D4.5 – Second 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	30.09.2017
Actual date of submission	09.11.2017
Dissemination level	Public
Nature	Report
Version	3 (after internal review)
Work Package	4
Lead beneficiary	GCF
Responsible scientist/administrator	Sarah Wolf
Contributor(s)	Andreas Geiges, Steffen Fürst, Enrico Ubaldi, Margaret Edwards, Jette von Postel
Internal reviewers	Ralf Schneider, Patrik Jansson
Keywords	Health Habits, Green Growth, Global Urbanisation
Total number of pages:	77

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://coegss.eu/

Version History

	Name	**Partner**	**Date**
From	Sarah Wolf	GCF	
First Version	for internal review		09/2017
Second Version	for second review		10/2017
Third Version	for submission		11/2017
Reviewed by	Patrik Jansson Ralf Schneider	Chalmers HLRS	09/2017 10/2017
Approved by	ECM	UP, HLRS, ATOS	11/2017

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABM	Agent-Based Model
CES	Constant elasticity of substitution
CoeGSS	Centre of Excellence for Global System Science
CO_2	Carbon dioxide
EU	European Union
GDP	Gross Domestic Product
GIS	Geographic Information System
GSS	Global Systems Science
HDF5	Hierarchical Data Format 5 (a smart data container)
HLRS	High-Performance Computing Centre Stuttgart (a partner in CoeGSS)
HPC	High Performance Computing
HPDA	High Performance Data Analytics
ID	Identifier
IEA	International Energy Agency
I/O	Input/output
LAD	Local Authority District
LSE	Least square error
MIDAS	Models of Infectious Disease Agent Study
MoTMo	Mobility Transition Model
MPI	Message Passing Interface
MTMs	methods, tools and mechanisms
NBH	Niedersachsen, Bremen and Hamburg (federal states of Germany)
NDA	Non-disclosure agreement
NHS	National Health Service
NUTS	Nomenclature d'Unités Territoriales Statistiques
ONS	Office for National Statistics
pse	Parameter space exploration package of R
PV	Photovoltaics
SEDAC	Socioeconomic Data and Applications Center
SIS	Synthetic Information System
UK	United Kingdom
WP	Workpackage

Abstract

This deliverable presents the status of the three pilot studies of the Centre of Excellence for Global Systems Science – Health Habits, Green Growth, and Global Urbanization – at the end of the second 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 made in the second project year is presented for the synthetic information system of each pilot, and for the Future Applications task that completes WP4.

1 Introduction

This deliverable presents the progress made in work package 4 (WP4) of the Centre of Excellence for Global Systems Science (CoeGSS) throughout the second project year. The main focus is on the three pilot studies: T4.1 Health Habits, T4.2 Green Growth, and T4.3 Global Urbanisation; a further task in the work package is T4.4 Future Applications.

The emerging research field of Global Systems Science (GSS) combines data-driven simulation modelling with engagement of stakeholders and citizens to provide decision support by developing understanding and evidence on global challenges. CoeGSS brings together GSS and the power of High Performance Computing (HPC) and High Performance Data Analytics (HPDA). The pilot studies address three selected global challenges – respectively, the global epidemics of smoking, the diffusion of electric vehicles in the global car fleet, and the two-way relationship between transport infrastructure and real estate pricing – with a two-fold purpose:

- to derive requirements from pilots – as prototypical applications – to steer the evolution of the centre, and
- to develop success stories / examples for turning global challenges into business and policy opportunities.

First, in deriving requirements for GSS modelling and simulation on HPC, the pilots look for structural similarities in the work, so that the workflow, methods, or tools can be applied to more than one case, and hence to potential future applications as well. We refer the reader to the requirement deliverables D4.1, D4.2, and later this year also D4.3 for details. However, also specificities of single pilots are dealt with. For example, the Global Urbanisation pilot builds on pre-existing proprietary code, wherefore non-disclosure agreements (NDAs) have been set up.

Second, the development of success stories for global challenges with the help of HPC points out the practical relevance of research and development for merging HPC and GSS. Here, pilots aim at confirming the conjecture that GSS can obtain better results with the help of HPC and HPDA. These allow to use higher resolution data sets, to develop larger scale models (both in terms of complexity and in terms of geographical scale) and to analyse larger sets of output data from ensemble simulation runs more deeply. The pilots explore how GSS can benefit from these opportunities.

To these two ends, the pilots are developing HPC-based synthetic information systems for their three example global challenges. On synthetic information systems, see D4.1 and D4.4; very briefly, these are agent-based models initialised with synthetic populations for observing potential evolutions of the system under study with the help of model simulation runs. Progress made on these is described in the pilot sections (2 - 4), after the remainder of this section provides more detail on the motivation behind merging HPC and GSS from the pilots' perspective. Then, Section 5 reports on the task for future GSS-HPC applications, before Section 6 concludes.

1.1 Motivation

Earlier this year, when the EU celebrated the 60th anniversary of the Treaties of Rome, it was pointed out that High Performance Computing is an essential tool for addressing societal challenges such as climate change or health, and at the same time that the related research and innovation challenges can drive the future development of HPC in Europe (Digital Day 2017; Viola and Smits 2017). D4.4 has outlined expected benefits of using HPC for GSS, that can be summarised as "getting a closer look" (in more detailed models) and "getting a wider overview" (in deeper analysis of potential dynamics simulated) on a global system. These ideas shall be deepened here to point out mutual potential benefits of merging HPC with GSS.

Global systems are composed of many agents interacting in complex network structures embedded in a shared environment. Several features of global systems imply that aggregate models cannot generally replace models at the individual level:

- networks between agents can have an important influence on the modelled outcome
- effects to be modelled may concern a small subgroup of a population, as for example an initially small group of adopters in addressing the question whether a certain "green" innovation can trigger a sustainability transition
- path dependency can lead to differing outcomes resulting from small changes at the micro-level of the system

A closer look at the system is provided in agent-based modelling, that helps not to "aggregate away" such important elements in the evolution of global systems. Agent-based models (ABMs) can be compared with models of molecular dynamics, where many particles interact, for which HPC methods, tools and codes exist. However, the following differences make research and development between HPC and GSS necessary:

- heterogeneity of different types of agents needs to be accounted for,
- heterogeneous types of data are resolved in a spatial and individual data structure; a synthetic population can be viewed as a "social coordinate system" (e.g., Marathe 2017)
- social networks between agents are not regular grids; they go beyond spatial proximity relations, come with a set of network properties such as heavy tailed degree distributions, hierarchical structures and assortativity, and the networks themselves co-evolve over time with the agents' actions, and
- the environment in which agents interact is spatially differentiated and also co-evolves with the agents' actions; for example there may be an interplay between agents' actions and niches the agents are located in.

To obtain a wider overview on potential future evolutions of the thus composed global systems, the models require thorough exploration due to

- Complexity and unpredictability of modelled processes, particularly
  - Human decision, coming with a level of uncertainty and unpredictability

  - Uncertainty concerning value distribution and variability of some feature (e.g. initial state, individual parameter value), sometimes mimicked over stochasticity, requiring replications and analysis of not one but a set of simulations as such.

- Number of parameters, particularly when broken down to the level of individual agents, often leading to non-linear influence and unexpected combined effects, leading to potentially high dimensional parameter spaces to be explored.
- Tailoring to a specific modelling purpose and questions, which usually evolve.

Steps in model exploration include

- Initially, calibration or adjusting parameter values following available data.
- Exploring simulations to understand all the practical implications of modelling choices, including under unusual conditions (e.g. evolution at boundary conditions, crises, …) and to validate a model against given specifications[1].
- Finally, when simulating the validated model: to test various possible scenarios (and different influence factors) of interest to stakeholders (including risk assessment, for instance).

This requires ensembles of many stochastic simulation runs that produce large amounts of output data needing to be analysed and visualised. However, every simulation is time and resource costly. Here, a next step in pilot work shall be to explore how HPC can play a role by providing support for intelligent parameter space exploration (see Section 6.1). When many parallel simulation runs can communicate in order to interactively zoom in on areas of interest in huge parameter spaces, HPC can provide advantages over purely parallel, non-communicating execution of simulation ensembles. In GSS, such areas of interest will often be "turbulent zones" in the sense of rapid changes in a social system. Such change may be desired in some cases, as for example when looking for policy measures that can induce a transformation to sustainable mobility. In other cases, the goal will be to steer the system in a safe distance from turbulent zones, as for example an epidemics taking off.

Some fields in GSS are already using individual-level models; for example, epidemiology, to represent statistically correct dynamics of encounters between individuals for analysing the spread of a virus. For other topics such models are being developed. The three pilots of CoeGSS provide three examples. The process of model development in GSS is iterative: the stages of model definition, data collection, implementation, testing, running and analysing simulations (see also D4.4, Section 2.3) are carried out in repeated loops. The CoeGSS pilots have begun to include HPC in this loop. Tools that facilitate GSS modelling and analysis work on HPC, chosen or developed based on the experience made in this process, will support further iterations towards more powerful, larger scale global systems models. Attracting other

[1] Thus, specification itself is an important part of simulation based science, to have something to validate against (Ionescu and Jansson 2013)

global systems scientists and modellers, the development of such tools could open up a new field of HPC users.

A longer-term effort in research and development is needed to produce detailed large-scale models for many global challenges, and it can only be approached by GSS and HPC together. In this effort, typical structures in global systems described above may help shape future developments in HPC and HPDA. The complexity of human agents and their interactions, including decision-making procedures, vastly exceeds that of molecules. A goal for HPC and GSS together is to reflect this complexity in refined representations of human individuals as agents in these models. Ultimately, the need for computational power in addressing societal challenges may then increase far beyond what is needed in molecular modelling.

2 Status of the Health Habits pilot

The development and research work on the Health Habits pilot mainly focused on the improvement of the model definition and on the implementation of the full stack of the data management and analysis part.

2.1 Framework

The implemented model is a compartmental, agent-based model describing the smoking spreading in a population. While the details of the implementation are reported in the pilot's agent specification document in Appendix A, we here report the relevant novelties introduced in the model. The selected modelling framework fits the description of the smoking behaviour dynamics for different reasons: *i*) the process that an individual undergoes when she starts smoking can be seen as a complex contagion process (Lang, Abrams, and Sterck 2015; Beheshti and Sukthankar 2014; Sharomi and Gumel 2008; Levy, Cummings, and Hyland 2000), that can be modelled leveraging on the well-established framework of epidemic models (Barrat, Vespignani, and Barthelemy 2008; Castellano, Fortunato, and Loreto 2009), *ii*) it features a limited number of parameters that can be determined by means of data that are commonly available in most of the countries (namely smoking prevalence, statistics on the number of quitting attempts and relapse rate), and, *iii*) this particular framework can be easily implemented in an HPC-compliant framework (like for the global model of the Green Growth pilot (see Section 4.1.3 in D4.4), the Pandora framework (Rubio-Campillo 2014) was used) and can be extended to account for an agent graph model in the next software development steps.

In our model, each agent i stores the information on its current smoking status C_i and the time q_i that she spent in that status since the last transition. Each smoking status corresponds to a compartment in the model. We define three different compartments in our model (Castillo-Garsow, Jordán-Salivia, and Rodriguez-Herrera 1997; Levy, Cummings, and Hyland 2000):

- N: never smokers, i.e., agents that never smoked before and thus may start smoking;
- C: current smokers, i.e., agents that are currently smoking;
- E: ex smokers that quit the smoking habit. These agents may either relapse and start to smoke again in the future or permanently stay as ex-smokers until they leave the system (die), accordingly to the quit relapse mechanism explained below.

Note that by using this compartmental description we are doing some approximations with respect to the real-world scenario. For example, the current smokers compartment C comprehends all the possible types of smokers, from the regular to the occasional ones. This is done because data describing the distinction between the two types of smokers are lacking for many regions and time intervals. Another approximation is that, as we will see below, the quit relapse mechanism does not depend on the agent traits (such as age or time spent smoking) but is the same for every person in the system. Also in this case we preferred to test and calibrate a minimalistic model to later include more complex mechanisms.

Each agent can undergo one of the following transitions between the different compartments found in the system (see Fig. 2.1(a) for a pictorial representation of the model):

- $N + C \xrightarrow{\beta} 2C$: initiation process, a never smoker gets in contact with a current smoker and starts smoking at rate β. The latter generally depends on many factors, both global and agent-dependent (e.g., an agent's propensity to interact with others). We leave these details for next versions of the model. Just note that the parameter β accounts for all of the underlying mechanisms setting the infection rate. In other words, β sets how likely it is for the smoking habit to spread to a never-smoker when she gets in contact with a regular smoker. Note that we do not allow for spontaneous initiation process $N \xrightarrow{\beta} C$ at this stage. This mechanism would add a free parameter (a spontaneous initiation rate) difficult to measure in real world data and then to calibrate. Indeed, since this spontaneous initiation would add a second $N \rightarrow C$ channel, it would be very hard to disentangle the two mechanisms using real world data, where not even the number of people starting to smoke in one year is available.
- $C \xrightarrow{\gamma} E$: quitting process, occurring when a current smoker sets a quitting date, thus quitting smoking, at rate γ. This transition is spontaneous as we did not yet implement global awareness or social pressure to stop smoking in the model.
- $E \xrightarrow{\delta} C$: relapse mechanism, going in the opposite direction of the quitting process and occurring whenever a former smoker falls back to the smoking habit at rate δ. Actually, this rate is time-dependent as the relapse probability for an agent varies in time since he tries to quit smoking. For now, this relapse mechanism is calibrated on real world data regarding the outcomes of a one-year follow-up of two thousand patients in Great Britain (Ferguson et al. 2005). So, even if it is a spontaneous relapse, we can reasonably assume that all the mechanisms concurring in the relapsing process are accounted for in the implemented relapse rate. We will cover these details in Section 2.3.

To make the model more realistic we have to add demography (mortality). Indeed, so far an agent can only go from the N compartment to the $C \leftrightarrow E$ loop of quit and relapse. By adding demography to the population, one allows the agents to continuously enter (be born) and leave the system (die). At the current stage, this is done in the simplest possible way, i.e., we set natality and mortality rates to be equal and so that their value μ is comparable to the statistical records of the national health statistics service.

The natality/mortality process governed by the mortality rate μ sets the number of new agents added to the system in the N compartment (natality) and the number of agents removed from the system.

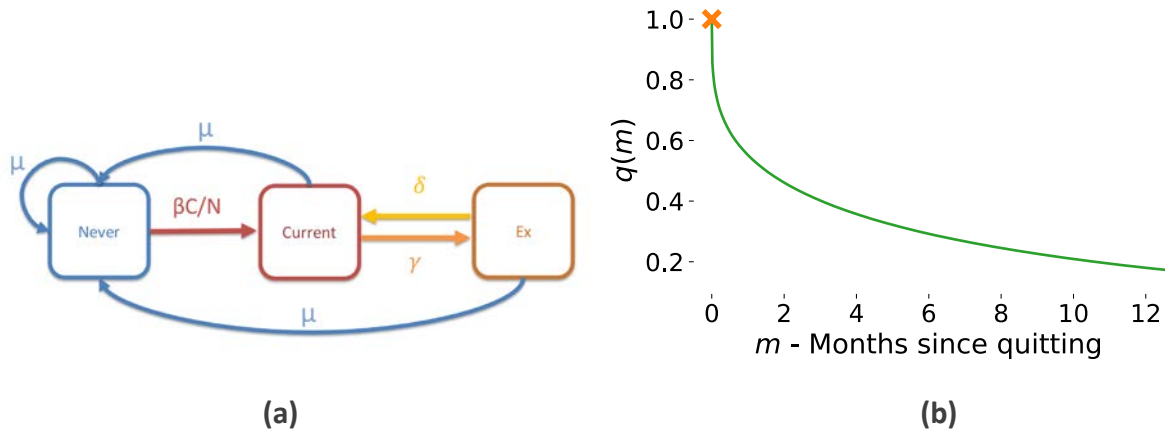


Figure 2.1: (a) The pictorial representation of the epidemiological compartment model with all the possible transitions. A newborn agent (one for every agent that died) enters the system in the N compartment, being a never smoker. Mortality introduces also the $N \to N$ transitions that just reset the age of an agent. (b) The fraction $q(m)$ of successful quitters after m months since their quitting date as reported for $m = 1$ by the 2013 National Health Survey (Public Health England 2016) (blue dot), $m \in [1, 12]$ by the study in (Ferguson et al. 2005) (orange crosses), and the fitted survival function $q(m) = r(m)$ when using a $\chi^2(m)$ time to first failure distribution with $k = 0.55$ degrees of freedom.

The mortality rate μ is the same for all the agents in the system, regardless of their current compartment. From this perspective, an agent enters the system in the N (never-smoker) compartment and then leaves the system at rate μ, regardless of the compartment she is at that time. The compartmental system can be written as a set of equations describing it in continuous time (see also Appendix A):

$$\begin{aligned} \frac{dN}{dt} &= -\beta S \frac{C}{P} + \mu(P - N), \\ \frac{dC}{dt} &= \beta S \frac{C}{P} - \gamma C + \delta E - \mu C, \\ \frac{dE}{dt} &= \gamma C - \delta E - \mu E, \end{aligned} \tag{2.1}$$

where $P = (N + C + E)$ is the total population of the system.

Note that the mortality rate μ sets the number of individuals entering and leaving the system in an infinitesimal time interval dt, as $\mu P(t)dt = \mu(N(t) + C(t) + E(t))dt$ is the number of people born (dead) in the $(t, t + dt)$ time interval. For a finite time interval of Δt length, we have that $\mu P \Delta t = B = D$ equals the number B (D) of people born (dead) as measured from census data.

We then set the mortality rate to be $\mu = M/P = \langle D, B \rangle / P$, where $M = \langle \dots \rangle$ is the arithmetic average of the measured number of deaths D and births B. For instance, in England and Wales together there has been $D = 529655$ and $B = 697852$ in 2015 over a population of $P \simeq 5.7 \cdot 10^7$, so that we can set $\mu = 0.0107\ \text{year}^{-1}$ (Office for National Statistics 2016).

Following a similar data pre-processing procedure, it is possible to estimate the value of the quitting rate γ and the parameters regulating the relapse mechanism, so far reported using

the relapse rate δ for simplicity. We briefly present this procedure in Section 2.4.2, while details can be found in the health habits model report in Appendix A.

2.2 Agent-based model definition

We are now ready to define the discrete-time ABM based on the just outlined continuous-time modelling framework. As we pass from a continuous compartment model to a discrete agent based model, we have to transform the continuous time rates μ, β, γ, to their discrete time probabilities counterparts. For simplicity, let us assume that we are given a function *rate2prob* that converts $m = \text{rate2prob}(\mu)$, $b = \text{rate2prob}(\beta C(t))$, $g = \text{rate2prob}(\gamma)$ (see Appendix A for details). Note that the rate b actually depends on the continuous rate β and the instantaneous prevalence of smokers $C(t)$, so that the initiation probability of an agent varies in time as $C(t)$ changes.

Also, note that we will implement the relapse mechanism indicated by δ using a truly agent-based approach, i.e. each agent will store personal information on her quitting status.

The parameters of the system are the natality (or mortality) probability m, the influence probability b, and the spontaneous quitting probability g. In addition, we have the parameters $\mathbf{p}$ describing the distribution of the relapse time. These parameters set the probability per single step for each agent to move from one compartment to the other. In addition, we have to define:

- P, the number of agents in the system (we assume without losing generality that P equals the real population P_{real} of the country under investigation);
- $\mathcal{C}$, the number of cells in the simulations, where each cell c corresponds to a particular area of the country under investigation and belongs to a specific geographical (or administrative) region r;
- $\mathcal{T}$, the time step of the ABM, i.e. each evolution step represents the evolution of the system for $\mathcal{T}$ units of time (could be days, months or years).
- the number n_T of steps to reproduce, so that $T = n_T\mathcal{T}$ is the total period of time simulated;
- $n(t=0) = N(0)/P$, $c(0) = C(0)/P$, and $e(0) = E(0)/P$, the initial conditions of the system, i.e. the fraction of the population that is found in every compartment at the beginning of the simulation;

Note that, depending on the data, we may let the rate γ (and thus also the probability g) vary from one administrative area to the other. For example, in Great Britain the National Health Service (NHS) provides measures of the quit/relapse rates on a regional scale (Public Health England 2016). We also define the population of the system to be divided into $\mathcal{C}$ cells, defined accordingly to a selected partition of the geographical space, in this case we selected the SEDAC cell division of the entire globe in $1x1$ km cells (SEDAC 2016). In this model definition, we assume that agents may interact only with other agents present in the same cell, leaving as a future development the inter-cell interactions of the agents.

Given these assumptions, and the parameter definitions given above, we can define the model as follows:

1. for each time step, cycle over the cells in the system;
2. for each cell compute the smoking prevalence $c_c(t)$ and, accordingly, the probability b of the smoking transmission from a current smoker to a never smoker;
3. cycle over the agents[2] of that cell c and update their status accordingly to the probabilities:

$$\begin{matrix} N & \overset{b}{\rightarrow} & C \\ C & \overset{g}{\rightarrow} & E \end{matrix} \tag{2.2}$$

4. if a current smoker tries to quit, draw the number of months $\overline{q}_i$ that she will spend in the E compartment from the time distribution $\mathcal{P}(q)$; if, instead, the agent is an ex-smoker, she will increase the q_i counter by $\mathcal{T}$ months, with $\mathcal{T}$ the number of months per discrete simulation step; if, after this incrementation, $q_i > \overline{q}_i$ the agent relapses going back to the C compartment;
5. replace with probability m an agent in the system (mortality) with a never-smoker in the N compartment (natality), so as to keep the population constant in time.

A visual description of the model is given in Figure 2.2.

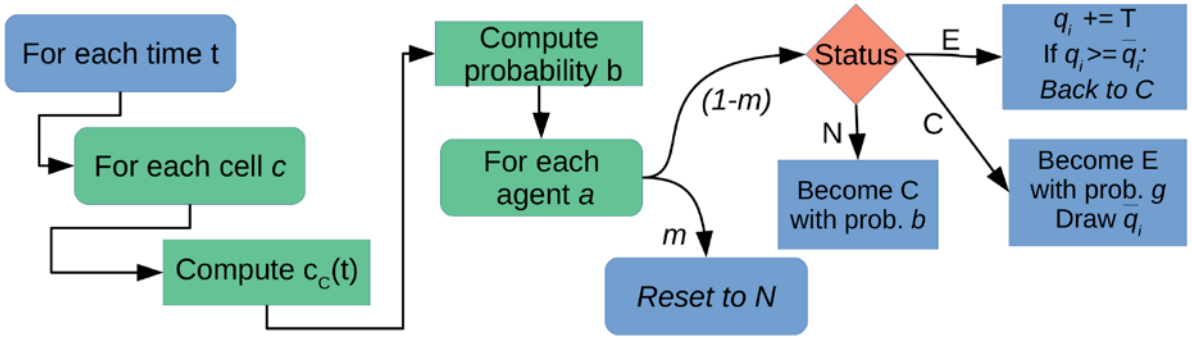


Figure 2.2: A visual representation of the compartmental model transposed to the ABM framework. From top left following the arrows: for each simulated time step t and for each cell c we compute the smoking prevalence $c_c(t)$, i.e., the fraction of current smokers in the cell population. We then compute the initiation probability b using this prevalence value. For each agent with probability m we reset it to the N compartment (mortality). With probability $(1-m)$ we check its status: if the agent is N it starts smoking and becomes C with probability b; if it is C it quits with probability g and draws a quit time $\overline{q}_i$ and sets $q_i = 0$. If it is E and after incrementing the counter $q_i + T >= \overline{q}_i$ then it relapses to C.

