International Earth Science Information Network*.

Bandura, A. (1977) *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall. Available at: http://www.bibsonomy.org/bibtex/2ac8005ab3acbb005a8481cffeaba453d/tobidiplom.

Forey, B., Hamling, J., Lee, P., and Wald, N. (eds) (2002) *International Smoking Statistics: A Collection of Historical Data from 30 Economically Developed Countries (Oxford Medical Publications)*. Oxford University Press.

Battiston, S., Mandel, A., Monasterolo, I., Schütze, F. and Visentin, G. (2016) 'A Climate Stress-Test of the Financial System'. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2726076

Beheshti, R. and Sukthankar, G. (2014) 'A Normative Agent-based Model for Predicting Smoking Cessation Trends', in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems (AAMAS '14), pp. 557–564.

Bicchieri, C. (2005) *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z. and van der Torre, L. (2001) 'The BOID Architecture: Conflicts Between Beliefs, Obligations, Intentions and Desires', in *Proceedings of the Fifth International Conference on Autonomous Agents*. New York, NY, USA: ACM (AGENTS '01), pp. 9–16.

Caldarelli, G., Capocci, A., De Los Rios, P. and Muñoz, M. A. (2002) 'Scale-free networks from varying vertex intrinsic fitness.', *Physical review letters*, 89(25), p. 258702.

Castellano, C., Fortunato, S. and Loreto, V. (2009) 'Statistical physics of social dynamics', *Rev. Mod. Phys.* American Physical Society, 81(2), pp. 591–646.

Centola, D. (2011) 'An experimental study of homophily in the adoption of health behavior', *Science*. American Association for the Advancement of Science, 334(6060), pp. 1269–1272.

Chao, D. L., Halloran, M. E., Obenchain, V. J. and Longini Jr, I. M. (2010) 'FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model', *PLoS Comput Biol*. Public Library of Science, 6(1), pp. 1–8.

Christakis, N. and Fowler, J. H. (2007) 'The Spread of Obesity in a Large Social Network over 32 Years', *New England Journal of Medicine*, 357(4), pp. 370–379.

Cimini, G., Squartini, T., Garlaschelli, D. and Gabrielli, A. (2015) 'Systemic Risk Analysis on Reconstructed Economic and Financial Networks.', *Scientific reports*, 5, p. 15758.

Cimini, G., Squartini, T., Musmeci, N., Puliga, M., Gabrielli, A., Garlaschelli, D., Battiston, S. and Caldarelli, G. (2015) 'Reconstructing Topological Properties of Complex Networks Using the Fitness Model', *Lecture Notes in Computer Science*, 8852, pp. 323–333.

Collier, N., Ozik, J. and Macal, C. M. (2015) 'Large-Scale Agent-Based Modeling with Repast HPC: A Case Study in Parallelizing an Agent-Based Model', in Hunold, S., Costan, A., Giménez, D., Iosup, A., Ricci, L., Gómez Requena, E. M., Scarano, V., Varbanescu, L. A., Scott, L. S., Lankes, S., Weidendorfer, J., and Alexander, M. (eds) *Euro-Par 2015: Parallel Processing Workshops: Euro-Par 2015 International Workshops, Vienna, Austria, August 24-25, 2015, Revised Selected Papers*. Cham: Springer International Publishing, pp. 454–465.

Dargay, J., Gately, D. and Sommer, M. (2007) 'Vehicle Ownership and Income Growth, Worldwide: 1960-2030', *The Energy Journal*, 28(4), pp. 143–170.

Deutsche Automobil Treuhand GmbH (2015) 'DAT-Report 2015'.

Deutsche Automobil Treuhand GmbH (2016) 'Highlights aus dem DAT-Report 2016'.

DiClemente, C. C. and Prochaska, J. O. (1982) 'Self-change and therapy change of smoking behavior: A comparison of processes of change in cessation and maintenance', *Addictive Behaviors*, 7(2), pp. 133–142.

Epstein, J. M. (1999) 'Agent-based computational models and generative social science', *Complexity*. John Wiley & Sons, Inc., 4(5), pp. 41–60.

Epstein, J. M. (2014) *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science* . Princeton University Press.