Let us stress that the model was implemented using this discrete-time, agent based formulation of it. The continuous time definition of it was presented to explain how to

[2] Note that "touching" all agents although not all update their state is necessary due to limitations of the ABM-framework used. The same holds for the following step.

measure the continuous rates of infection and quitting to later translate them to discrete transition probabilities.

2.3 Relapse mechanism

Once a current smoker is set to quit smoking she draws a relapse time q (measured in months) from a given time distribution $\mathcal{P}(q)$. The latter is derived from real world data reporting the quitting and relapse behaviour of people in England (Public Health England 2016; Ferguson et al. 2005) that we present in detail in Section 2.4.1.

These statistics report both the fraction of smokers trying to quit and the fraction r of these quitters that relapse after one month for each one of the nine regions of England. This fraction is measured both as individuals self-reporting the quitting status after one month, and the actual numbers of individuals who have been found negative to the carbon monoxide test (thus being CO-validated). Data are reported in Table 2.1.

Indicator	**Quitters**	**Successful**	**Successful CO-validated**
Region			
East Midlands region	7232	4056	2421
East of England region	7557	4112	2918
London region	7401	3844	2694
North East region	9532	4393	3519
North West region	8572	3714	1935
South East region	6030	3299	2432
South West region	6647	3444	2676
West Midlands region	8379	4403	3525
Yorkshire and the Humber region	5558	2994	2260

Table 2.1: The number of smokers per $100,000$ smokers setting a quit date (Quitters). The number of these individuals still being successful quitters after one month (Successful, self reported) and the subset of those who where also validated via carbon-monoxide test (CO-validated). Data from Public Health England (2016).

We expect that the reported quitters' success rate gives an over estimation of the real figures, as these numbers are bound to decrease as time passes from the quitting date. That is why we incorporated in the relapse analysis the data of a follow-up study tracking the quitters' relapse behaviour for one year (Ferguson et al. 2005). We report the findings of this study in Fig. 2.1(b).

We model this relapse mechanism by means of the hazard function $h(t)$, i.e., the conditional probability to fail (relapse) at time t given no previous failure, and survival (or reliability) function $s(t)$, the probability of no failure before time t. The hazard function $h(t)$ is then defined as:

$$h(t) = \frac{r(t) - r(t + \Delta t)}{\Delta t\, r(t)} = \frac{f(t)}{r(t)}, \tag{2.3}$$

where $f(t)$ is the time-to-first-failure distribution. Eq. (2.3) can be thought as a conditional probability for an agent to fail during a discrete time interval $[t, t + \Delta t]$ and it is measured as the number of failures observed in such time interval divided by the number $r(t)$ of individuals that never failed up to time t. The latter can be expressed in terms of the complementary cumulative distribution (CCDF) of $f(t)$, i.e., $r(t) = CCDF(f(t)) = 1 - CDF(f(t)) = 1 - F(t)$, where $F(t)$ is the cumulative of $f(t)$. We find the best fit to the empirical survival probability function $r(t) = q(m)$ (where $q(m)$ is the fraction of quitters still successful and without failures up to month m) to be a $\chi^2(t)$ time to first failure distribution with $k = 0.55$ degrees of freedom, as shown in Fig. 2.1(b).

Hence, the relapse mechanism is implemented by letting a quitting agent draw a relapse time $\overline{q}$ from the χ^2 distribution whose parameters are set by the previous analysis. Then, after $\overline{q}$ months are simulated, the agent will move back to the current smoker compartment C. We will also assume that if an agent draws a relapse period $\overline{q} > 12$ months we assume her to remain in the ex-smoker compartment forever (or until she dies), so that she will never relapse. This assumption is due to the limited time span of the quitters' follow-up and may be refined when longer follow-up studies will be available.

2.4 Data management

An essential part of the pilot work is to collect the data needed to compute the fixed parameters, generate the simulations' input and calibrate the model based on the simulations' output. In this section, we outline the methods and procedures developed by the pilot to address these requirements, starting from the data import and pre-processing up to the handling of the output.

In order to speed-up and generalize the pre- and post-processing procedures, one of the pilot outcomes has been the deployment of a geographical database and the development of a high-level interface to store and retrieve geographical data in a quick and simple way.

All these procedures allow us to combine different layers of information (census, health, etc.) into the model input as we show in Fig. 2.3(b). A more complete and detailed description of this work can be found in CoeGSS D3.3, Section 4.

2.4.1 Data pre-processing

The pre-processing part covers the data handling in the phases right before their importing to the database that we will present later.

The typical format of the data so far found as input is the *c*omma *s*eparated *v*alue (*csv*) format in which data are presented as a table whose columns contain the different values to be included in the work.

Census data As an example, the United Kingdom (UK) census data retrieved from the Office for National Statistics (ONS [3]) and the EuroStat website [4] present statistics on population, age structure of population and mortality/natality. For instance, the age structure of the population is given in a tabular form like:

CODE	NAME	ALL AGES	0	1	2	3	…
K02000001	UNITED KINGDOM	65,110,034	776,769	786,125	806,704	834,076	…
K03000001	GREAT BRITAIN	63,258,413	752,520	761,640	781,686	808,221	…
K04000001	ENGLAND AND WALES	57,885,413	696,519	704,962	723,873	748,880	…
E92000001	ENGLAND	54,786,327	662,977	670,993	688,932	712,587	…
E06000047	County Durham	519,695	5,340	5,569	5,708	5,768	…
E06000005	Darlington	105,389	1,231	1,211	1,319	1,354	…
E06000001	Hartlepool	92,493	1,015	1,068	1,094	1,155	…
E06000002	Middlesbrough	139,509	1,947	1,967	2,008	1,998	…
E06000057	Northumberland	315,263	2,745	2,911	3,096	3,259	…
E06000003	Redcar and Cleveland	135,275	1,472	1,482	1,623	1,674	…
E06000004	Stockton-on-Tees	194,803	2,363	2,362	2,510	2,573	…

Table 2.2: Excerpt of the UK census data set.

where the columns report the Nation or **L**ocal **A**uthority **D**istrict (LAD) code, the name of the region, the total population and then the people of each age in the corresponding column.

Health data The data about the smoking habit report the number of people that are current, ex or never smokers. These figures describe the situation at different geographical resolutions, covering different time periods. We find both a coarse national prevalence from 1974 onwards [5], and a detailed LAD-resolved prevalence starting in 2012 [6].

Besides the smoking prevalence, we leverage on other socio-economic indices reported by the ONS, e.g., the affordability of tobacco during the 1980-2014 period, where affordability is a measure of the price of tobacco in relation to the consumer's income level.

Another source of data exploited comes from the *Local Tobacco Control Profiles for England* [7] as shown in Fig. 2.1 and Table 2.1, that use parts of this data.

[3] https://www.ons.gov.uk/
[4] http://ec.europa.eu/eurostat
[5] See the *Adult Smoking Habits in Great Britain* from the ONS dataset.
[6] Time series from 2012 to 2015 available in the *Smoking habits in the UK and its constituent countries* dataset from the ONS.
[7] See http://www.tobaccoprofiles.info.

The data report the number of smokers per 100'000 smokers who actually set a quit date during the year (quitters) and then the number of such quitters that is still successful after 1 month. To extend the statistics on the relapse mechanism we include also a study reporting a one-year follow-up of quitters that reports their success rate at three months intervals (Ferguson et al. 2005).

These data allow for a more realistic modelling of the quit and relapse process as described in Section 2.3.

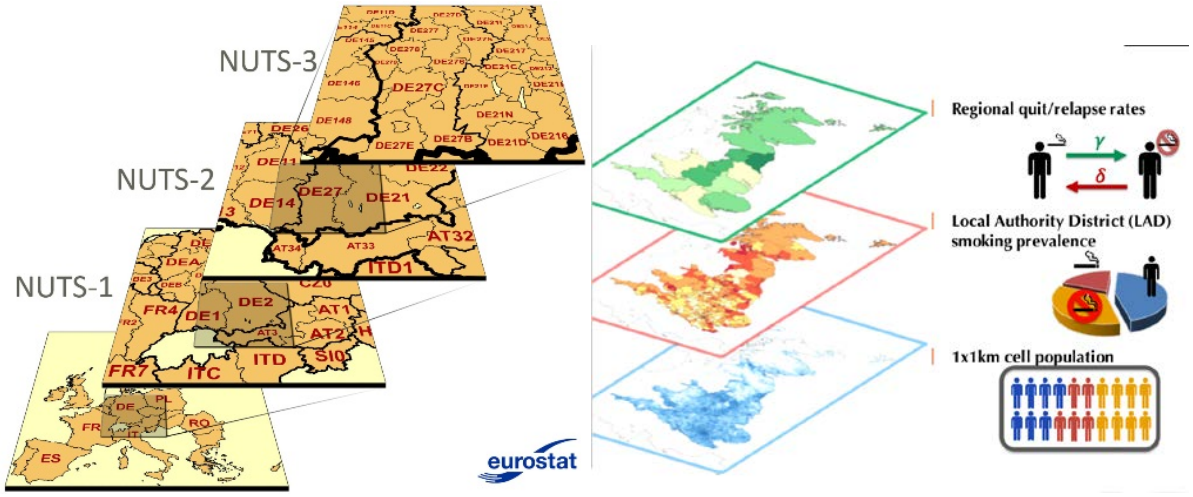
Geographical and boundaries data The geographical data about boundaries and mapped observables are mainly found in two formats, vector and raster.

The former class comprehends the shape-file and the geo-JSON formats and stores data as lists of polygons (the boundaries of different nations or regions) with a list of attributes attached (e.g., figures regarding population and health conditions, the name of the region, its administrative code etc.). While this representation is not memory demanding and it allows for a hierarchical representation of the administrative boundaries, it does not provide a detailed spatial description of space-dependent indicators (such as the smoking prevalence in a given area) below the resolution of the boundary itself.

That is why we have to integrate the vector data representation with a finer one, i.e., with the second type of geographical data: rasters. These can be thought as grids of points covering a particular region and reporting the numerical values of a particular indicator, e.g., population count in a given square.

In this work, we use the SEDAC UN-adjusted raster, reporting the population count for each one square kilometre cell globally, [8] mainly because it reports data in a precise grid of fixed latitude-longitude steps, which is valuable when working with parallel simulations where different nodes process agents belonging to different cells.

[8]See http://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-adjusted-to-2015-unwpp-country-totals.



(a) (b)

Figure 2.3: (a) The NUTS hierarchical levels explained. Image from http://ec.europa.eu/eurostat/web/nuts/overview. (b) The different layers combined to prepare the simulation input. From top to bottom the regional statistic of quit/relapse (green), the LAD-level data on smoking prevalence (red) and the SEDAC cells containing the population count (blue).

Region nomenclature

NUTS - The European Commission's *Nomenclature d'Unités Territoriales Statistiques* (NUTS) geographical division is a regional code standard of the European Commission that naturally provides a hierarchical organization of regions and areas. Indeed, the typical NUTS code reads CC123, where CC are two characters identifying the country, while 1, 2, 3 are three alphanumeric characters that are present in the first, second and third NUTS level, respectively (i.e., CC12 is a code of level 2 for country CC which is a child of the CC1 level 1 NUTS). A schematic visualisation of the NUTS code organization is shown in Fig. 2.3(a).

LAD - Though the NUTS code provides 4 levels of geographical detail, in our pilot implementation we have Great Britain's data on smoking prevalence given at the Local Authority District (LAD [9]) level, which, surprisingly, is at a higher resolution with respect to the NUTS-3 level. To this end we created in our geographic database a fourth, custom defined level of NUTS by appending a character to the NUTS-3 code corresponding to each LAD's parent NUTS-3 boundary.

The main task during pre- and post-processing is to match the data contained in the raster cells to the vector boundaries, and vice-versa. Indeed, while in the data pre-processing one has to project the fraction of agents belonging to each compartment for every boundary to the simulation cells belonging to it, in the post-processing part the task is to aggregate for each boundary the number of agents for each compartment for all the simulated cells lying within the given boundary.

[9] See http://geoportal.statistics.gov.uk/datasets?q=LAD%20Boundaries%202015&sort=name

Rasterization The final step of the pre-processing phase is the creation of the rasters (one for each simulated compartment and one for each-space dependent parameter) containing either the numbers $C(i,j)$ of agents belonging to compartment C in the (i,j)-th cell or the value of the generic parameter $r(i,j)$ in the same cell.

For example, the region-dependent quit rate $\gamma(R)$ setting the propensity of current smokers to set a quit date varies from one region R to the other. Thus, all the raster cells belonging to the region R will contain the corresponding value $\gamma(R)$.

On the other hand, the agents' raster cells contain the $C(i,j) = P(i,j) \cdot p(C)$ value which is the result of a combination of the SEDAC raster population count $P(i,j)$ at cell (i,j) and the prevalence value $p(C)$ of compartment C of the boundary containing the cell.

To summarize, we obtain the input rasters by performing the following procedure that is also outlined in Figure 2.4:

- get all the shapes of the boundaries composing the area under investigation (UK) and derive the longitude-latitude box that contains it;
- retrieve the SEDAC transformation geometry matrix and set the transformation matrix of the new raster accordingly;
- retrieve the cells contained in the area bounding box and compute the desired value in combination with the boundary value;
- export the resulting raster.

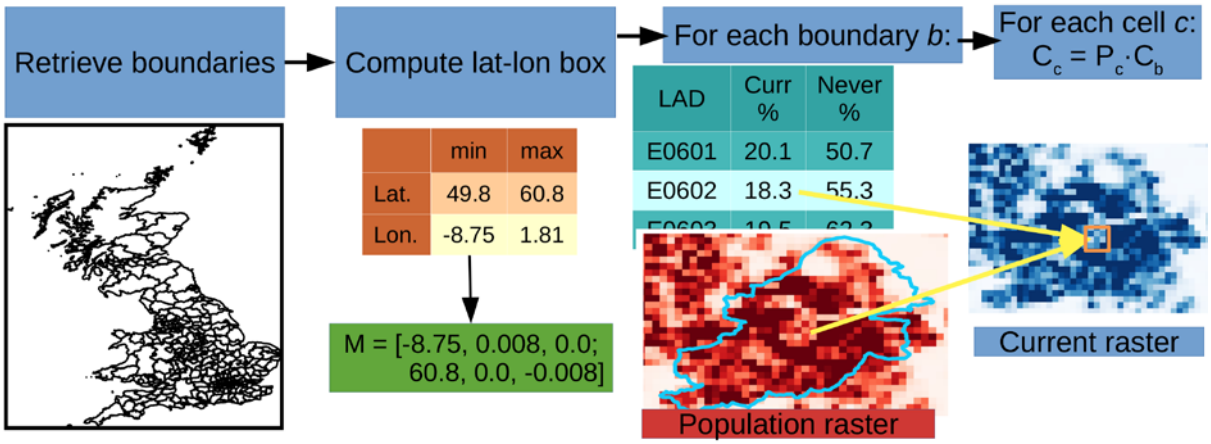


Figure 2.4: A visual example of the rasterization process. From left to right: we retrieve the boundaries at a given level from our geographic database, computing the latitude-longitude extension of the box containing all the boundaries. This allows us to compute the transformation matrix M that is used to align the SEDAC raster and the vectorial boundaries. Then, for each boundary b we query for all the cells $c \in b$ overlapping with b. For each cell $c \in c$ we compute the output cell value (in this case the number of current smokers C_c in the cell c) as $C_c = P_c \cdot C_b$, where P_c is the number of people living in the cell c and C_b is the prevalence of smokers in the boundary b.

As a quick example, the generic value $C(i,j)$ of the current smokers raster's cell (i,j) can be evaluated as:

$$C(i,j) = P(i,j) \cdot f_{\text{curr}}\big(B \supseteq (i,j)\big), \tag{2.4}$$

where B is the boundary containing the (i,j)-th raster cell and $f_{\text{curr}}(\cdot)$ is the fraction of current smokers in the administrative unit represented by the boundary B.

2.4.2 Parameter estimation

The parameters that can be estimated per LAD are the quit rate γ together with the initial conditions on the N, C and E numbers of agents falling in each compartment. At a national level, we can instead estimate the parameters of the time to first relapse distribution $\chi^2(m)$ and the natality (mortality) rate μ. The transmissivity β is left free as the parameter to optimize during model calibration.

The evaluation of γ_{LAD} for each LAD is done by using the fraction $q_{\text{LAD}}(t)$ of current smokers that tries to quit (the column *Quitters* in Table 2.1). Then,

$$\gamma_{[0,\Delta t]} = -\frac{1}{\Delta t}\ln\left(\frac{C(\Delta t)}{C(0)}\right) = -\frac{1}{\Delta t}\ln\big(1 - f_{\text{quit}}\big), \tag{2.5}$$

where f_{quit} is the fraction of smokers trying to quit in the $[0, \Delta t]$ period (usually $\Delta t = 1\text{yr}$). By applying Eq. (2.5) for every LAD value we populate the *gamma* column with the corresponding γ_{LAD}.

Finally, the $\mu = 0.0127\ \text{yr}^{-1}$ and the time to first relapse distribution $\chi^2(m)$ parameter $k = 0.55$ are derived nation-wide as discussed in the previous section.

2.4.3 Geographic database

As already mentioned, pre- and post-processing tasks can be strikingly simplified by using a space-aware database interface such as the one provided by using mongoDB and the Python module shapely together. The technical details can be found in CoeGSS D3.3 Section 4.

2.4.4 Simulation output

The data post-processing consists in the management of the simulations' output data, i.e., the rasters containing the counter of agents falling in each model compartment for each simulated time step. For example, we start from the number of smokers in a given SEDAC cell.

The smoking prevalence in a given LAD or region is computed by summing the number of current smokers for all the cells falling within the region's boundary, thus reversing the boundary to cells procedure outlined in the rasterization process.

Specifically, we want to aggregate for each boundary B the number $C(B)$ of agents falling in the compartment C for all the cells $(i,j) \in B$, possibly accounting for the overlap $\phi_{(i,j),B}$ of the cell with the boundary (i.e., the cell may not entirely lie within the boundary), so that:

$$C(B) = \sum_{(i,j)\in B} C\,(i,j) \cdot \phi_{(i,j),B}, \tag{2.6}$$

with $0 < \phi_{(i,j),B} \leq 1$.

Once this back-aggregation is done, we load the result into the database by using the developed interface:

```
geoClient.updateBoundary(boundaryID,
 {"$set":
  {'properties.simulations.health.smoking.simulation2012-2020.2012.CurrSmok':
        numberOfCurrentSmokers}})
```

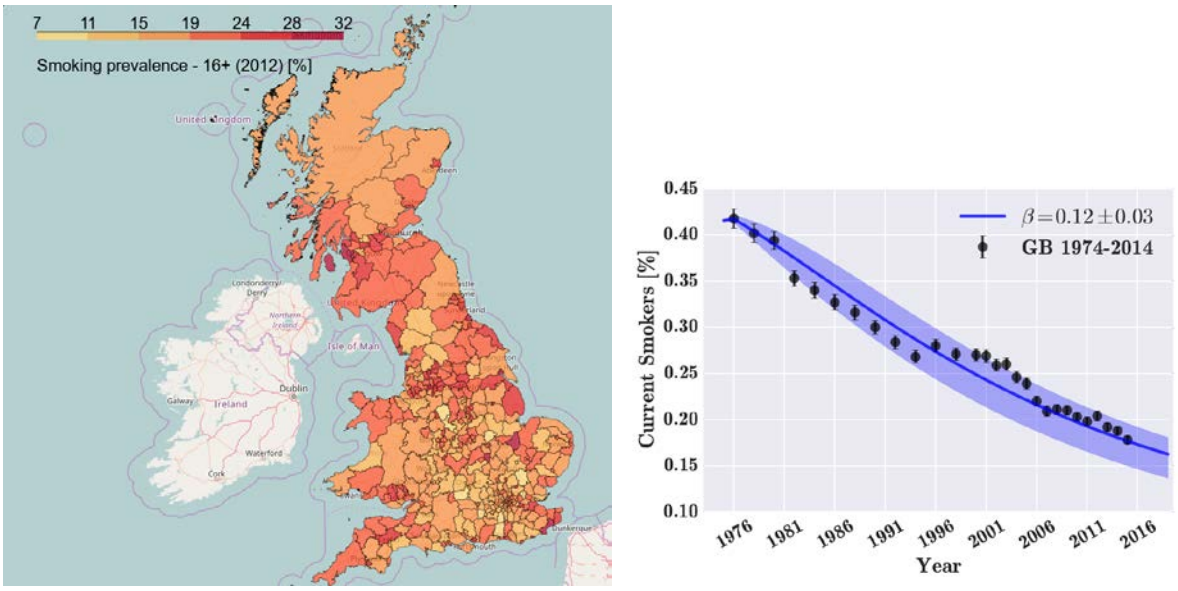
where numberOfCurrentSmokers is the vector containing the simulated number of smokers in the given boundary, i.e. we add one value for each run of the simulation. Also, the data will be saved under the simulation2012-2020 key, that is used as a reference to recall different simulation runs done in order to reproduce different results.

As for the prevalence data we can now aggregate the results of the simulation (that are loaded in the database at the NUTS 4 level) so that we can reproduce the regional and national prevalence just issuing

```
AggregateCountryLevel(countryCode="UK
        levelStart=3,
        levelStop=0,
        mode="sum",
        valfoo=np.mean,
        valueField=
   "properties.simulations.health.smoking.simulation2012-2020.*.*")
```

where we sum the results for all children of a NUTS 2 level (as the simulation output is the absolute number of smokers and not a prevalence), and where the final asterisks denoted the keys over which aggregation should be repeated. In this case we want to aggregate this data for each simulated year and for each different smoking indicator.

Once this procedure is complete, we have in our database both the empirical and the simulated smoking prevalence time series, see Fig. 2.5(b), and we can proceed with model calibration.



(a) (b)

Figure 2.5: (a) The smoking prevalence at LAD level resolution in Great Britain in 2012. (b) The empirical (black dots) and simulated (blue solid line) smoking prevalence in UK from 1974 to 2014. The optimal value of $\beta = 0.12 \pm 0.03$ is shown together with the confidence interval (blue shaded area).

2.5 Model calibration

We run a preliminary parameter sweep for model calibration, meaning that all the model's parameters are fixed except the influence rate β. In order to calibrate the model, we retrieve the empirical and the simulated national prevalence for the 1974-2014 time-interval from the database, for each simulated β value as shown in Fig. 2.5(b).

Then, for each value of the influence rate parameter β we compute the discrepancy between these two time series as the $\chi^2(\beta)$ sum of squared residuals for each health status compartment, i.e.

$$\chi^2(\beta) = \frac{1}{(Y-1)} \sum_{\text{status}} \sum_{y=1974}^{2014} \left[f_{\text{empirical}}(\text{status}, y) - f_{\text{simulation}}(\text{status}, y, \beta) \right]^2, \qquad (2.7)$$

where Y is the number of years simulated, $f_{\text{empirical}}(\text{status}, \text{year})$ (and $f_{\text{simulation}}(\text{status}, \text{year}, \beta)$) are the empirical (simulated) prevalence of a given health habit *status* at a given *year* (and for a given value of the β parameter for the simulations), respectively.

The optimal $\overline{\beta}$ is then defined as:

$$\overline{\beta} = \min_{\beta} \chi^2(\beta). \qquad (2.8)$$

2.5.1 External indicators

As one can see from Fig. 2.5(b) the predictions of the model are in agreement with the empirical data but feature a smooth, regular descent, whereas the empirical data show a more noisy behaviour.

The empirical signal of smoking prevalence is the result of the interplay of many different complex socio-economical processes, of varying smoking restriction policies and collective awareness. To integrate all of the possible mechanisms in a single model would be infeasible and would result in a cumbersome model whose calibration would be difficult.

Instead of complexifying the model beyond our analytical power, we can parsimoniously insert in the simulations a few data layers accounting for the most important processes leading the smoking prevalence evolution. To this end, the tobacco affordability index $I(t)$, with $t \in [1980,2014]$, accounts for the economic situation and the active policies in time to evaluate how easy and likely it is to afford tobacco, on average, for the population.

We add the $I(t)$ index to the model in a simple but reasonable way, i.e., it modulates in a non-linear way the absolute value of the β influence rate as

$$\tilde{\beta}(t) = \beta S'\big(I(t)\big) = \beta \cdot \frac{1}{2}\left[1 + \exp\left(I(t) - \frac{\bar{I}}{\epsilon}\right)\right], \tag{2.9}$$

where we set $\bar{I} = 1$ and $\epsilon = .2$ with $I(t = t_0) = 1$.

The results of the simulations run with this varying influence rate are shown in Fig. 2.6(a). We now observe that the simulated prevalence has a non-monotonic behaviour with respect to the previous case. In particular, simulations are able to capture the peak of smoking prevalence during the early 90s.

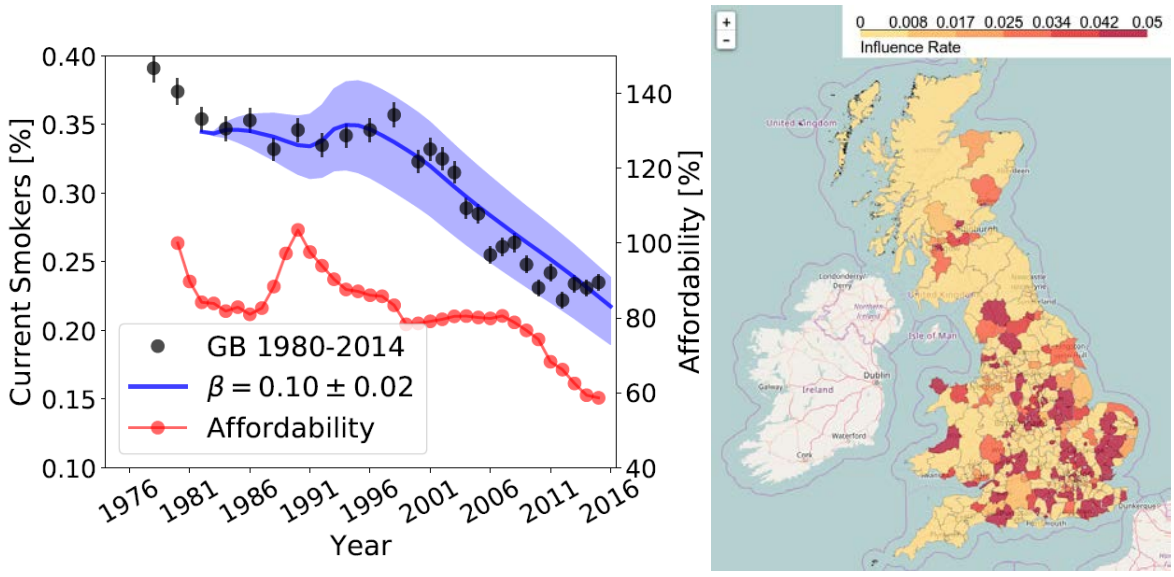


Figure 2.6: (a) The empirical (black dots with confidence interval) and simulated (blue solid line and shadows for the confidence interval) smoking prevalence in the 1982-2015 period. The simulated prevalence is obtained using the time-varying influence rate $\widetilde{\boldsymbol{\beta}}(t)$ depending on the tobacco affordability index $I(t)$ (red dots) as shown in Eq. (2.9). (b) The map of the $\overline{\beta}_{\mathrm{LAD}}$ for Great Britain's LAD (colourbar is shown). The optimal $\overline{\beta}_{\mathrm{LAD}}$ is derived following Eq. (2.10).

2.5.2 Regional model calibration

The procedure shown in Equations (2.7) and (2.8) can be applied at an arbitrary geographical (NUTS) level (as long as we have empirical data refined at the chosen NUTS level). For example, we can calibrate the model with a LAD-dependent $\beta(l)$ in the 2012-2015 period while we can only find a national β in the previous period.

When we simulate the 2012 onward period in Great Britain we initialize the simulations with the prevalence resolved at the LAD level, evolve for different values of β and finally compute the simulated smoking prevalence for each LAD, separately.

Then, we can back aggregate the smoking prevalence for each LAD and find the optimal beta for each LAD, by generalising Eq. (2.7) to:

$$\chi^2_{\mathrm{LAD}}(\beta) = \frac{1}{(N-1)} \sum_{\mathrm{status}} \sum_{y=2012}^{2015} \left[f_{\mathrm{emp}}(\mathrm{status}, y, \mathrm{LAD}) - f_{\mathrm{sim}}(\mathrm{status}, y, \mathrm{LAD}, \beta) \right]^2, \quad (2.10)$$

$$\text{s.t.} \quad \overline{\beta}_{\mathrm{LAD}} = \min_{\beta} \chi^2_{\mathrm{LAD}}(\beta).$$

An example of such analysis is reported in Fig. 2.6(b), where we show the national map of the $\overline{\beta}_{\mathrm{LAD}}$ in Great Britain: this analysis reveals the areas where the smoking epidemics is more severe and should then be addressed with the highest priority. Visualisation here is also obtained with the help of the geographic database interface as described in CoeGSS D3.3.

2.6 Future directions

The obtained results show that the smoking prevalence modelling task can be tackled using epidemic-like models. Starting from real world data we initialized a compartmental agent-based model and calibrated it with empirical time series.

Though the model is simple and implements homogeneous mixing between the agents, the results are in good agreement with the empirical data. The parsimonious modelling approach also allows including additional mechanisms and external indicators shaping the evolution of the system. Indeed, we showed that, by accounting for the tobacco affordability index, we get a more realistic description of the system evolution.

Moreover, the developed analysis also allows for a fine fit of the model that reveals the problematic areas of a country where the propensity toward the adoption of the smoking habit is higher.

Nonetheless, there are still different mechanisms and refinements to introduce in the model, defining the future work and efforts of the pilot. First, we have to overcome the homogeneous mixing approximation by inserting contact matrices and a population structure into the model, so as to fully leverage the agent-based approach. This can be done by using a more detailed synthetic population accounting for social interaction networks among the population, e.g., by grouping people in households and communities so as to better model their social contacts.

Besides, our results call for a better understanding of the relevant processes shaping the interactions and the complex-contagion process to include into the simulations.

Another important direction is the improvement of our analytical framework so as to be able to pinpoint the most important changes in tobacco policies for a given country and to explain how a single policy may have very different outcomes when applied in two different countries. To tackle these problems, we also have to extend the simulated geographical region, by including additional datasets and countries in our simulations.

3 Status of the Green Growth pilot

3.1 Approach

The Green Growth Pilot studies electric mobility in view of green growth (see D4.4 and Appendix B). In order to do so, the car centred global system has been defined (see Section 3 of this appendix) and two models have been developed.

1. A global model with spatially explicit data input on a fine grid and a basic innovation diffusion dynamics. Section 3.2 highlights the progress made in this line of work since the status report in D4.4.
2. A complex agent-based model for a specific region in Germany (Niedersachsen, the federal state home to the Volkswagen company, together with the two cities Bremen and Hamburg, also a federal state each). Structures for this model have been specified in D4.4, here, Section 3.3 reports about model definition, implementation, and simulations.

The aim of this approach is to use the smaller scale in the complex model for prototyping and testing mechanisms in the agent-based model setup. A smaller region has been chosen for this work so as to keep data collection and pre-processing activities manageable. Once consolidated, elements from the complex model can be integrated into the global model, for those regions where data are available.

3.2 Global model

3.2.1 Conceptual model

The conceptual model has been described in D4.4. Introducing a country specific policy parameter to take into account possible measures like a state-subsidised buyer's premium has extended it. This country specific parameter γ_c is multiplied with the GDP per capita, so that e.g. the buyer's premium would be represented in the model as an increased GDP. A broader discussion of this policy parameter can be found in Section 4.2 of Appendix B.

3.2.2 Data

Data used to construct the global model has been described in D4.4. In order to calibrate the model against the new registrations of battery electric cars by country, data from the International Energy Agency (IEA 2016, Table 9) was used. This covers yearly data from 2005 till 2015 for 16 countries. About 95% of battery electric cars worldwide were registered in these countries.

3.2.3 Simulations and calibration

In the case with data for 16 countries, there are 18 parameters that can be used to calibrate the model, $\gamma_1 \dots \gamma_{16}$, η and κ. To simplify the process, the deterministic version of the model was used and two properties of the model where exploited:

- As γ_c and η are both factors for the GDP per capita, we can combine them into a single country specific parameter η_c to get rid of η.
- Due to the rather limited direct influence range of cells of maximally 10 km, there is only a marginal influence of the value of η_c on the value of η_{c^*} for c and c^* neighbouring countries (and even less for non-neighbours). Therefore, it is not necessary to run simulations for different combinations of the η_c.

So, at the end it was sufficient to exploit a two-dimensional parameter space, where the parameters are κ and η. An exploration task was created in OpenMOLE (https://www.openmole.org/), that scanned κ between 0.4 and 0.7 and η between 5e-06 and 3e-04. After this, κ was fixed to 0.7 and additional runs with η up to 0.02 were performed, as the optimal value of γ_{Norway} was outside the given range. Overall, about 250 simulations where carried out.

While this work can be seen as a "manual calibration", supported in part by the OpenMOLE exploration task, there is scope for multi-dimensional optimization over model parameters in similar cases (see Section 6).

3.2.4 Insights from preliminary results and outlook

Figure 3.1 and Figure 3.2 show the differences between the data from the International Energy Agency report and the simulation results after calibrating the data for some selected countries. In addition to the data the simulation results of (up to) two different simulations are shown. One displays the simulation with the minimal least square error compared to the data, the second one restricts the parameter space to one dimension by fixing κ to 0.7. This restriction is necessary as only η is modified by the country specific parameter.

As can be seen, the aggregate number (which is labelled as "World" in Figure 3.1) of battery electric car sales matches the data. But this is not the case for countries like Norway and China, which have an above-average growth in 2014 and 2015. In the current version, the parameters are the same for the whole simulation period. A possible solution would be to add a yearly γ_c, which could represent the introduction of policy measures at different points in time, but without further restrictions this would overfit the model.

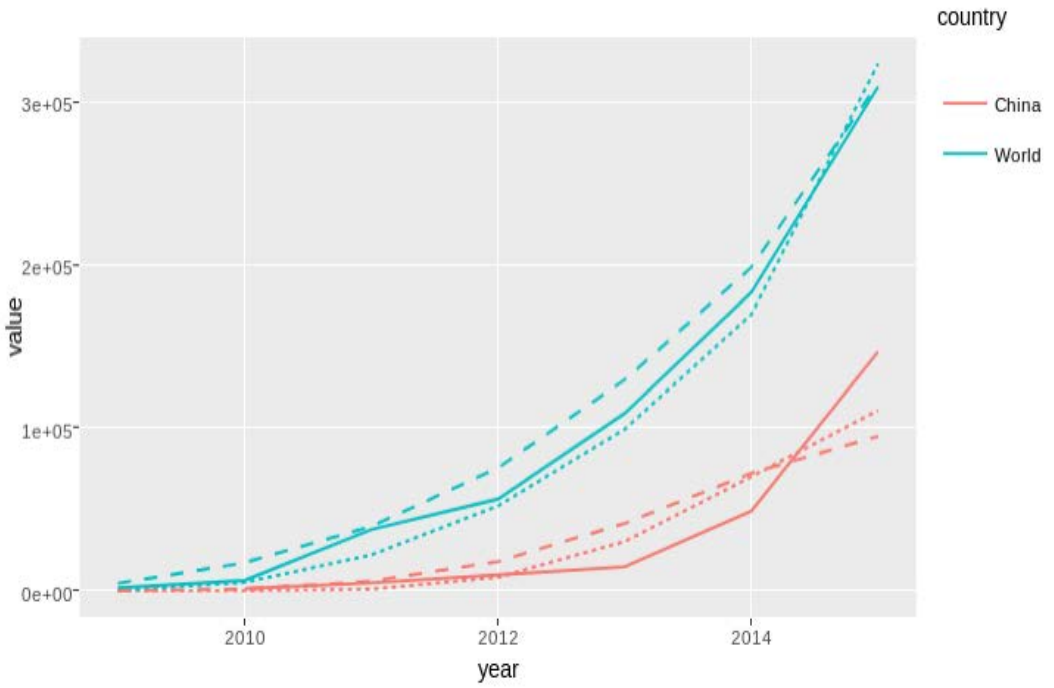


Figure 3.1: Compare calibrated simulation results with data of sold electric cars for China and the World. The solid lines show the data, the dotted lines the simulations with the least square error (LSE) and the dashed lines the simulations with the LSE restricted to the subset of simulations with κ = 0.7.

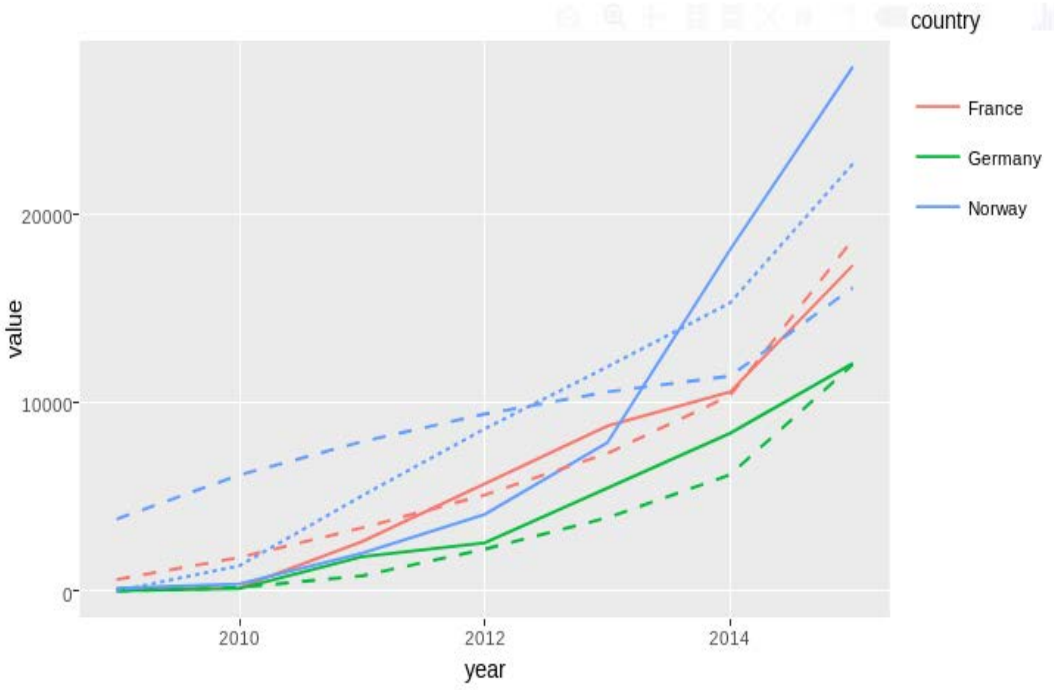


Figure 3.2: Compare calibrated simulation results with data of sold electric cars for France, Germany and Norway. The solid lines show the data, the dotted lines the simulations with the least square error (LSE) and the dashed lines the simulations with the LSE restricted to the subset of simulations with κ = 0.7. In the case that the simulation with the LSE has κ = 0.7, no dotted line is visible

For a discussion of the results from the calibrated simulation runs and an outlook see Appendix B.

3.3 Complex model

3.3.1 Conceptual model structure

The conceptual model has been described in D4.4; in short, the main elements are that we focus on households as potential car buyers and model their decision making with the help of expected utility maximisation. They use a learning mechanism to update expectations given information from their social network.

The agent-based mobility transition model we implemented is called MoTMo (Mobility Transition Model). It currently distinguishes between three types of mobility ("brown" (high-emission) cars, "green" (low-emission) cars, and "other") and is able to simulate a temporal development of the prevalence of these three types based on expectations, experiences, and social learning of individual actors. At the moment, it is run using individuals from a synthetic population of the German federal states Niedersachsen, Bremen, and Hamburg (NBH) (NUTS-1 regions), which are displayed in Figure 3.3. The synthetic population was originally generated by the MIDAS project and adapted as described in Section 3.3.2.

MoTMo is implemented in an object-oriented manner as a graph-based ABM of many interacting heterogeneous entities (agents) of a few different types (persons, households, locations). Each entity type is implemented in separate class with properties and methods. The population in the three regions consists of about 10 Mio. people, which are currently reduced in the simulation by a factor of 20. Thus, the we compute the model using about 500,000 persons, allocated to about 250,000 households, at specific locations, and two unique aggregate global entities (world and market).

The world class represents the environment, and controls the simulation step, and the market is an economic and technological representation of the mobility sector. Households are allocated on a regular grid network of locations; presently, the NBH area is divided in 3,833 locations. The locations represent data points of the SEDAC gridded world population data set as described in D4.4. There is another network of social interactions (friendship) between persons, which is created in the initialization phase (for details, see the initialization phase description below). It connects persons with similar priorities that live in the same or a neighbouring location.

The system evolves in discrete time-steps; therefore all entities proceed in time by an internal step function. Thus, the simulation workflow can be described the individual step functions of all entity types and in which order these functions are called. The order in which the step functions are executed is defined in the step function of the world class.

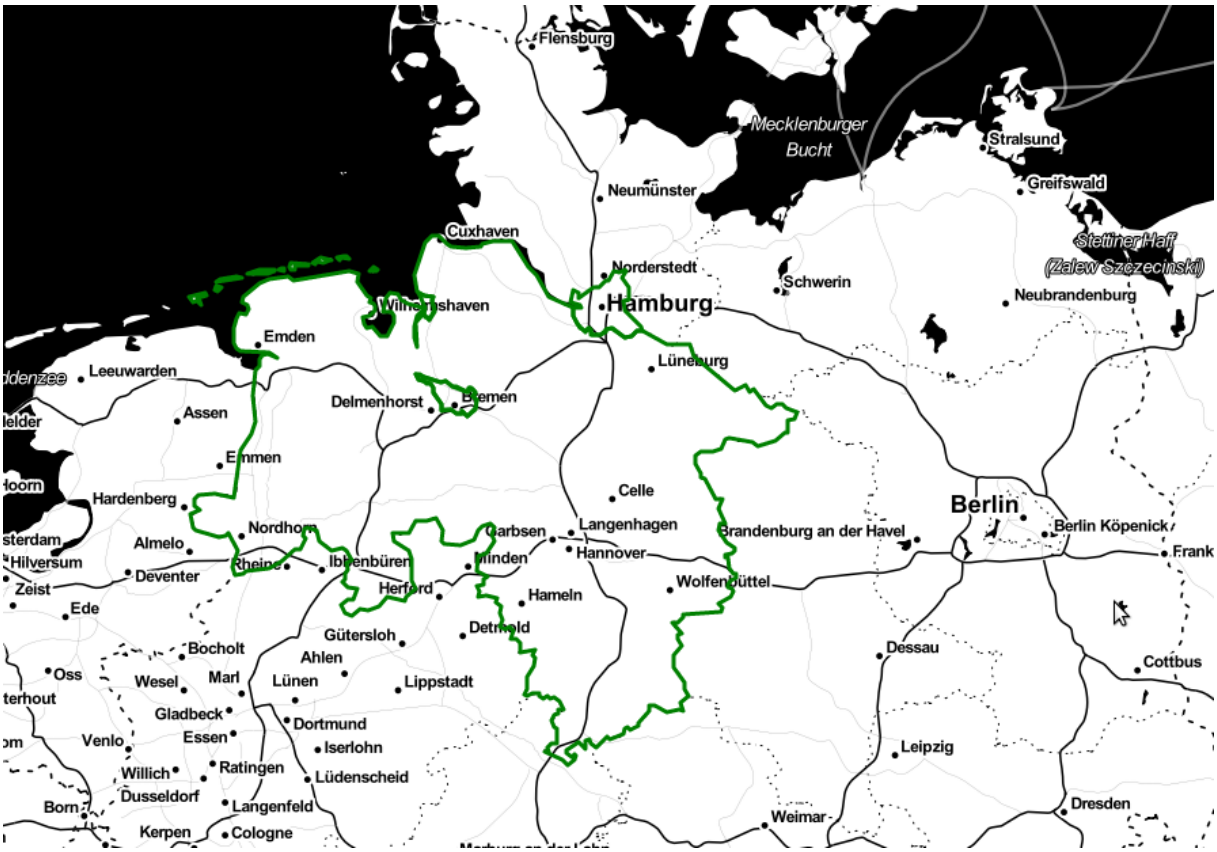


Figure 3.3: Model region including Niedersachsen, Bremen, and Hamburg.