Garlaschelli, D. and Loffredo, M. I. (2008) 'Maximum likelihood: Extracting unbiased information from complex networks', *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(1), pp. 1–5.

Göhlmann, S. (2007) *The Determinants of Smoking Initiation - Empirical Evidence for Germany*. techreport.

Granovetter, M. S. (1973) 'The Strength of Weak Ties'', *American Journal of Sociology*, 78(6), pp. 1360–1380.

Gualdi, S., Cimini, G., Primicerio, K., Di Clemente, R. and Challet, D. (2016) 'Statistically similar portfolios and systemic risk', http://arxiv.org/abs/1603.05914.

Holme, P. and Saramäki, J. (2012) 'Temporal networks', *Physics Reports*, 519, pp. 97–125.

International Monetary Fund (2016) 'World Economic Outlook, GDP per capita based on purchasing-power-parity (PPP) in current international dollars'. https://www.imf.org/external/pubs/ft/weo/2016/01/weodata/weoselgr.aspx

Jarvis, A., Vincze, M. P., Falconer, B., Garde, A., Geber, F., and Daynard, R. (2008) 'A Study on Liability and the Health Costs of Smoking', *DG SANCO*.

Karni, E. (2005) 'Savage's Subjective Expected Utility Model'.

Kreps, D. (1988) *Notes on the Theory of Choice*. Westview Press.

Lang, J. C., Abrams, D. M. and Sterck, H. De (2015) 'The influence of societal individualism on a century of tobacco use: modelling the prevalence of smoking', *BMC Public Health*, 15(1), pp. 1–13.

Lee, R. D. and Carter, L. R. (1992) 'Modeling and Forecasting U. S. Mortality', *Journal of the American Statistical Association*, 87(419), pp. 659–671.

Levy, D. T., Bauer, J. E. and Lee, H. (2006) 'Simulation Modeling and Tobacco Control: Creating More Robust Public Health Policies', *American Journal of Public Health*. American Journal of Public Health 2006, 96(3), pp. 494–498.

Levy, D. T., Cummings, K. M. and Hyland, A. (2000) 'A simulation of the effects of youth initiation policies on overall cigarette use.', *American Journal of Public Health*, 90(8), pp. 1311–1314.

Levy, D. T. and Friend, K. B. (2000) 'A simulation model of tobacco youth access policies', *Journal of Health Politics, Policy and Law*. Duke Univ Press, 25(6), pp. 1023–1050.

Luck, M., Mahmoud, S., Meneguzzi, F., Kollingbaum, M., Norman, T. J., Criado, N. and Fagundes, M. S. (2013) 'Normative Agents', in Ossowski, S. (ed.) *Agreement Technologies*. Dordrecht: Springer Netherlands, pp. 209–220.

Mendez, D. (2011) 'Results from a Population Dynamics Model of the Consequences of Menthol Cigarettes for Smoking Prevalence and Disease Risks', *Appendix A of the Tobacco Products Science Advisory Committee MentholReport*.

Mendez, D., Warner, K. E. and Courant, P. N. (1998) 'Has Smoking Cessation Ceased? Expected Trends in the Prevalence of Smoking in the United States', *American Journal of Epidemiology*, 148(3), pp. 249–258.

Merler, S. and Ajelli, M. (2010) 'The role of population heterogeneity and human mobility in the spread of pandemic influenza', *Proceedings of the Royal Society B: Biological Sciences*. The Royal Society, 277(1681), pp. 557–565.

Murphy, J. T. (2011) 'Computational Social Science and High Performance Computing: A Case Study of a Simple Model at Large Scales'. https://computationalsocialscience.org/wp-content/uploads/2011/10/JTMurphy.pdf

Murphy, J.T. (2014) 'High Performance Agent-Based Modeling in Repast HPC', *Computation Institute, Chicago*. https://www.ci.uchicago.edu/events/high-performance-agent-based-modeling-repast-hpc.

OICA (2015) 'Vehicles in use'. http://www.oica.net/category/vehicles-in-use/

Park, J. and Newman, M. E. J. (2004) 'Statistical mechanics of networks', *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 70(6 2).