3.3.1.1 Input parameters

In the current setup, a scenario is defined by the input parameters that are listed in Table 3.1

Parameter	Value	Description
nSteps	460	number of simulation steps
burnIn	100	number of burn-in time steps
burnInTimeFactor	2.5	factor that increases time speed during burn-in phase
omniscientBurnIn	10	number of time steps (at the beginning of burn-in phase) with omniscient agents
connRadius	3.5	radius within which locations are connected
minFriends	60	min number of friends per person
maxFriends	200	max number of friends per person
utilObsError	50	observation error to compare utilities
mobNewPeriod	60	time period in which a mobility choice is considered new and will thus not be changed
randomCarProp DeviationSTD	0.01	individual random deviation of car properties

urbanThreshold	20,000	population density threshold that separates urban and rural
urbanCritical	40,000	population density for minimal convenience
convA	0.8	maximum convenience of brown and green cars
convC	0.3	maximum convenience of mobility type other
convD	0.06	rate how fast convenience changes with population density
kappa	-0.4	initial green infrastructure disadvantage
innoPriority	0.15	weight of priority of innovation
mobIncomeShare	0.1	share of the household income that is spent for mobility
individualprio	0.33	individual random component of priorities
charIncome	5,000	characteristic normalizing constant for income
minIncomeEco	2,500	minimal income boundary for ecologic priority
initialGreen	2,000,000	inertia of green technical change (for full population)
initialBrown	80,000,000	inertia of brown technical change (for full population)
initialOther	40,000,000	inertia of other technical change (for full population)
radicality	1	variable to control the opinion strength
util	cobb	persons' utility function (Cobb-Douglas or CES-function)
reductionFactor	100	population reduction factor
selfTrust	3	how much more an agent values his own opinion

Table 3.1: Listing of all relevant scenario input parameters.

In the following we describe the classes world, market, location, person and household.

3.3.1.2 The world class

The world class is a global structure within which all other entities are embedded. It contains the graph that structures all interactions between agents. The graph represents multiple entity types, location, households and persons as well as different types of connections. It represents spatial relations between neighbouring locations and their distance. It connects the households to their location, providing the spatial component. It further connects the persons the household they live in. At last a social connection between individual persons is represented as last connection type. Other non-graph-related properties of the world class are listed in Table 3.2 contains the properties of world class. In addition, all input parameters from Section 3.3.1.1 are available in the world class, which contain the most important model related parameters.

Variable	Description
timeStep	current time step of the simulation
globalRecord	structure for all global values that are stored
glob	structure for global variables (MPI communication)
graph	structure that represent all agents' interactions

Table 3.2: State variables of the world class

For progressing a single time step, the function world.step() is called. Within this step, first the time is advanced by one unit and it is checked whether the simulation is in the burn-in phase or not. During the burn-in phase, the time can be scaled by a factor (burnInTimeFactor) for faster convergence towards an initial stable state, which means time can be increased by this factor. Next, global variables are updated, synchronized (in the case of multiple processes) and stored for the record. After that, the step function of the market is called to get the new state of the mobility market (see 3.3.1.3). The step function is called for all locations to update the spatial related infrastructure and environment variables (see 3.3.1.4). Then, all households are iterated in random order and their step function is called. This loop evaluates and executes actions and decisions as described in 3.3.1.6. Dependent on the settings, two steps are distinguished, an omniscient step and a normal step. The omniscient step differs from the normal step, in that the household and persons take decisions while having all required information available, while in the normal step, only expectations over certain relations and consequences are available.

The final loop iterates over all persons to covers all social interactions between persons (see Section 3.3.1.5.). At the end of each step, global data is synced between processes.

3.3.1.3 The market class

The market class represents the production and development component of the model. It updates the properties of mobility types over time depending on their market shares.

Table 3.3 contains the variables of the market entity. In addition, all input parameters from Section 3.3.1.1 are available, which contain the most important model related parameters.

Variable	Description
sales	list of current sales per mobility type
meanEmis	mean emissions
stdEmmis	standard deviation of emissions
meanPrc	mean mobility price
stdPrc	standard deviation of prices
mobilityProp	dictionary that maps mobility type ID on the respective properties
mobilityGrowthRates	current growth rates per mobility type

techProgress	current productivities per mobility type
allTimeProduced	aggregated sum of selected mobility types
currKappa	current green infrastructure parameter

Table 3.3: State variables of the market class

The market step function market.step() is called in the global (world) step function. In this function, first the current growth rates g_j (for mobility type j) are determined and the resulting technological progress η_j (for mobility type j) is calculated as

$$\eta_j^{(\text{new})} = \eta_j^{(\text{old})} \cdot (1 + \max\{0, g_j\}).$$

The new green infrastructure parameter currKappa is computed from the input parameter kappa divided by η_{green}. The parameter allTimeProduced is updated by the current sales. Last, the current market statistics are updated, i.e. the new emissions and price mean and standard deviation are calculated.

3.3.1.4 The location class

The location class is not an acting agent, but rather an entity representing the spatial aggregation in the model. It contains all spatial properties that are equal for all agents connected to the locations (e.g., population). The current step of the locations is used to represent the local infrastructure for different mobility types over time and accordingly update the convenience of the different mobility types at each location. The state variables of the location class are listed in Table 3.4.

Variable	Description
position	the (x, y)-position of the location
population	the number of person from the synthetic population living at the location
convenience	the local convenience of mobility types related to the infrastructure

Table 3.4: State variables of the location class

3.3.1.5 The person class

Table 3.5 summarizes the state variables that are necessary to describe the state of a person. Persons have opinions about all mobility types, these opinions are formed based on their own experiences, i.e. the utility they obtain(ed) from the respective mobility mode, and on the opinions of their friends. The variable commUtil is the vector of current (weighted) opinions of the person's friends about all mobility types, selfUtil is the vector of current and past experiences (if applicable). From these two, a person can calculate his expected utility for the different mobility types as a mixture of own and community experience (controlled by the input parameter selfTrust). The expected utility is the criterion for a person to optimize her mobility mode, but after selecting a new mobility mode, the actual utility is evaluated.

The actual utility is a function of the person's priorities $\vec{\alpha}$ and the consequences of the mobility choice. Currently two possible functions can be used,

$$u(\vec{x},\vec{\alpha}) = \sum_i x_i^{\alpha_i} \quad \text{(Cobb} - \text{Douglas)},$$

$$u(\vec{x},\vec{\alpha}) = \left(\sum_i (\alpha_i x_i)^{\frac{s-1}{s}}\right)^{\frac{s}{s-1}} \quad \text{(CES)},$$

where s is an elasticity constant that controls how well different contributions to the utility can be replaced by others. The input parameter "util" (see Section 4.3.1) controls in the simulation is given by the. All properties of the person class are listed in table 3.5

Variable	Description
hhID	ID of the person's household
priorities	tuple of four priorities $\vec{\alpha}$ with α_1: convenience, α_2: ecology, α_3: money, α_4: innovation
gender	gender of the person
age	age of the person
util	utility $u(\vec{x},\vec{\alpha})$, function of the person's priorities $\vec{\alpha}$ and the consequences $\vec{x}$ of the (current) mobility type
commUtil	current social opinion about utility of all mobility types
selfUtil	person's overall opinion (expected utility) of all mobility types
expectedUtil	current expectations for each mobility type
mobType	current selected mobility type of the person
prop	current properties of the current mobility type (emissions and price)
consequences	currently the emissions and price of the selected mobility type
lastAction	time stamp at which the agent has changed/set its mobility type last
ESSR	effective relative number of friends (n_{eff} weighed friends / n friends)
peerBubble Heterogeneity	weighted standard deviation of priorities of friends

Table 3.5: State variables of the person class

The step of the persons accounts for social interactions. The person evaluates all its connections to friends with regard to how useful their input has been to predict the expected utility of different mobility types. To do so, the person compares its current utility, given the current mobility mode, with the opinion of its friends. The differences between the respective utility of the friends and the own utility is normalized by observation error distance (input: utilObsError) and then used to compute a likelihood that opinions are equal. The likelihood is then used to update the connection weights by applying Bayes' Theorem on the prior weights. Thus, over time, the effective network structure changes, since connections with near-zero weights do not contribute anymore to the interactions between persons. Currently, it is optional to delete near-zero weights and to replace them with new connections.

3.3.1.6 The household class

Table 3.6 summarizes the state variables that are necessary to describe the shared state of a household. The persons living in the household contribute to the state as well, but are encoded in the individual person's state.

Variable	Description
pos	household's (x, y)-position
hhSize	number of humans in household
nKids	number of children
income	household's income
expUtil	expected utility as function of the possible mobility type combinations
util	current utility (sum of utilities of all persons in the household)
expenses	current expenses for mobility

Table 3.6: State variables of the household class

In the beginning, for each person in the household it is evaluated whether the person's mobility choice is old enough that she is looking for alternatives to her current mode. This represents the awareness of a person that better mobility choices might be available. The probability p_a that a person is aware depends on the time passed since the last taken actions t_l and the parameter t_{new}, which indicates how long a decision is not contested:

$$p_a = \min\left(1, \max\left(0, \frac{\left(t_l - \frac{t_{new}}{10}\right)}{t_{new}}\right)\right).$$

Within one household, the mobility options of all persons lead to a list of possible combinations, each resulting in a different overall utility for the household.

Each combination is evaluated by summing over the persons' expected utilities for the combination. The combination with the highest sum (expected utility, $U_{e,new}$) is compared with the current utility of the household (U_{curr}). If the expected utility is 5 % above the old one, the new combination is accepted with a probability of $\min(1, (U_{e,new}/U_{curr}) - 1)$, otherwise no action is taken. In case that the combination is accepted, all persons in the household take action to obtain the new mobility type combination, i.e., some acquire a new mobility type. This is registered in the market, which aggregates in the individual sales.

To obtain the actual utility of the household, actual utilities of all persons are summed. A person's utility is a function of consequences and the person's priorities (see Section 3.3.1.5). In the household step, consequences are re-computed according to the current choice of mobility, the location, and the current market state. The vector of consequences $\vec{x}$ consists of the entries convenience (x_1), ecology (x_2), money (x_3), innovation (x_4). Table 3.7 provides the computation formulas.

The consequence "convenience" (x_1) measures the convenience that a mobility mode provides. It depends on the technological state and on the location, in particular, on the degree of urbanisation, which is modelled using population density ρ. A defined density threshold ρ_0 is given as an input parameter, which separates urban and rural areas, as well as an infrastructure parameter for "green" mobility κ and shape parameters A, B, C and D.

The consequence related to "ecology" (x_2) is solely based on the CO_2 emissions produced by the mobility type. ε_G, ε_B, ε_O are the emissions of green and brown cars and other mobility types, respectively, and are dependent on the technological progress at the time of purchase.

The consequence "money" (x_3) is the remaining budget for the household. The budget of the income that is spent for mobility $y(\vec{\psi})$ is controlled by the input parameter mobIncomeShare. Expenses for all mobility types p_G, p_B, p_O are functions of the technological progress at the time of purchase, which is a function of sectoral growth rates (see Section 3.3.1.3). Expenses for mobility of all persons in the household are summed up and the sum is used to compute the remaining share of that money as given in Table 3.7.

The consequence "innovation" (x_4) exemplifies how much the agent has the feeling of using a new innovative technology and is, thus, related to a degree of technical maturity M. Currently this maturity is is defined as the all-time sales of a mobility type S_i, relative to the sum of all-time sales of all mobility types S_{all}:

$$M_i = \frac{S_i}{S_{all}} \quad i \in \text{green, brown, other}$$

Using this maturity factor, the consequence x_4 is calculated as given in the last column of Table 3.7. We assume that the different mobility types have different levels of technical maturity in the beginning of the simulation, and initialize by setting the parameters initialGreen, initialBrown, and initialOther.

type \ x_i	convenience	ecology	money	inno-vation
green	$A - B(\rho - \rho_0)^2 - \kappa(m, \vec{s})$	$\frac{1}{1 + \exp(\sigma_\varepsilon^{-1}(\varepsilon_G(m) - \mu_\varepsilon)}$	$1 - \frac{p_G(m)}{y(\vec{\psi})}$	$1 - \sqrt{M_G}$
brown	A if $\rho < \rho_0$ $A - B(\rho - \rho_0)^2$ if $\rho \geq \rho_0$	$\frac{1}{1 + \exp(\sigma_\varepsilon^{-1}(\varepsilon_B(m) - \mu_\varepsilon)}$	$1 - \frac{p_B(m)}{y(\vec{\psi})}$	$1 - \sqrt{M_B}$
other	$\frac{C}{1 + \exp(-D(\rho - \rho_0))}$	$\frac{1}{1 + \exp(\sigma_\varepsilon^{-1}(\varepsilon_O(m) - \mu_\varepsilon))}$	$1 - \frac{p_O(m, s)}{y(\psi)}$	$1 - \sqrt{M_O}$

Table 3.7: Individual contributions of mobility types a_i to the three consequences x_i

With the new consequences, the new actual utilities for all the persons are computed and summed up to the household's overall utility.

3.3.1.7 Model initialization

At first, the global entity world is instantiated, which is the structure that connects all other entities, controls the time and simulation work flow. In addition, it provides the functionality to create a basic network of locations. The data input for this generation method is a regular lattice of locations which is used to create a network of locations (which might be distributed over different parallel processes). This network is generated connecting locations within a defined interaction radius (connRadius). In the model, established connections between entities represent interactions and, thus, exchange of information. After that, the second global instance, the market, is created and the initial set of mobility types is created. Thereby, each mobility type is defined by price, emissions and a convenience function.

As a next step, the population of each location is imported as a map, and the population numbers are reduced by a given reduction factor (reductionFactor, currently 20). Each location then loads the required number of persons and households from the synthetic population file, which is described in Section 3.3.2.

Households are connected to their location. Each region has its own synthetic population, thus holds a different household and income structure. Similarly, the person entities are created and connected to their households. For each person, a set of priorities (convenience, ecology, money, innovation) is generated randomly according to the properties from the synthetic population (age, gender, income and the number of children).

The last step of the initialization is the generation of the interaction network between persons. This network defines which persons share information with each other. Therefore, for each person i a random sample of persons j in a given spatial distance δ is created. The probability distribution for the sampling process is inversely proportional to the combined distance measure of space Δ^{spat}, priorities Δ^{prio} and income Δ^{inc}:

$$p_j \sim \frac{(\Delta_{i,j}^{spat} + \Delta_{i,j}^{prio} + \Delta_{i,j}^{inc})}{\sum_j^{\{j|\Delta_{i,j}^{spat} \leq \delta\}} (\Delta_{i,j}^{spat} + \Delta_{i,j}^{prio} + \Delta_{i,j}^{inc})}$$

For each person, a random number of social connections between an upper and a lower bound (minFriends, maxFriends) is added. The resulting network forms smooth interconnected niches/communities of similar agents.

Simulations start in the year 2005 with a state that mainly consists of combustion engine cars (“brown” cars) and users of public transport. To distribute mobility mode among persons, an initial dynamic burn-in phase is required to reach a feasible starting point. Thus, before the first time-step, each person is initialized with a random mobility type. This is followed by a few (omniscientBurnIn) omniscient steps, where all persons evaluate all options while fully knowing all consequences. At the end of this phase, all agents are close to their optimal choice. (They might not be at their optimum, because what is optimal for one person depends on the behaviour of all the others.) The short omniscient period ensures that there will not be dramatic fluctuations in the beginning. However, the aim of MoTMo is to represent a system with limited information, and thus this omniscient burn-in phase is followed by a longer burn-

in phase with optimal choice under limited information. These burn-in steps are equivalent to the normal simulation steps, just without progressing in time and without technical change.

3.3.2 Data and Synthetic population

The complex model considers the population of the German federal states Niedersachsen, Hamburg and Bremen. This synthetic population was available from the MIDAS-project as described in D4.4.

Since the offered synthetic population does not include data on people's income, a synthetic income had to be assigned to each household in the population. Data on income distributions is provided by different sources. The database Eurostat provides data on mean equalised net income by age and sex in the European Union from 2005 to 2016.

The World Bank provides distributions that give income decile as percent of the total income in Germany from 1991 to 2010. Thus, the key challenge is to join both types of statistical data without violating any cross-relation. To assign realistic income values to the given households, data from both sources was combined in the following way. From the Eurostat data we have the mean income over all ages and sexes. Multiplied with the number of people in the given population, we get the total income of the population. Applying the World Bank distribution to this, we get a total income for every decile of the population and the mean income, respectively. Every household in the synthetic population is now assigned a household income in two steps. First, the decile the household belongs to is randomly picked where all ten deciles are uniformly distributed. Then within the decile we pick an income value from a normal distribution with the decile mean income as mean and a third of the difference to the next decile's mean income as standard derivation.

So far these randomly assigned household incomes do not take into account household properties like number of people living in it and their sex and age. For this reason, groups are formed according to the labels in the Eurostat data for sexes male and female as well as the age classes 'less than 6y.', '6-11y.', '12-17y.', '18-24y.', '25-49y.', '50-64y.' and '65y. or over'. To consider these, the picked household incomes are reassigned within the list of all households in a way that minimizes the error between the current and expected total income in every age group.

The expected total income per group is the mean income of the group from the Eurostat data multiplied with the number of people in the group. The current total income is calculated as the sum of household incomes for all people belonging to the group. This can be adapted by interchanging the assigned household incomes.

The algorithm therefore switches sets of household incomes, then computes the resulting total income for every age group and compares it with the income per age group coming from the Eurostat data. This continues until the total error is small enough.

PERSION-ID	AGE	SEX	HOUSEHOLD-ID	HH-SIZE	HH-INCOME	LONGITUDE

niedersachsen-1-1	35	1	niedersachsen-1	5	5246.42704	11.34354858
niedersachsen-1-2	33	2	niedersachsen-1	5	5246.42704	11.34354858
niedersachsen-1-3	13	1	niedersachsen-1	5	5246.42704	11.34354858
niedersachsen-1-4	11	1	niedersachsen-1	5	5246.42704	11.34354858
niedersachsen-1-5	6	2	niedersachsen-1	5	5246.42704	11.34354858
niedersachsen-2-6	21	1	niedersachsen-2	1	10079.7359	9.406152273
niedersachsen-3-7	38	1	niedersachsen-3	4	17646.4856	7.60766319
niedersachsen-3-8	37	2	niedersachsen-3	4	17646.4856	7.60766319
niedersachsen-3-9	16	2	niedersachsen-3	4	17646.4856	7.60766319
niedersachsen-3-10	15	2	niedersachsen-3	4	17646.4856	7.60766319
niedersachsen-4-11	58	2	niedersachsen-4	1	9582.93293	8.659073776
niedersachsen-5-12	67	1	niedersachsen-5	2	7127.47095	8.085025180
niedersachsen-5-13	65	2	niedersachsen-5	2	7127.47095	8.085025180
niedersachsen-6-14	41	1	niedersachsen-6	4	13749.6673	7.270101507

Table 3.8: Excerpt of the synthetic population file

As the result of this process we have a synthetic population file which consists of a set of information about every person in the population joint with the information about its household. Different identification numbers are used to identify the person within the population. The most important aspects as given in Table 3.8 are the number of persons living in the household including the person itself, the exact position with longitude and latitude, age, sex, level of education and a variable for the total household income.

3.3.3 Model Implementation

For rapid prototyping, the model was developed within python, using accessible modular tools for parallelization. The ABM-prototyping framework thus developed at the same time bridges the gap between early model prototyping and early-stage production on HPC-resources. The python-based framework enables fast model development for GSS-modellers. When a model has reached a more mature state, it can be implemented in HPC-efficient manner.

The following is not a complete framework description, but rather summarizes the main features. The framework is implemented in python 2.7 and is built on the C-based package numpy, igraph and the HDF5 format for distributed file I/O.

The object-oriented code is structured in a graph-based approach, where nodes are action entities and connections represent interaction. The framework provides a control class called world and two base classes to derive higher level entity types: locations and agents.

These base classes are used to derive agent classes, like regions, cities, households and persons. Ghost classes are available for each basic class, which duplicate “real” agents on other processes for communication between different parallel processes.

Currently, automatic functionality for syncing ghost agents, global variables and computations of global statistics are implemented. To do so, statistics (like, mean, variance, standard deviation, sum) on the individual nodes are computed and the relevant information is shared so that the global statistics are computed and shared between all processes. In addition, automatic storage of agent properties is implemented as collective file I/O using MPI and HDF5. The partitioning of the agents is currently done in a pre-processing step based on locations and their interactions by using the partitioning software Metis.

Figure 3.4 shows exemplarily the runtime required for a single process, split in computing time, time for syncing between processes, time for waiting for other processes and runtime required for I/O.

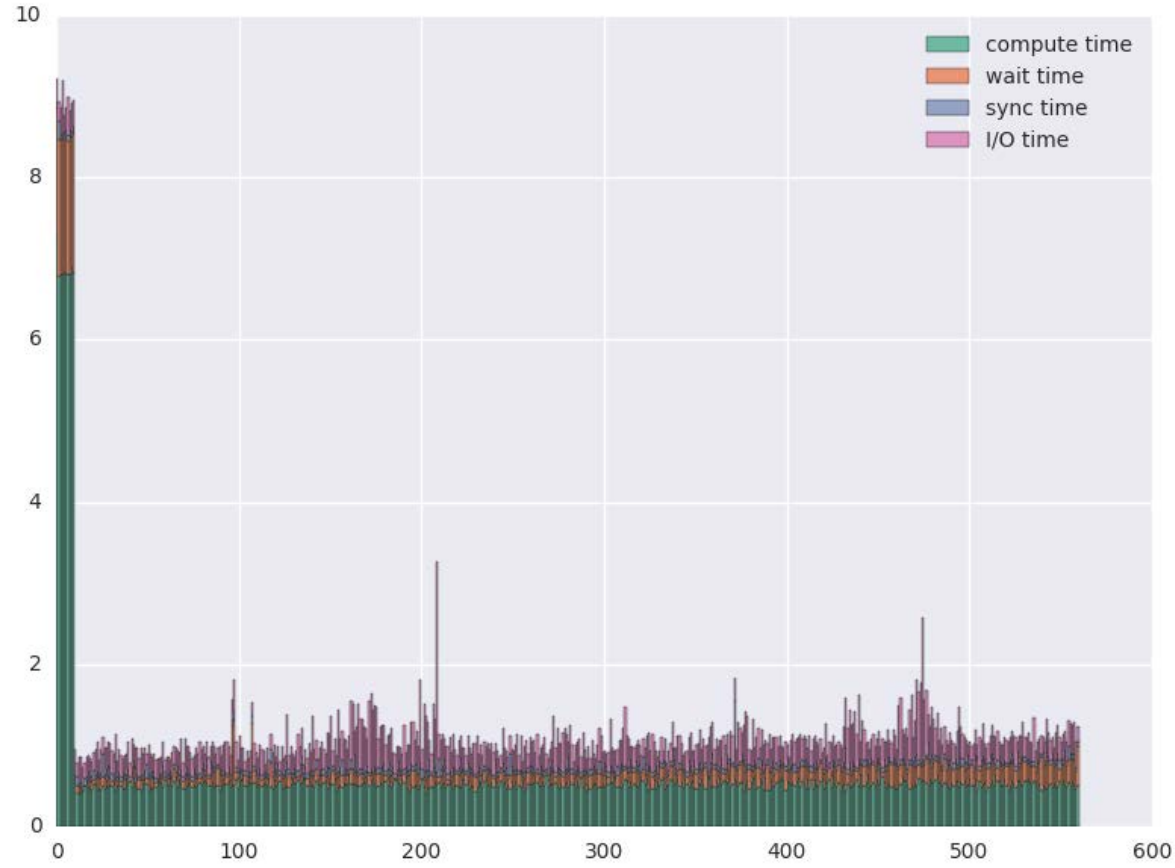


Figure 3.4: Runtime per time step of a single process. The time is separated between computing time (green), waiting time for other processes (orange), time for syncing between processes (blue) and the time that is required for writing input/output (pink). The high times for the firsts ten step are required by the omniscient time step, with a more complex optimization procedure.

3.3.4 Simulations and discussion

Figure 3.5 shows the ratios of selected mobility types for the simulation period (2005–2035) together with the initial burn-in period. The initial distribution is developed in ten omniscient time steps, which is followed by a redistribution due to the switch to the normal simulation

step. Incomplete information leads to different mobility shares and stabilizes till the beginning of the actual simulation start. Only from the year 2020, a fast rise in the numbers of electric cars can be observed.

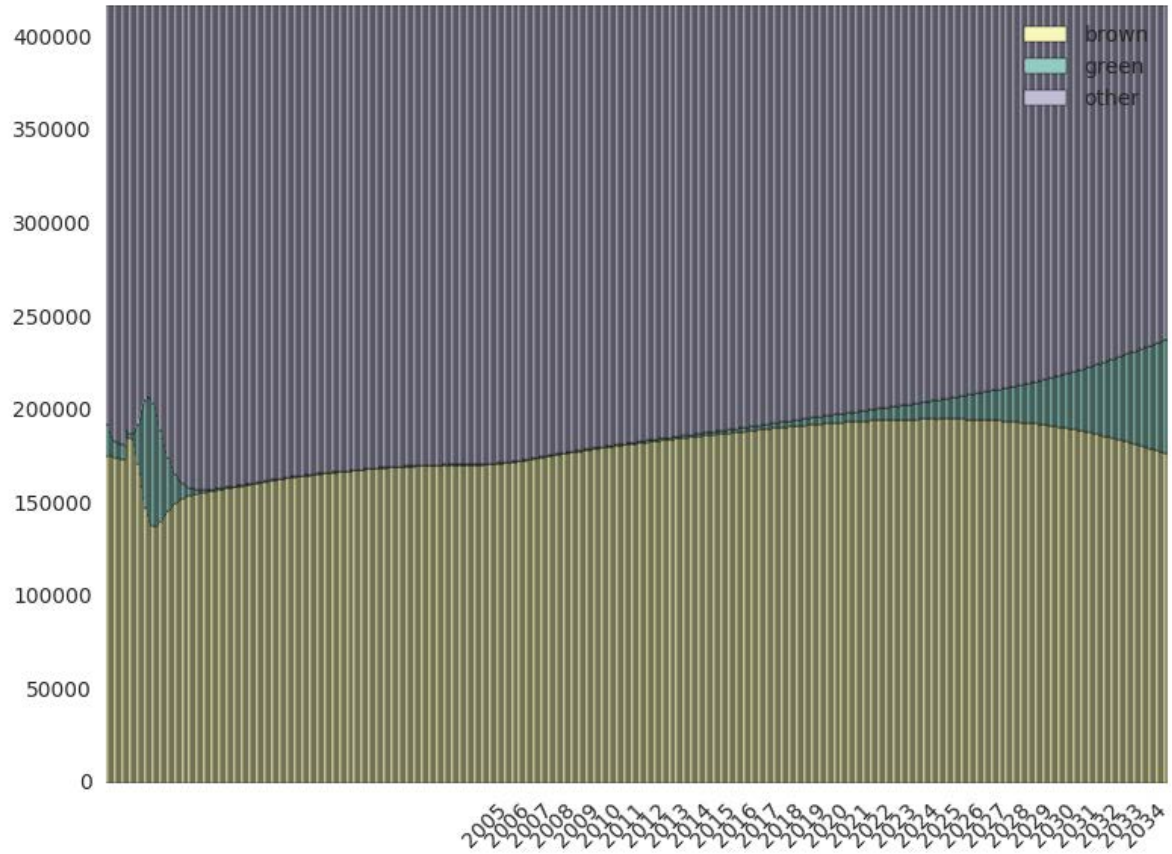


Figure 3.5: Temporal development of the distribution of all mobility types from 2005 until 2035. Left of 2005 the burn-in phase is shown.

Figure 3.6 shows the same numbers separated for the three sub-regions. The scale of the y-axis is logarithmic to allow visualising the very small numbers of electric vehicles. The simulation output is not yet match the data, however, certain important differences between the regions can already be observed and the order of magnitudes resemble the real values. Of course, a calibration process will be required to fit the data more closely, but dynamics and overall pattern can be observed. For example, the urban regions Bremen and Hamburg show higher numbers of electric cars, since the electric infrastructure in urban regions is more developed. For the same reason, the utility of combustion engine cars remains higher in the region of Niedersachsen and the related number of combustion cars does not drop at later stages of the simulation as it can be observed in the urban regions.

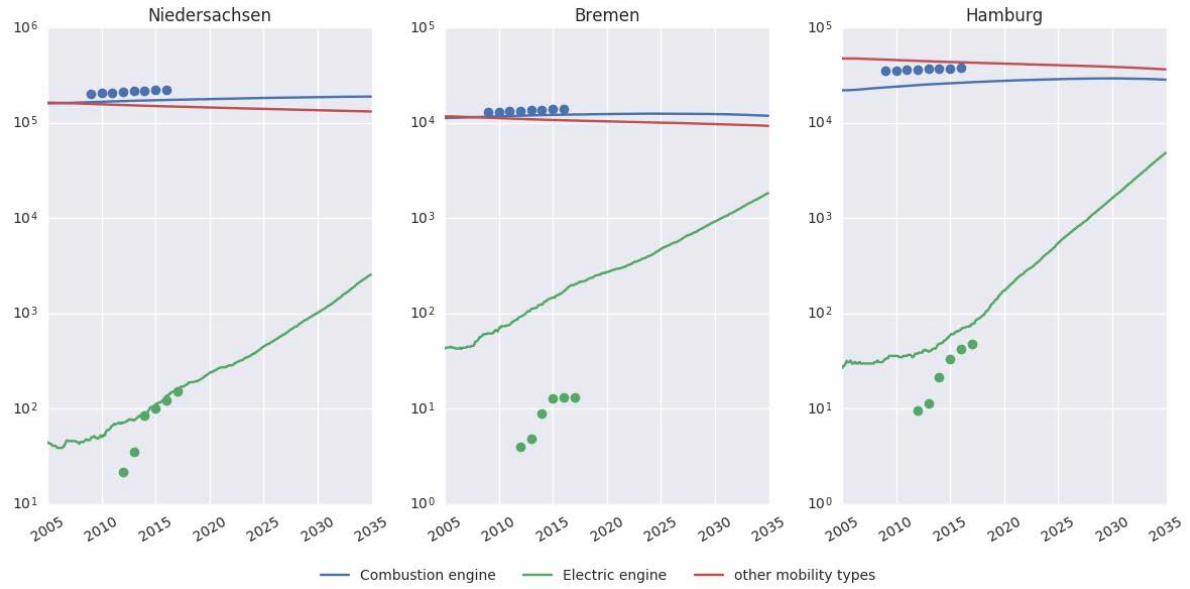


Figure 3.6: Temporal development separated by sub-regions of all mobility types from 2005 until 2030. Calibration data is depicted by dots.

Figure 3.7 shows the spatial distribution of electric vehicles. The four maps depict the development over time that highlights how electric mobility first develops within cities and expand to the surrounding rural areas in later stages. Dominant focal points are the cities of Hamburg, Bremen and Braunschweig, whereas in Hannover, the very dense population favours more the use of other (public) mobility.

Figure 3.8 shows the expected utility of all persons about the respective mobility types. Solid thick lines depict overall averages, thin lines averages over the different preference types, i.e. over all people that consider convenience, ecology, money, or innovation most important. As expected, people with a priority for convenience have on average a better opinion about brown cars than the general mean. It is worth noting that initially the electric mobility is not driven by the ecologically minded people, but by the innovators, since the electric car has initially a worse CO_2-balance than the mobility type “other”. Only with technical improvement, emissions and price of electric cars surpass the alternative and from year 2025 on, ecologically thinking people have higher expectations about the utility of electric cars than the innovators.

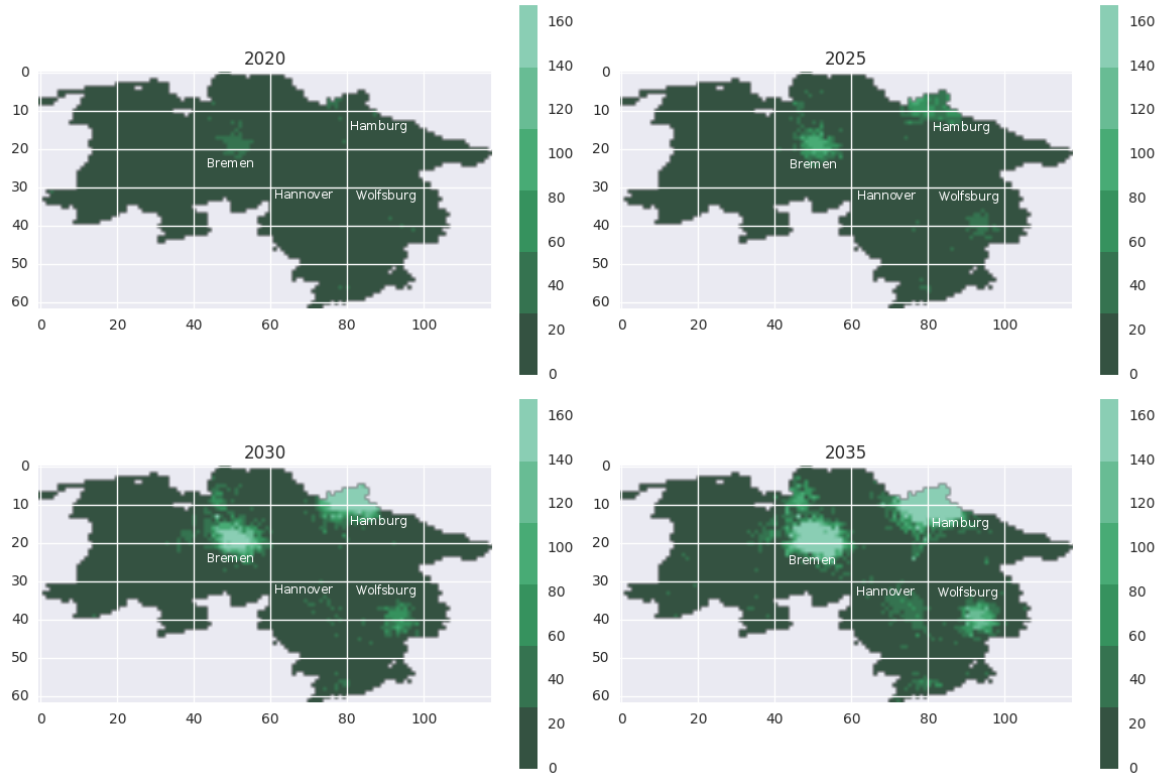


Figure 3.7: Spatial distribution of green cars in the NBH area

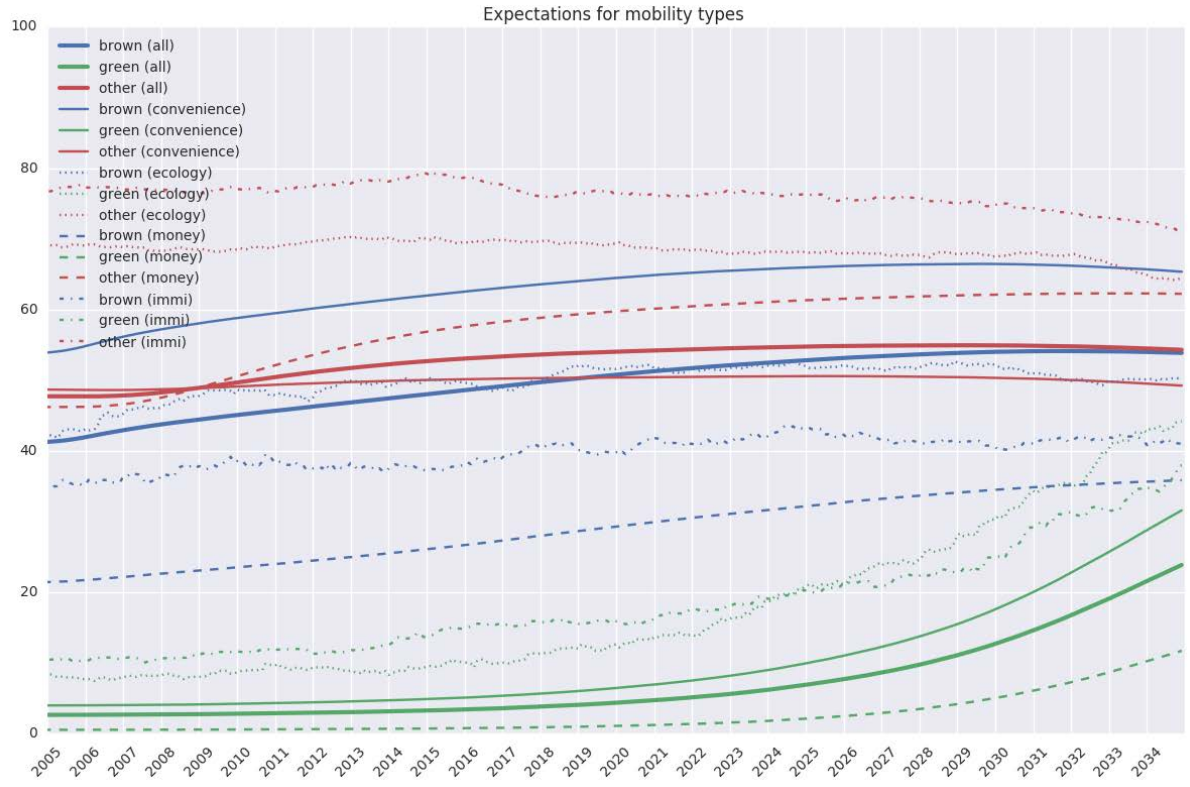


Figure 3.8: Mean expectations about all mobility types: Thick solid lines show the expectations about the three different mobility types averaged over all persons, all thinner lines show averages over people of the same priority type

Figure 3.9 shows in exploratory manner how the synthetic population properties can be used again to show which part of the population is using what mobility mode. Due to the high initial price, only people with high income are able to afford an electric vehicle and therefore, the sub-group of electric car owners (green) has statistically a higher income. Only when prices become comparable, people with lower income switch to electric cars and therefore the average income of people having electric cars decreases. Generally, other (public) mobility is used by people with lower than average income.

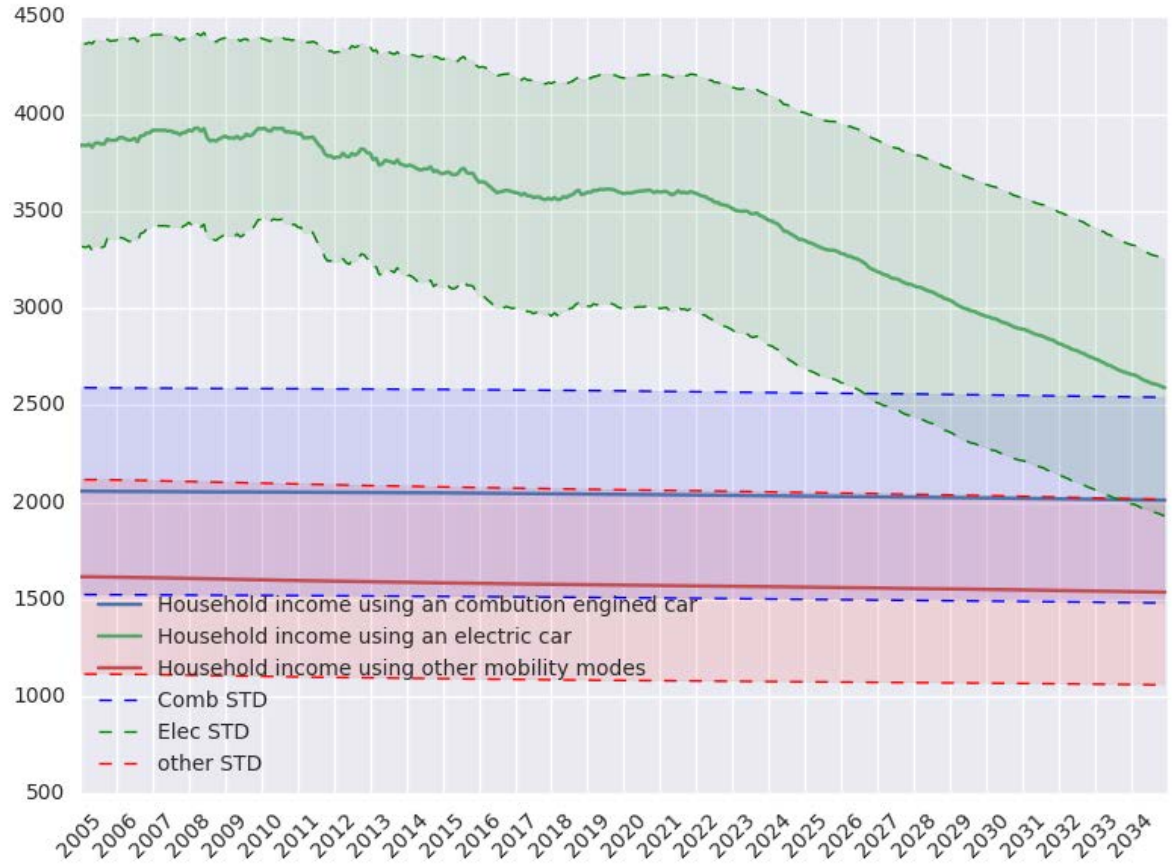


Figure 3.9: Mean income and income spread per mobility type

3.4 Future directions

Next steps for this pilot are ensemble simulations of the complex model to identify sensitive parameters, evaluate the current uncertainty of the model and to test different changes in the available/exchange of information. Thus, the results will be augmented with different scenarios, uncertainty measures and the sensitivity to input parameters. In addition, effects of incentives and political regulations can be tested and evaluated in the following. Furthermore, the load balancing of the current model has to be improved with the help of HPC-experts. The description of the model, the simulation results and the underlying conclusions shall be published. Similarly, the upcoming use of big-data methods for model output analysis shall be described and published in a journal.

There are then various possibilities of refining the model. For example, the social network between agents could be improved in cooperation with the group from IMT. Instead of 3 types of mobility choices (brown, green, and other), further technical detail on mobility options could be included (see e.g., Section 4.2.1.4 in D4.4). Charging infrastructure for electric vehicles could be explicitly included in the information about location, and many more details could be added. Further steps to be taken shall be decided based upon the outcome of the just described next steps envisaged.

4 Status of the Global Urbanisation pilot

4.1 Model overview

This model, further detailed in the requirement deliverables, is based on a synthetic population of inhabitants defined by an income (a fixed input parameter), a housing (which evolves in the course of the simulation) and a working location (assumed not to change), and a transport mode preference (for now: either car or public transport).

While commuting they generate pollution and noise at a level depending on their transport mode choices. These impact the real estate pricing, which is re-evaluated accordingly, leading to relocation of inhabitants (following their income) and therefore to change in commuting travels.

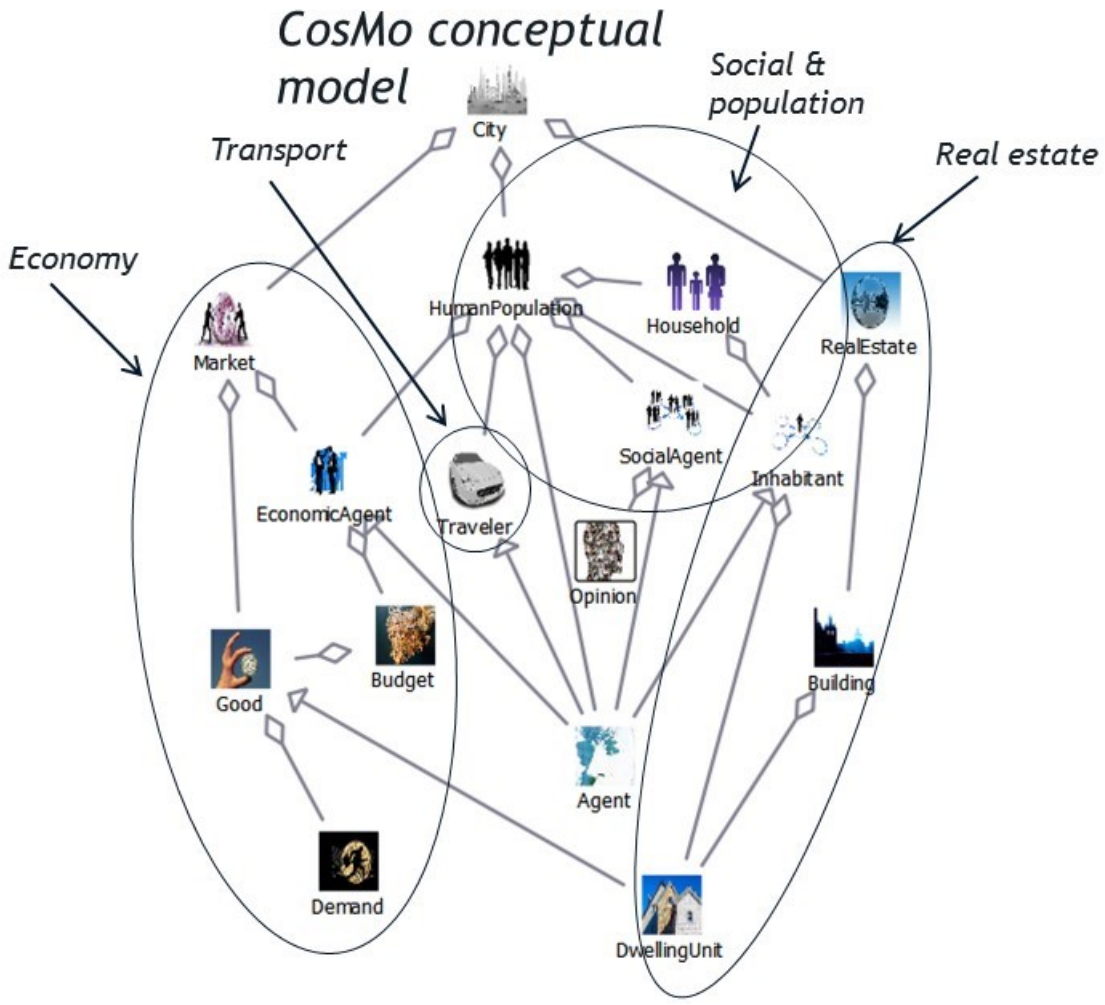


Figure 4.1: Urbanisation conceptual model

The space is split into a regular grid. Every spatial unit of the city is further characterized now by a specific public transport offer, is part of a district (which is a data relevant aggregate spatial unit), holds housing units (with inhabitants) and sees its pollution level (and eventually its specific public transport offer) re-evaluated at every time step.