Przywara, B. (2010) *Projecting future health care expenditure at European level: drivers, methodology and main results*. https://ideas.repec.org/p/euf/ecopap/0417.html

Rubio-Campillo, X. (2014) 'Pandora: A Versatile Agent-Based Modelling Platform for Social Simulation', in *SIMUL 2014: The Sixth International Conference on Advances in System Simulation*.

Saracco, F., Cimini, G., Quattrociocchi, W. and Squartini, T. (2016) 'Inferring communities of Facebook users by Likes on different argument posts', *In preparation*.

Saracco, F., Di Clemente, R., Gabrielli, A. and Squartini, T. (2016) 'Inferring monopartite projections of bipartite networks: an entropy-based approach', pp. 1–20. http://arxiv.org/abs/1607.02481

Schuhmacher, N., Ballato, L. and van Geert, P. (2014) 'Using an Agent-Based Model to Simulate the Development of Risk Behaviors During Adolescence', *Journal of Artificial Societies and Social Simulation*, 17(3), p. 1.

Schwartz, S. (2006) 'A Theory of Cultural Value Orientations: Explication and Applications', *Comparative Sociology*, 5(2), pp. 137–182.

Schwarzer, R. (2008) 'Modeling Health Behavior Change: How to Predict and Modify the Adoption and Maintenance of Health Behaviors', *Applied Psychology*. Blackwell Publishing Ltd, 57(1), pp. 1–29.

Sharomi, O. and Gumel, A. B. (2008) 'Curtailing smoking dynamics: A mathematical modeling approach', *Applied Mathematics and Computation*, 195(2), pp. 475–499.

Socioeconomic Data and Applications Center (SEDAC) (2016) 'Gridded Population of the World, v4'. http://sedac.ciesin.columbia.edu/data/collection/gpw-v4

Squartini, T., Caldarelli, G. and Cimini, G. (2016) 'Stock markets reconstruction via entropy maximization driven by fitness and density'. http://arxiv.org/abs/1606.07684

Squartini, T., Fagiolo, G. and Garlaschelli, D. (2011) 'Randomizing world trade. I. A binary network analysis', *Physical Review E*, 84(4), p. 46117.

Thun, M. and Myers, D. (1997) 'Age and the exposure-response relationships between cigarette smoking and premature death in Cancer Prevention Study II', *Smoking and Tobacco Control Monograph No. 8 - Changes in Cigarette-Related Disease Risks and Their Implications for Prevention and Control*. Citeseer, (Chapter 4), pp. 383–413.

U.S. Department of Commerce, Census Bureau (2014) 'Selected Economic Characteristics, 2006-2010 American Community Survey'.

U.S. Department of Transportation, Bureau of Transportation Statistics (2011) 'National Transportation Statistics, Table 1-17 - New and Used Passenger Car Sales and Leases'.

U.S. Energy Information Administration (2016) 'International Energy Outlook 2016, World GDP per capita by region expressed in purchasing power parity'.

United Nations Department of Economic and Social Affairs (2014) 'World Urbanisation Prospects, the 2014 Revision, File 21: Annual Percentage of Population at Mid-Year Residing in Urban Areas by Major Area, Region and Country, 1950-2050'.

United Nations Department of Economic and Social Affairs (2015) 'World Population Prospects, the 2015 Revision'.

Verzi, S. J., Brodsky, N. S., Brown, T. J., Apelberg, B. and Rostron, B. (2012) 'An agent-based approach for modeling population behavior and health with application to tobacco use', *Sandia National Laboratories*.

Voelcker, J. (2014) '1.2 Billion Vehicles On World's Roads Now, 2 Billion By 2035: Report'. http://www.greencarreports.com/news/1093560_1-2-billion-vehicles-on-worlds-roads-now-2-billion-by-2035-report

Wadud, Z., MacKenzie, D. and Leiby, P. (2016) 'Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles', *Transportation Research Part A: Policy and Practice*, 86, pp. 1–18.

Wheeler, L. (1966) 'Toward a theory of behavioral contagion', *Psychological Review*, 73(2), pp. 179–192.

World Bank (2015) 'GDP per capita (current US$)' http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

World Bank (2015a) ‘Land area (sq. km)’
http://data.worldbank.org/indicator/AG.LND.TOTL.K2