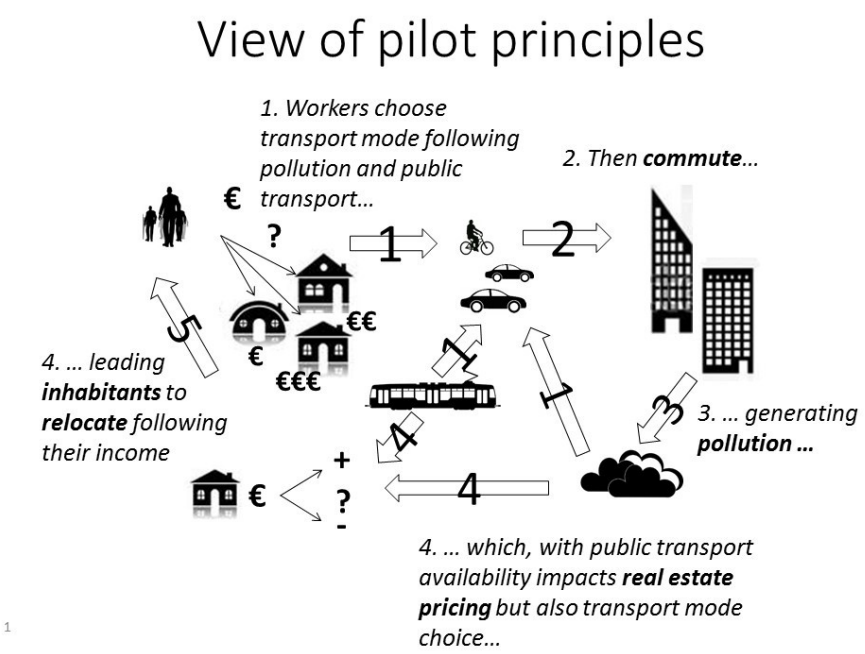


Figure 4.2: Urbanisation model principles

4.2 Model data

This model is based on population data for the city of Paris concerning the number of commuters, and their flows between districts. Pollution generation is based on average car / bus generation values. The link to data concerns therefore the initialization rather than the evolution of prices or transport mode choice.

Indeed, while the focus of this pilot is the two way relation between real estate prices and transport offer, studies show that these interactions are not necessarily the only influence factors of their evolution, driven by general real estate market price raise and characteristics of housing units such as surface, story, neighbouring amenities. Typically the following study (Poulhès 2015) based on Paris data downplays the influence of public transport accessibility and car nuisance on real estate pricing. If after surface and story, it cites the accessibility to workplaces (i.e. percentage of jobs accessible in half an hour) as significant in the price (up to 2% to increase accessibility by a value corresponding to its overall standard deviation), it concerns car accessibility rather than accessibility over public transport (non significant). Nuisance due to traffic (noise) and local distance to public transport (anyway mostly near in Paris as in possibly most cities) on the final price of real estate don't appear to play a major role. Another study on the influence of public transport improvement on real estate pricing in Paris over eight use cases (Nguyen-Luong, Boucq, and scientific advisor: F. Papon 2011) shows mixed results of the influence of public transport evolution on the real estate prices. It highlights the crossed influence of context and various specific parameters, such as the kind of public transport, additional local city developments coming with the works, previous public transport availability, local (social, economic, historical) context but also global real estate market trends). These results make a real-estate precise price validation difficult based on the

only dynamics of the model. Similarly, pollution in a city are due to a great variety of sources (including various travel motivations, local deliveries, transit traffic, heating, industry, ...)

Therefore, the purpose of this model appears rather to allow exploring given scenarios following assumed influence factors, while ignoring price or pollution evolution due to other factors.

4.3 Model evolution

We made the model evolve by improving the transport sub-model, in different ways.

4.3.1 Model evolution purpose

We extended the model following two purposes.

- First, we sought to increase its modelling interest and the challenge of its study (particularly by reinforcing possible feedback), hoping to highlight possible synergies of HPC and GSS, by:
  - increasing the number of parameters to increase the number of possible different evolutions
  - and complexifying further the model to promote feedback and favour complex behaviours which are difficult to predict

  These two kinds of evolution extend the parameter space to be explored. They potentially increase its unpredictability; they therefore promote further the need for refined and thorough exploration facilitated by HPC.
- Second, we aimed at varying the themes, scales and possible scenarios to be studied, and consequently the kinds of stakeholders they might interest. For instance, the present model allows to investigate the influence of
  - institutional decisions such as the offer of public transport
  - social (indirect) interdependence, for instance over real estate prices, generated pollution
  - individual choices such as ecological awareness, transport mode choice, housing preferences, ...

  This opens to test different kinds of scenarios and of leverage on the overall evolution, for instance
  - the influence of public transport availability (increasing or decreasing it)
  - ecological awareness enhancement (for instance over information campaigns)
  - individual choice factors (for instance over incentives)

4.3.2 Model evolution description

4.3.2.1 Overview

The model evolves by reinforcing the role of transport, in different ways.

- Firstly, the real estate price is impacted by transport in two ways. Not only is it negatively impacted by car traffic (noise and pollution) but it is now also positively impacted by the availability of public transport. The influence of these two features is defined by specific parameters allowing to balance them.
- Secondly, the choice of the transport mode by the commuters is now not only an initially fixed parameter, but an individual choice updated at every time step, depending on both observed pollution and public transport availability.
- Thirdly, the public transport offer is open to adapt to demand and evolve endogenously, thereby reinforcing the possibility of feedback loops.

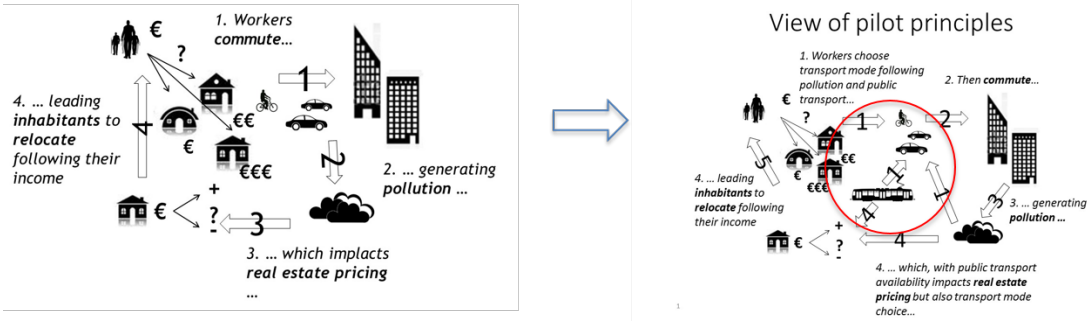


Figure 4.3: Urbanisation model evolution

4.3.2.2 Equations

Price evolution

The evolution of the real estate prices now takes into account the value of a local indicator of public transport availability.

$$price_{i,j}(t+1) = price_mini,j(t+1) + (1 + E_{i,j}(t)) * price\_difi,j(t+1)$$

where

$$price\_min_{i,j}(t+1) = price_{i,j}(t) * (1 - p_{price\_malus})$$

$$price\_max_{i,j}(t+1) = price_{i,j}(t) * (1 + p_{price\_bonus})$$

$$price\_dif_{i,j}(t+1) = price\_max_{i,j}(t+1) - price\_min_{i,j}(t+1)$$

$$E_{i,j}(t) = \frac{p_{ITA} \cdot T_{i,j}(t) - p_{IN} \cdot N_{i,j}(t)}{p_{ITA} + p_{IN}}$$

with the following variables: $E_{i,j}(t)$ is the calculated environment influence on the price, $T_{i,j}(t)$ the public transport availability and$N_{i,j}(t)$ is the nuisance level due to traffic. p are parameters of the model and more particularly, p_{ITA} is the parameter defining the influence of public transport availability and p_{IN} is the parameter defining the influence of nuisance due to traffic (noise and pollution)

Transport mode choice

An agent a switches over to public transport if

$$GreenIncentive_a(t) > p_{GreenThreshold}$$

with

$$GreenIncentive_a(t) = \frac{p_{PTO} \cdot T_a(t) + p_P \cdot N_a(t)}{p_{PTO} + p_P}$$

where p_{PTO} and p_P are parameters defining the influence of the Public Transport Offer and of traffic nuisance, while $Ta(t)$ is the public transport offer as perceived by the agent (average of the offer at his residence and working place) and $Na(t)$ is the traffic nuisance he perceives at his housing location.

A parameter limits the number of changes per iteration to a certain percentage of the population (no limit if it is set to 100%).

Another parameter defines whether reverting to car commuting is possible. If so, it is triggered when the previous condition is no longer met.

Public transport offer

Every spatial unit of the grid holds a level of public transport offer. We allow the public transport offer to adapt to the demand, following a simple equation.

$$T_{i,j}(t+1) = T_{i,j}(t) + p_T \cdot \max(0, (G_{i,j}(t) - T_{i,j}(t)))$$

where $T_{i,j}(t)$ is the public transport offer in the unit (i,j) of the spatial grid, $G_{i,j}(t)$ is the percentage of green commuters living in this spatial unit, and p_T is a parameter defining the adaptability of public transport offer to the demand.

4.4 Simulations

Figure 4.4: City: Paris

4.4.1 Simulation purpose

The city studied is Paris. Our purpose is to explore the influence of the new features on the dynamics, and particularly increased feedback. Therefore, we study three different kinds of simulations following the new features.

First, we assume that real estate and transport mode is influenced principally by public transport availability, defined as a fixed parameter of the model.

Second, the choice of the transport mode by the commuters is now not only an initially fixed parameter, but an individual choice updated at every time step, depending both on observed pollution and public transport availability.

Third, the public transport offer is open to adapt to demand and evolve endogenously, thereby reinforcing the possibility of feedback loops.

4.4.2 Simulation results

4.4.2.1 First simulation set: transport mode influenced principally by public transport availability

In this first simulation set, the transport mode is influenced by public transport availability. We test different levels of ecological awareness (which influences how easily commuters choose to move over to public transport). Furthermore, we simulate different scenarios where real estate pricing is influenced mainly by the proximity of public transport availability, or by the level of pollution, or by both.

Appropriate parameter values (high enough sensitivity to pollution and public transport availability) allow green transport mode choice to spread and consequently to reduce pollution.

In the following two figures, we show how ecological awareness firstly influences the final percentage of 'green' commuters preferring public transport to their car. Secondly how it allows consequently decreasing pollution. This influence, if intuitively monotonic, is not linear. Furthermore, we can see that prices influenced mainly by pollution lead to higher percentages of commuters choosing public transport than prices influenced mainly by public transport offer; the scenario where prices are influenced both by the public transport offer and pollution leads to intermediate values.

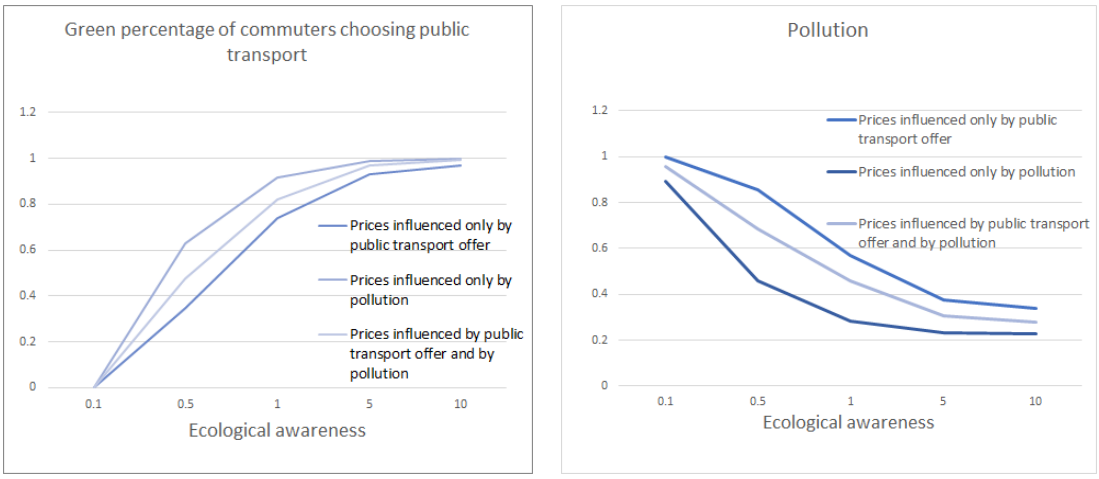


Figure 4.5: Public transport mode choice and pollution, following the ecological awareness, for difference pricing scenarios, in the first simulation set

4.4.2.2 Second simulation set: introducing the influence of evolving pollution (weak feedback)

Overview

In this second set of simulations, the choice of the transport mode evolves during the simulation following the level of pollution, leading to possible feedback and self-regulation. We also assume that commuters do not revert from choosing public transport (typically because they have subscribed over a long period to minimize daily price).

We display the results when setting the parameter value high enough to lead to effective feedback.

In this second simulation, we so observe a successful feedback loop leading to decrease the pollution in the city and to the choice of public transport to predominate.

This kind of scenario allows assessing for instance the influence of an awareness campaign promoting public transport to help reduce pollution.

Detailed simulation results

Hereafter we show the evolution of various indicators over monthly time steps in the simulation.

Pollution

The pollution varies spatially following the commuting flows. During the simulation, it changes following the relocation of commuters and the evolution of their daily travel, but also following the spreading of green choice of public transport, leading to less car traffic and therefore pollution. Here we can see that its peak moves spatially before diminishing significantly, starting with the areas where public offer is the highest. (The diagonal pollution pattern can be interpreted as a bias due to major commuting flows and the spatial grid description.)

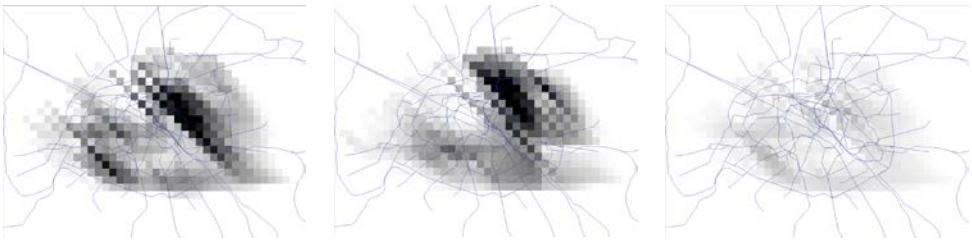


Figure 4.6: Evolution of simulated pollution in the city over time (months 1, 4 and 7) (second simulation set)

Green commuters

The level of pollution and public transport availability leads commuters to choose increasingly public transport over car, starting with the centre, near to the most polluted areas before spreading to polluted areas south west and then more widely. The commuters with the best-connected public transport offer (in the North West) are the first ones to become green, leading to locally decrease the pollution. The choice of public transport increases around, and more particularly in the polluted South West, with slightly less public transport offer. High commuting flows between North and East, only covered by lower public transport offer, lead first to consequent pollution. However, as the percentage of commuters to choose public transport increases, the pollution diminishes consequently.

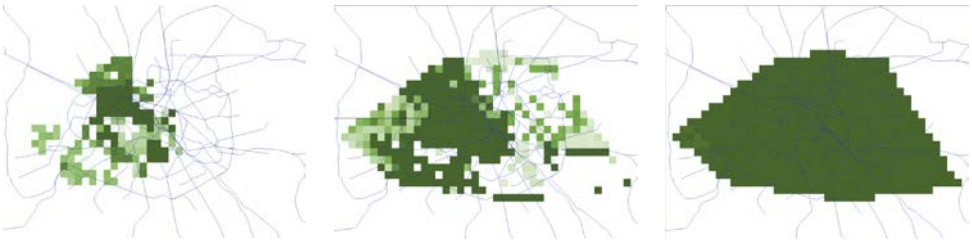


Figure 4.7: Evolution of simulated green commuters in the city over time (months 1, 4 and 7) (second simulation set)

Real estate price

The real estate prices do no vary here much, not playing an essential role in the dynamics and not being much influenced by them. However, while being highly influenced by their district, due to other pricing factors, we can observe that they change following the influence of observed pollution. They first increase in the areas where public transport availability is the highest, encouraging commuters not to take their car and therefore diminishing pollution. The increase in prices spreads with the decrease in pollution, even to the North East.

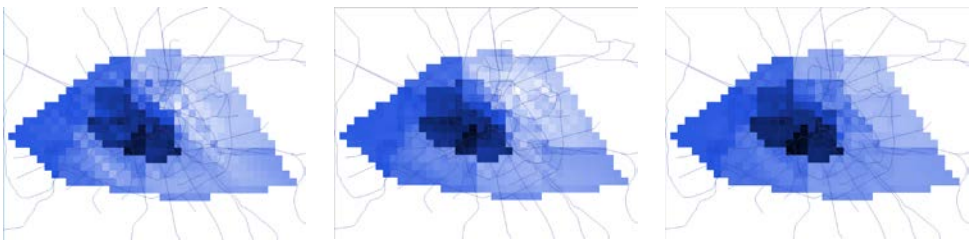


Figure 4.8: Evolution of simulated real estate prices in the city over time (months 1, 4 and 7) (second simulation set)

4.4.2.3 Third simulation set: endogenous public transport offer leading to reciprocal influence loops (strong feedback)

Overview

In this new set of simulations, we observe two new kinds of evolutions (by setting appropriately parameter values).

First, we assume that the offer of public transport adapts to the demand. As previously, choosing public transport is promoted by a good public transport offer. Second, we allow green commuters to switch back to car transport.

Therefore, after having observed in the second set of simulations how green transport mode choice spreads simply over the city, we observe here how (in the same time frame) it progressively converges towards an intermediate value, displaying spatial heterogeneity.

Detailed simulation results

Green commuters

The awareness of pollution impact and the availability of public transport lead commuters to prefer green transport mode to individual car. However, due to possible return to car commuting, green mode choice is less widespread than previously, and its increase is spatially more varied. Here we see again that well connected public transport in the West encourage commuters to prefer public transport over their car, increasing from North to South, before spreading to the East.

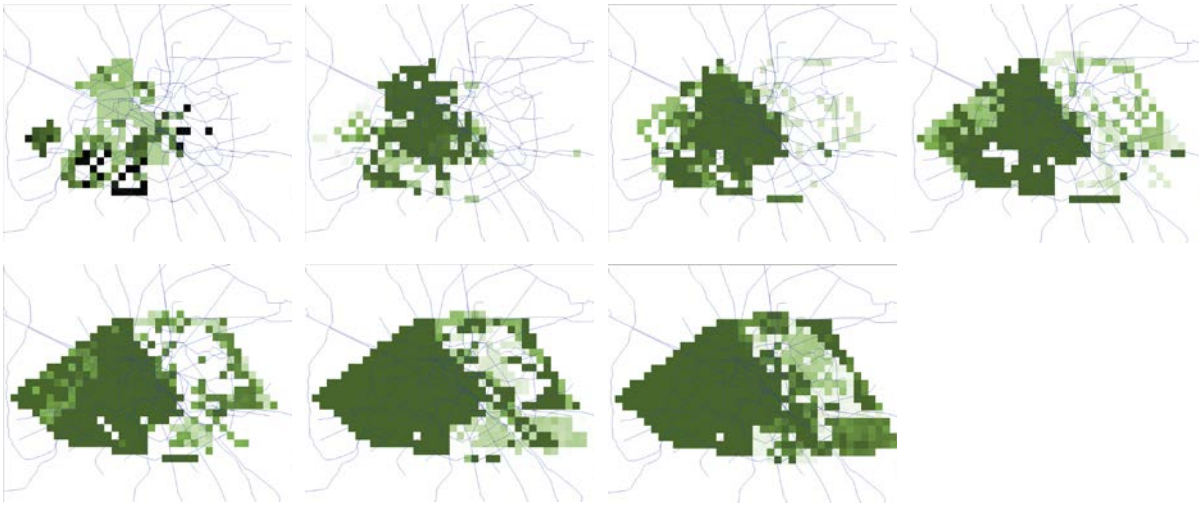


Figure 4.9: Evolution of green transport mode choice in the city over time

Public transport offer

Due to increased demand for public transport, the offer adapts, spreading from the centre to the outskirts. The public transport offer follows the demand of commuters. Therefore, it first increases in the West, from North to South, then to the East, levelling the discrepancies and favouring therefore a switch of the commuters from their car to public transport. Here, at the opposite to the second set of simulations, commuters can choose to return to car commuting, for instance if pollution decreases. However, we can see that the improved public transport offer allows to stabilize the preference for public transport and consequently to lastingly reduce pollution.

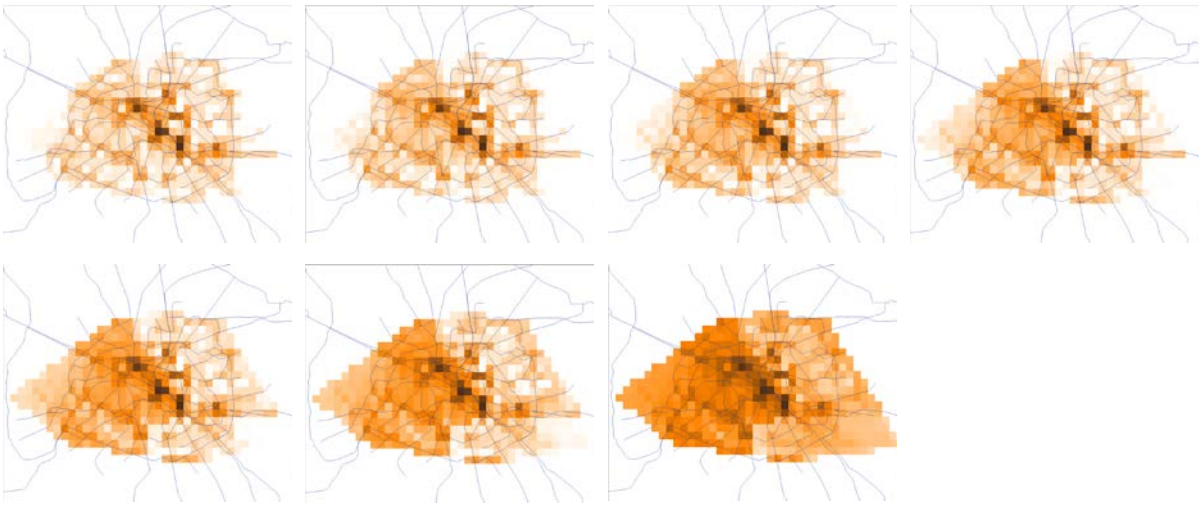


Figure 4.10: Evolution of public transport offer in the city over time

Pollution

The pollution evolves and diminishes due to less car commuting, but stays longest present in the East, where there are less green travellers, and where, consequently, the public transport offer has increased less than in other parts of the city. It decreases however finally significantly. Due to the possibility for commuters to revert back from public transports to car commuting, the decrease is less quick than in the second simulations. However, it stabilizes, similarly to the preference for public transport, thanks, particularly, to the adaptability of the public transport offer.

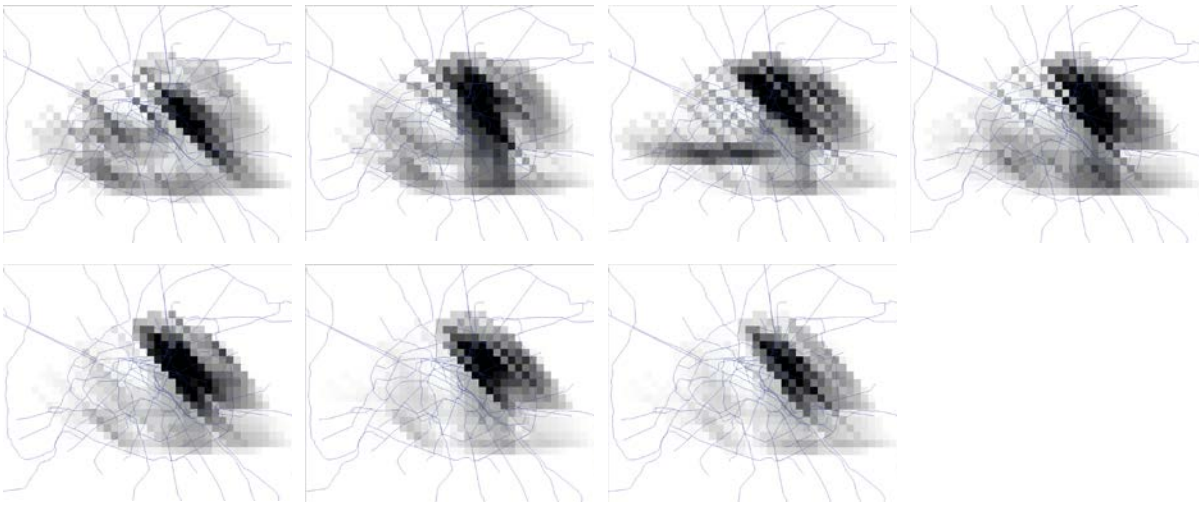


Figure 4.11: Evolution of simulated pollution in the city over time

Real estate price

The real estate does not change much in this simulation, due to parameter values focusing on mainly transport linked dynamics.

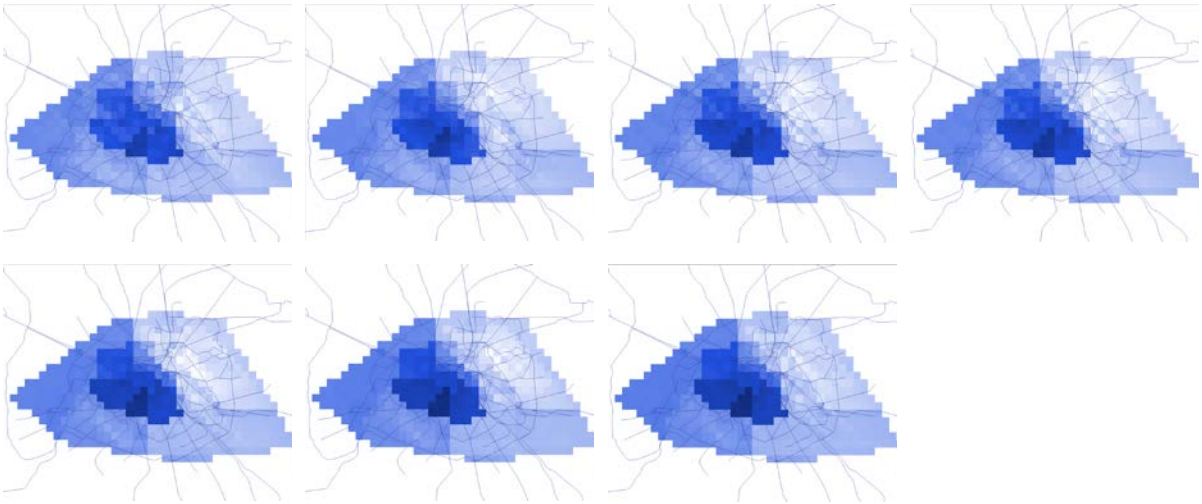


Figure 4.12: Evolution of simulated real estate prices in the city over time

4.4.3 Calculating various indicators

In this section, we present various high level indicators calculated to clarify simulation results at different scales, and to allow comparing them, putting into light possible benefits of HPDA.

4.4.3.1 High level indicators:

In this first section, we calculate different key performance indicators in the parameter space over a set of simulations. Their purpose is to highlight the model's overall behaviour and the influence of two key parameters on the evolution: at the individual level, the ecological awareness, and at the institutional level, the adaptability of the public transport offer to an increase in demand.

Key performance indicators over the parameter space 3D view

The following three-dimensional charts show the value of different key performance indicators (average pollution, real estate prices, percentage of green commuters, public transport offer) following the value of ecological awareness of individuals, and of institutional public transport adaptability to increased demand. This kind of visualization shows the sensitivity of a given indicator to various parameters, not only individually, but when combining their influence. For instance, here it allows to see whether an increase in ecological awareness is sufficient to reduce pollution, even without increased public transport offer, or how adaptability of public transport offer can increase the effect of enhanced ecological awareness.

Pollution

We can see that the pollution depends in the first place on ecological awareness, however not in a linear way. The adaptability of the public transport offer plays a role for intermediate values of ecological awareness, where it promotes green modes and leads to decrease pollution.

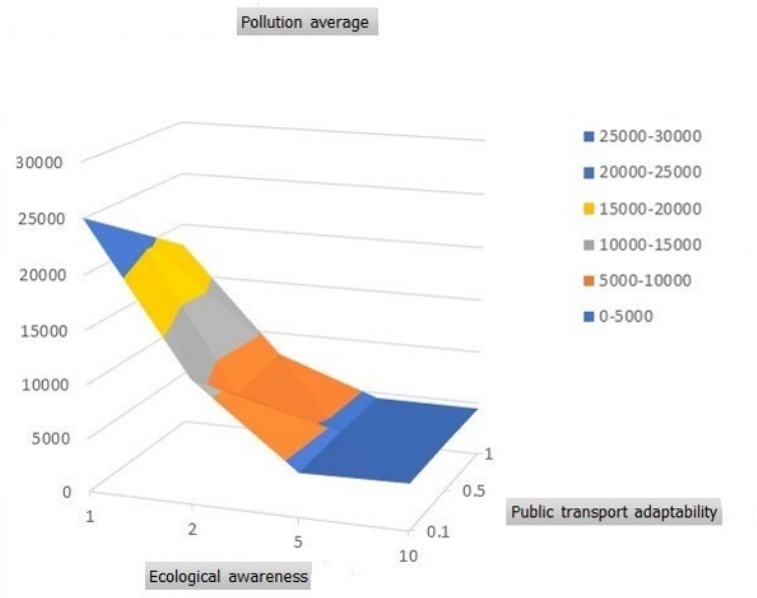


Figure 4.13: 3D parameter space view of pollution

Real estate prices

Similarly, real estate prices depend mainly on ecological awareness and generated pollution.

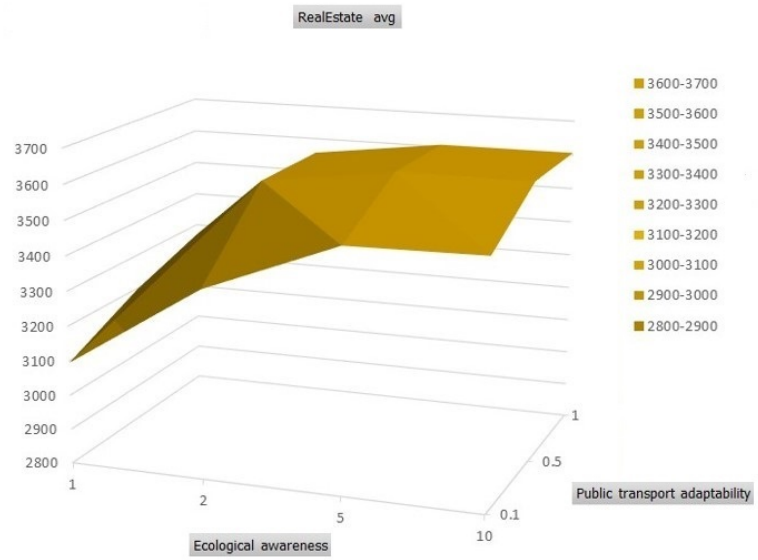


Figure 4.14: 3D parameter space view of real estate prices

Green commuters

Green commuters depend also mostly on ecological awareness level, however with a sharper increase and wider maximal value than pollution.

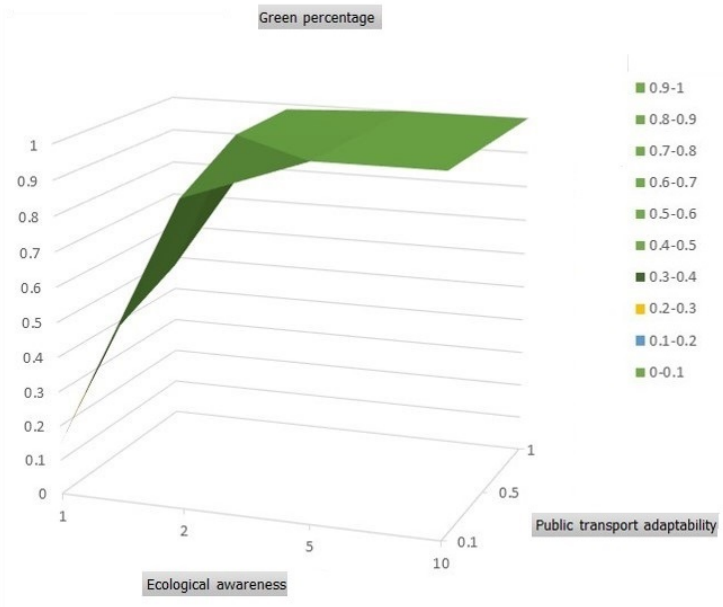


Figure 4.15: 3D parameter space view of green commuters

Public transport offer

In the following graph, we show the increase in public transport offer normalized by its initial value. The public transport offer depends as expected both on ecological awareness leading to increased demand, and on institutional adaptability.

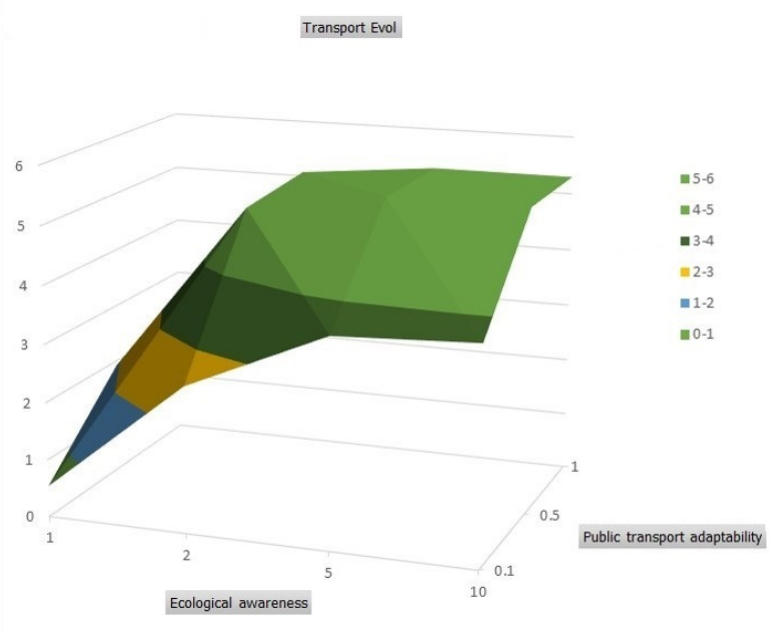


Figure 4.16: 3D parameter space view of public transport offer evolution

Key performance indicators over the parameter space 2D isometric view

In the following section, we show a two-dimensional view of the parameter space while tracing isometric values of the indicators. This kind of visualization allows to see more precisely which parameter values can lead to an 'acceptable' indicator value. It would allow to answer questions such as which strategies combining promoting ecological awareness and enhanced public transport offer can lead to acceptable levels of pollution or of real estate pricing evolution.

Pollution

In the following graph we show the ratio of the observed pollution for a given scenario, and its minimal value over all scenarios. The pollution depends mainly on the level of ecological awareness. However, we can see that for intermediate values of ecological awareness, public transport adaptability can contribute significantly in promoting green travel behaviours and reducing pollution.

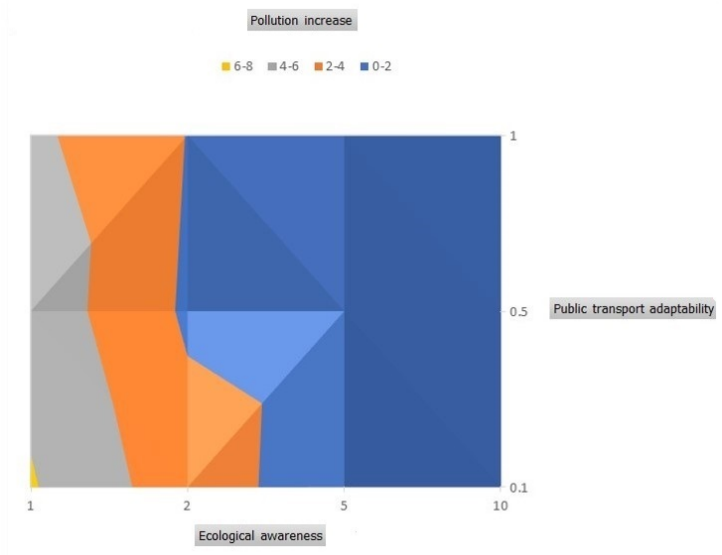


Figure 4.17: 2D isometric parameter space view of pollution addition

Real estate prices

In the following graph we show the evolution of the average real estate price between its value at the beginning and the end of the simulation, normalized by its initial value. Real estate prices are negatively impacted by pollution (influenced by ecological awareness) and positively by availability of public transport. Therefore, highest values need both to minimize pollution (and therefore, maximize ecological awareness), and maximize public transport offer.

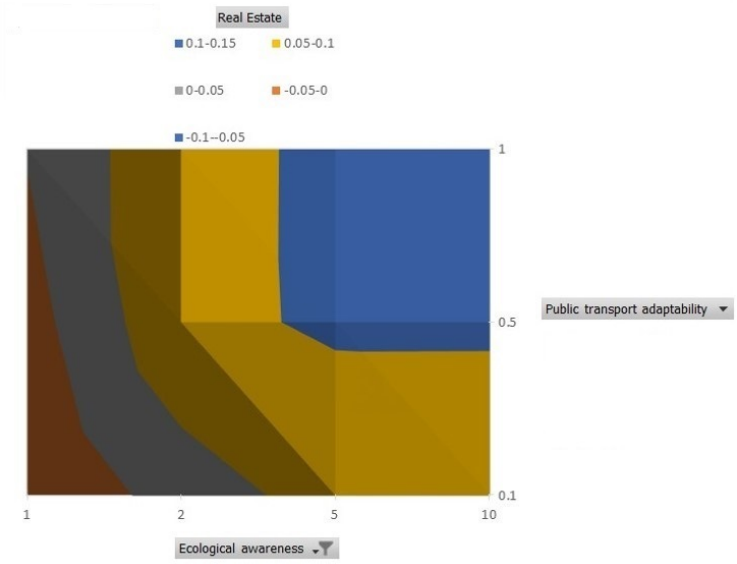


Figure 4.18: 2D isometric parameter space view of real estate prices

Green commuters

The percentage of green commuters depends expectedly on ecological awareness. However for intermediate values, it can be highly increased by an adaptable public transport offer.

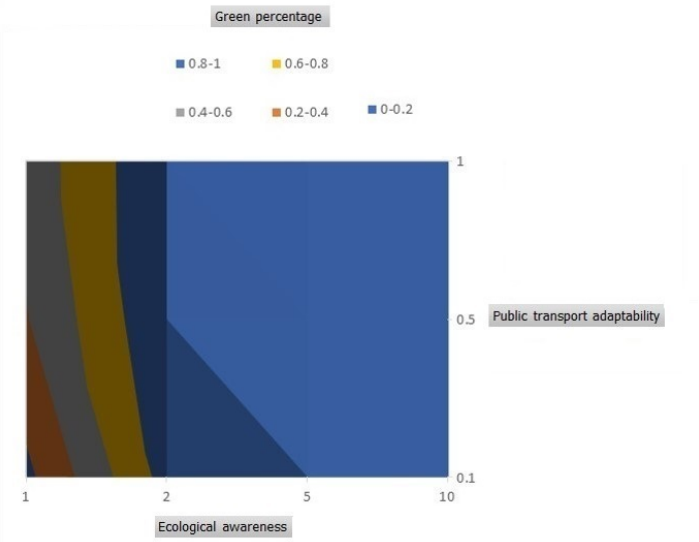


Figure 4.19: 2D isometric parameter space view of green commuters

Public transport offer

In the following 2D graph, we show an isometric view of the increase in public transport offer normalized by its initial value. The final public transport offer depends on its adaptability but is also importantly driven by the shift in behaviours allowed by ecological awareness.

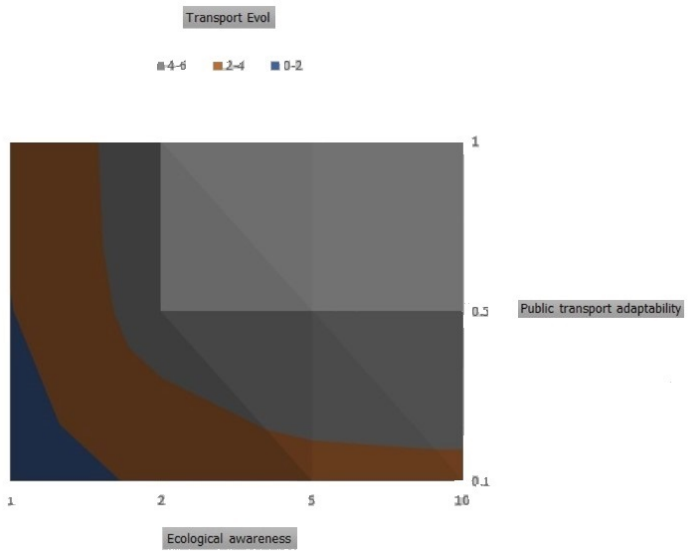


Figure 4.20: 2D isometric parameter space view of public transport offer evolution

4.4.3.2 Investigating further the difference between simulation results:

We have completed the simulations of the first year by exploring difference observed following the granularity of the initialization data for the real-estate prices: either per district, or interpolated over a regular grid. We explore further three transport scenarios: business as

usual (BAU), corresponding to 50% of car commuters, half BAU, corresponding to 25% of car commuters, and a scenario where all commuters take their car. We have calculated indicators at three different levels to assess this difference: the raw spatial difference at the end of the simulation, a first statistical overview on average and standard deviation values (following various scenarios), and finally statistical measures of spatial heterogeneity.

Spatial difference

These first observations highlight that the difference displays non-trivial spatial heterogeneity (either negative or positive (shades of red or blue)). It varies further following the agent grain and the transport mode scenario. To summarize, coarser agent grain favours greater differences between simulations corresponding to different initialization data grains. Furthermore, observed difference depends also on the transport mode scenario, which leads to more or less evolution, in a non-trivial way, calling to refine analytics. (Indeed, whereas in a 50-persons agent granularity a lower level of car commuters leads to increased observed heterogeneity, this is less so in a 10-persons agent granularity for instance.)

This puts into light the benefit of the calculation and visualisation of this indicator to better assess the implications of data and agent grains modelling choices. Furthermore, it puts into light the risk of approximating either the initialization data or the agent grain, emphasizing the benefit of HPC / HPDA allowing to refine them.

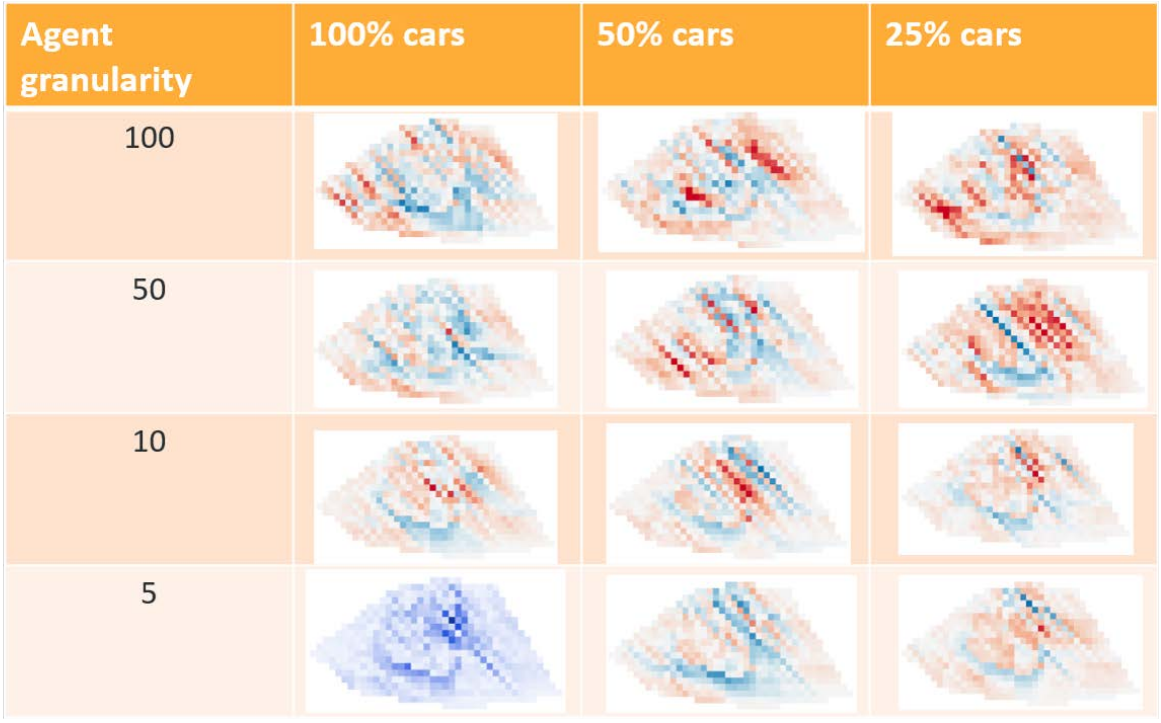


Figure 4.21: Difference in simulated pollution between interpolated and district initialization

Difference basic statistics

We then calculated basic statistics following parameter values, to assess further to which extent the initialization data, but also agent grain and transport scenario can influence the simulation results at a more aggregate scale, and whether difference might appear there to even out. In the scenario where all commuters take their car, the average difference increases

with the agent grain. This trend is less clear for the other scenarios, suggesting that the agent grain is less important when observing aggregate results, even if the standard deviation of the difference, revealing a fine-grained heterogeneity, appears in almost all cases to increase with the agent grain.

Here we see these basic statistic, too aggregated indicators might make us miss the spatial heterogeneity pointed by the previous finer indicators, and its sensitivity to agent grain.

These results emphasize the necessity to multiply fine grained indicators (here the spatial difference) rather than too coarse one (as the sole average and standard deviation here), and therefore the benefit of HPC / HPDA, which can easily retrieve detailed results and calculate various indicators over them.

Figure 4.22: Average and standard deviation of difference in simulated pollution between interpolated and district initialization, following agent grain and transport scenario

Spatial difference statistics

We further calculated C-Geary and Moran indicators, which indicate spatial heterogeneity.

Geary's C is defined as

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

Moran's indicator is defined as

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

The w_{ij} allow to define the neighbourhood to be considered.

This third set of indicators, by providing new insights, highlights further the need for GSS not only of detailed simulations, and detailed analysis of results but also of varied analytics, allowed by HPC / HPDA. This study therefore puts into light further their possible benefit to meet GSS specific needs.

We calculated these indicators of the difference here again for various agent grains and transport, but varying additionally the neighbourhood (N) (range of adjacent cells taken into account in the indicator calculation).

These results complement the previous ones, by calculating overall (so not just the first visual impression of the first indicator) yet fine grain (and not aggregate as for the second indicator) heterogeneity.

Particularly these indicators show the specificity of certain cases, such as the quarter cars transport scenario.

Figure 4.23: C Geary spatial heterogeneity indicator following calculation neighbourhood and transport scenario

Due to close values, they do not vary much. However, we can observe that their sensitivity to agent granularity depends on the scenario.

Figure 4.24: Moran spatial heterogeneity indicator following calculation neighbourhood and transport scenario

4.5 GSS and HPC synergies

4.5.1 Putting into light different GSS and HPC synergies

After having explored previously a first set of possible benefits of HPC for GSS (improving data and agent granularity) we have now focused on multiplying simulations of a model with an increasingly complex behaviour, and on multiplying scales, indicators, insights on simulation results.

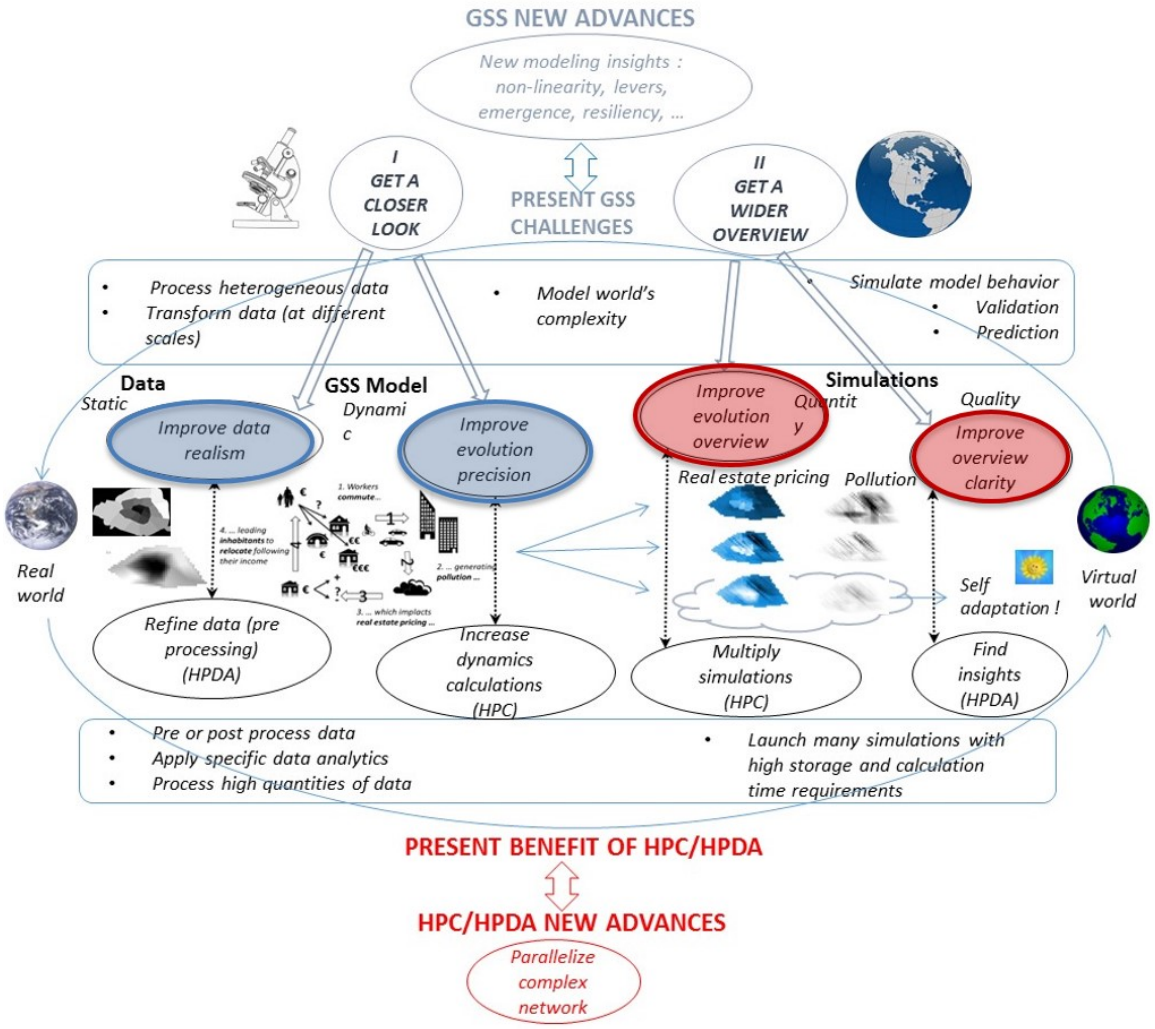


Figure 4.25: Synergies between GSS and HPC

4.5.2 HPC tests

After difficulties due to library versions, the first tests of cosmo simulation suite have successfully passed on Poznan's Eagle, validating most of more than 1000 tests. Time benchmarking is yet to be carried out.

4.6 Summary and future directions

4.6.1 Summary

We have focused on putting into light synergies between GSS and HPC focused on sets of simulations and their analysis rather than on refining data and agent grain as previously.

We have extended the model to

- Make it interesting to other stakeholders, by allowing studying more factors of influence and scenarios at different scales (institutional, social, individual).
- Complexify it to further put into light the benefit of HPC to simulate it.

We have explored this model following scenarios of increasing complexity.

We have calculated different kinds of indicators.

- Differences between simulations, monitored from different points of view:
  - spatial observed difference
  - basic statistics
  - statistical spatial heterogeneity
- Simulation results of various indicators (pollution, real-estate pricing, transport mode choice)
- Parameter space visualization for the value of various indicators
  - 2D visualization to define precise zone of acceptable values for a given indicator, following parameter values
  - 3D visualization to see how the influence of more than one parameter influenced

4.6.2 Future directions

Future directions are to focus on

- HPC scalability tests
- The possible benefit of calculating analytics over sets of simulation results, the systematically allowed by HPC opening the hope of uncovering insights more classical approaches would not have permitted even to investigate.

5 Future Applications

To identify needs and opportunities for future HPC applications to address global challenges, Task 4.4 has continued scoping exercises (Section 5.1) and taken a closer look at the previously identified potential future application on financial systems and applications of block chain technologies for global challenges (Section 5.2) in the second project year. Both streams of work shall be continued throughout the last year of the CoeGSS project. In particular, the next step will be the upcoming "International Conference: Computing Power for Global Challenges" (see D6.6) that shall foster discussions between HPC experts, GSS researchers and practitioners. The main topic will be how to apply HPC for obtaining better decision support around the global challenges of developing a sustainable and resilient global financial system, addressing the daunting risks of pandemics, transforming the fossil-fuel based global mobility system, and creating forms of democracy adequate to the age of digitalization.

5.1 Scoping

Several events where useful for looking into potential future application fields of HPC in GSS: the Social Simulation Conference 2016 in Rome, the Conference on Complex Systems 2016 in Amsterdam[10], and in particular, the International Conference on Synthetic Populations that CoeGSS co-organized in February 2017.

During the first two events, partners collected information in personal conversations and with the help of questionnaires. Several modellers, working on topics from complex social network analysis via statistical physics of complex networks and diffusion of innovations in social networks to household decision making on PV cells and insulation, pointed out the need to reduce model complexity due to limitations of memory or computing time, or stated they would like to use HPC to upscale a network or population under study.

Accelerating computations was also mentioned as a point of interest, in particular, real-time simulation was discussed as helpful for decision support; an example case discussed was helping to steer crowds around large sports events, using real-time data from a mobile phone company.

Combining the fact that (social) networks play important roles in the dynamics of most, if not all, social systems analysed in GSS, and the interest of modellers in larger computing power to research these networks, suggests that there may be many GSS applications that could benefit from being able to move their models to HPC.

A closer look at some potential applications was then obtained in the third event that is described in more detail in D6.6.

[10] Both these conferences took place in the last weeks of the first project year. As this did not leave the time to include the information obtained during these events in D4.4, it is reported here.

5.2 Financial sector and block chain

A session on financial systems during the International Conference on Synthetic Populations provided further insights as to how and why the CoeGSS approach of synthetic information systems can be relevant and useful for addressing challenges related to the global financial system.

In particular, to investigate the financial system in detail the synthetic population approach could fill important gaps. Due to the high confidentiality of data and high privacy, only partial data is available. Here, synthetic populations of banks and firms can preserve privacy, but yet provide statistical correctness, by generating plausible joint distributions of financial exposures, at the sector and firm level. Also, the generation of network architectures compatible with balance sheet constraints is desirable here. Network reconstruction methods are already applied in this field. Network effects matter in questions of financial stability (e.g., is risk diversified, are shocks absorbed or amplified?), of information asymmetries or collective moral hazard (e.g., how does the knowledge of being too connected to fail influence decision making?), and others. Therefore, modelling needs to explicitly consider the micro level, and one specificity of these networks is that different types of edges may exist between actors, resembling different kinds of financial contracts between them.

Experts on financial network modelling consider systematic modelling at the EU level necessary, and see an upcoming need of HPC in this field, related, for example, to laying microscopic foundations for understanding a complex system like trade, where many people may create unexpected behaviours of the overall system (like crashes) without central coordination. The topic shall be deepened in the upcoming International Conference Computing Power for Global Challenges in October with the help of a panel discussion and a workshop involving HPC experts on HPC, and researchers as well as practitioners dealing with the global financial system.

In personal communications, the block chain technology was further discussed as a technological advancement that has the potential to fundamentally change society. Beginning in the financial sector, block chain-based ICT is used to surpass central control instances (of trust) like banks. Bitcoin is currently the most successful example, but many other currencies have been created. Apart from economic implications (omission of banks and their influence), this also bears many social consequences, for example, that trading might be possible on much smaller and faster levels, as well as better transparency for customers. Since block chain allows intrinsic enforcement of rules, the opportunities of fraud and exploiting central systems might be drastically limited. GSS models on HPC scale might be able capture some insights of such decentralized systems by comparing decentralized and centralized ABMs of trade and banking. This topic shall also be followed up on during the upcoming GSS-HPC conference in Lucca in October 2017.

In addition, many currently centralized structures might change due to block chain. Currently, the very tightly controlled electric sector is changing due to renewable energy production, and decentralized marketing concepts might help to overcome some problems. The currently very

inflexible market structures might be replaced partly be direct and decentralized market structures which open new financing potential for small competitors and thus liberalize the market. The block chain technology is thus of interest for quite diverse applications within the field of GSS.

6 Conclusion and outlook

WP4 contributes the user perspective to the Centre of Excellence for Global Systems Science, by defining requirements GSS has in using HPC – the third iteration of the report on pilot requirements is due only two months after this present deliverable – and by beginning to develop something at the intersection between two previously largely unrelated fields. This report has collected the status of the three pilot studies in their endeavour towards HPC-GSS applications that point out how global challenges can be addressed and potentially turned into opportunities.

As shown throughout the chapters 2 to 4 of this document, the development of the pilots' synthetic information systems has progressed to a point from which model implementations can now be thoroughly explored and analysed. How GSS can benefit from HPC and HPDA in this task shall therefore be one of the WP4 concerns over the third project year. Two future directions are outlined in Sections 6.1 and 6.2 below.

Another focus of pilot work in year three shall be upscaling the models, for example by including further geographical regions or by including mechanisms tested in a regional scale model into a global model with initially simpler dynamics. A tight loop between model extensions and thorough exploration of the extended models, in order to find out which extensions work and may lead to new insights, shall be needed for this task.

In terms of results obtained, the current stage models indicate first policy relevant points, such as pointing out high-risk areas for the smoking epidemics that efforts should be concentrated on, or displaying spatial patterns for finding pilot regions if one wants to foster electric mobility. Pilot work in the coming project year, as well as the identification of potential future applications, shall actively pursue the aim of producing GSS-relevant example results for which the use of HPC or HPDA have played an essential role.

6.1 Parameter space exploration

The question of optimal parameter space exploration, allowing optimizing knowledge acquired over a number of simulations, has raised specific research questions (Santner 2003).

A first approach relies on optimal predefined exploration strategies (design of experiments) targeting deterministic or stochastic models, such as systematic exploration, latin hypercube sampling, and others. However, being defined a priori and not integrating information provided by simulation on the fly, they can often be improved upon.

Therefore, adaptive space exploration strategies have been developed (Picheny et al. 2010) for deterministic or stochastic models (Paninski 2009), in many different fields and different contexts, which appear relevant (sometimes when transposed) to the question of parameter space exploration. Just a few examples include spatially refining the calculation of fluid dynamics (Loseille and Löhne 2010) following the evolution of turbulent zones or zones of interest, picking points for optimal 3D visualization, exploration of high dimensional

parameter spaces in biology (Paninski 2009, Zamora-Sillero et al. 2011), adaptive parameter space exploration (Blondet et al. 2010) or adaptive mesh refinement.

There exist general implementations of parameter space exploration tools, such as the pse package of R (however mainly for latin hypercubes). There exist similar tools compliant with HPC, such as Dakota (https://dakota.sandia.gov/) or the Sparse Grid tools.

Some of these approaches request exchanging information in a decentralized way (e.g., for gradient calcuation) and would benefit from HPC (as compared with purely parallel, non-communicating Cloud simulations), while answering a real GSS need (see Section 1.1).

Requirements of the pilots in the field of parameter space exploration will be specified in the upcoming deliverable on pilot requirements, D4.3.

6.2 Big simulation data analysis

There is a broad understanding between the pilots that extensive model analysis is required to support the development of GSS models. Due to the large amount of possible output data and a similar structure of the output data, we plan to apply big-data methods and tools (like SPARK). This shall allow to analyse distributed output data efficiently and to facilitate well-established data analytics tools for post-processing, aggregation and visualization of model output data. An initial step will be the transfer of existing analysis steps done, for example, in python, to SPARK queries to use distributed systems. This will allow to rather easily extend the current analysis for one model run towards ensembles of simulations. Thus, the current results can be augmented with uncertainty measures like confidence intervals, standard deviations or probabilities.

Requirements of the pilots in the field of big data analytics methods for simulation output analysis will also be specified in the upcoming deliverable on pilot requirements, D4.3.

7 References

7.1 CoeGSS deliverables

D3.3: CoeGSS Deliverable D3.3; SECOND SPECIFICATION OF NEW METHODS, TOOLS AND MECHANISMS PROPOSED FOR THE SUPPORT OF THE APPLICATION USER AND PROGRAMMER; delivered to the European Commission; Patrik Jansson (editor), Marcin Lawenda, Burak Karaboga, Piotr Dzierżak, Oskar Allerbo, Enrico Ubaldi, Wolfgang Schotte, Cezar Ionescu, Michał Pałka, Eva Richter, Ralf Schneider, Michael Gienger

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

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

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

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

7.2 Literature

Barrat, A., Barthelemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks* (1st ed.). Retrieved from http://gen.lib.rus.ec/book/index.php?md5=EEFBCE3072D82D0F8A624188B53473BB

Beheshti, R., & 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* (pp. 557–564). Richland, SC: International Foundation for Autonomous Agents; Multiagent Systems. Retrieved from http://dl.acm.org/citation.cfm?id=2615731.2615822

Blondet, G., Boudaoud, N., Duigou, J. L., & Eynard, B. (2010). *Simulation data management for adaptive design of experiments: A literature review*. INRIA. Retrieved from https://www.rocq.inria.fr/gamma/gamma/Membres/CIPD/Adrien.Loseille/publication/loseille-remesh.pdf

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Rev. Mod. Phys.*, *81*(2), 591–646. https://doi.org/10.1103/RevModPhys.81.591

Castillo-Garsow, C., Jordán-Salivia, G., & Rodriguez-Herrera, A. (1997). Mathematical models for the dynamics of tobacco use, recovery, and relapse. *Public Health*, *84*(4), 543–547.

Digital Day. (2017). Panel I - HPC: A new drive in the European Digital Economy,

Ferguson, J., Bauld, L., Chesterman, J., & Judge, K. (2005). The English smoking treatment services: One-year outcomes. *Addiction*, *100*, 59–69. https://doi.org/10.1111/j.1360-0443.2005.01028.x

IEA. (2016). Global EV Outlook 2016. https://www.iea.org/publications/freepublications/publication/Global_EV_Outlook_2016.pdf.

Ionescu, C., & Jansson, P. (2013). Dependently-typed programming in scientific computing: Examples from economic modelling.

Lang, J. C., Abrams, D. M., & Sterck, H. D. (2015). The influence of societal individualism on a century of tobacco use: Modelling the prevalence of smoking. *BMC Public Health*, *15*(1), 1–13. https://doi.org/10.1186/s12889-015-2576-6

Levy, D. T., Cummings, K. M., & Hyland, A. (2000–8AD). A simulation of the effects of youth initiation policies on overall cigarette use. *American Journal of Public Health*, *90*(8), 1311–1314.

Loseille, A., & Löhne, R. (2010). *Anisotropic mesh generation application to high-fidelity simulation in cfd*. INRIA. Retrieved from https://www.rocq.inria.fr/gamma/gamma/Membres/CIPD/Adrien.Loseille/publication/loseille-remesh.pdf

Marathe, M. (2017). At-scale realistic synthetic social habitats: A unifying data structure for global system science. Presentation at the International Conference on Synthetic Populations. Retrieved from https://icspconference.files.wordpress.com/2016/12/06_marathe.pdf

Nguyen-Luong, D., Boucq, E., & scientific advisor: F. Papon. (2011). *Evaluation de l'impact du T3 sur le prix de l'immobilier résidentiel*. PREDIT, IAU, IFSTTAR, Ministère de l'écologie, de l'énergie, du développement durable et de l'aménagement du territoire. Retrieved from https://www.iau-idf.fr/fileadmin/NewEtudes/Etude_814/Evaluation_de_l_impact_du_T3_sur_les_prix_de_l_immobilier_residentiel.pdf

Office for National Statistics. (2016). People, population and community. https://www.ons.gov.uk/peoplepopulationandcommunity.

Paninski, L. (2009). *Efficient adaptive experimental design*. Columbia University. Retrieved from http://stat.columbia.edu/~yiannis/class/HOS/LP2.pdf

Picheny, V., Ginsbourger, D., Roustant, O., R.T.Haftka, & Kim, N.-H. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, *132*(91).

Poulhès, M. (2015). *Fenêtre sur Cour ou Chambre avec Vue? Les prix hédoniques de l'immobilier parisien*. INSEE. Retrieved from https://www.insee.fr/fr/statistiques/fichier/1304140/G2015-19.pdf

Public Health England. (2016). Local Tobacco Control Profiles for England, Smoking Quitters. http://www.tobaccoprofiles.info/search/quit.

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

Santner, T. (2003). *The design and analysis of computer experiments* (Springer, Berlin).

SEDAC. (2016). Gridded population of the world, version 4 (gpwv4): Population count adjusted to match 2015 revision of un wpp country totals. *Center for International Earth Science Information Network - CIESIN - Columbia University, Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)*. http://dx.doi.org/10.7927/H4SF2T42.

Sharomi, O., & Gumel, A. (2008). Curtailing smoking dynamics: A mathematical modeling approach. *Applied Mathematics and Computation*, *195*(2), 475–499. https://doi.org/http://dx.doi.org/10.1016/j.amc.2007.05.012

Viola, R., & Smits, R.-J. (2017). Why do supercomputers matter for your everyday life? - Rome edition. DSM blog post, 16 March 2017. Retrieved from https://ec.europa.eu/digital-single-market/en/blog/why-do-supercomputers-matter-your-everyday-life-rome-edition

Zamora-Sillero, E., Hafne, M., Ibig, A., Stelling, J., & Wagner, A. (2011). Efficient characterization of high-dimensional parameter spaces for systems biology. *BMC Systems Biology*, *5*(142). Retrieved from http://www.biomedcentral.com/1752-0509/5/142

Appendix A

Health Habits pilot model report

Appendix B

Slightly adapted version of a Green Growth pilot conference publication

Third Report on Model specification

Agent based model definition, numerical simulations and preliminary model calibration by the Health Habits Pilot

Enrico Ubaldi, Luca Rossi, Michele Tizzoni, Alessandro Vespignani

Introduction

The document updates the definition and implementation of our compartmental, agent based model describing the smoking spreading in a population, details the implementation of the model in `Pandora`, presents the post-process procedure and a preliminary calibration of the model.

With respect to the previous version we further developed the model introducing a refined quit-relapse mechanism in Section 2, we extend the model implementation accordingly in Section 3, and we detail the pre/post-processing workflow highlighting the tasks that can be improved by using of HPC tools and resources. Finally, we present a preliminary model calibration.

We recall that the modeling framework has been chosen for two main reasons: *i*) the process that an individuals undergo when they start smoking can be seen as an epidemic process [1, 2, 3, 4], or rather a complex contagion process, that can modelled leveraging on the well established framework of epidemic models [5, 6], *ii*) it features a limited number of parameters that can be set using data that are available for the most of the countries (namely, smoking prevalence, statistics on the number of quitting attempts and relapse rate) and *iii*) this particular framework can be implement and deployed using the *Pandora* framework, which is HPC ready and will feature the agent graph model in the next software development steps.

In the following, we first introduce epidemics models and the transitions that an agent can undergo during the system evolution. We then define the equations of the model dynamics and, finally, we show how to derive the simulations parameters from real world data.

The main topics and results presented in this work are the following:

- in Section 1 we start from an epidemic-like, compartment model as found in [7] and we extend it to comply with a discrete time dynamics and to the ABM modelling framework;
- in Section 2 we define the ABM model for smoking adoption, cessation, relapse and we give a pseudo-code implementation of it;
- Section 2.1 shows how to measure the model parameters from real world data and how to convert continuous time rates to the probabilities of the discrete time ABM.

- Section 3 then contains the preliminary results of numerical simulations and a quick outline of the pre-processing and post-processing steps done to generate the simulations' input and parameters and later analyze their output.

1 Compartmental model

In a compartmental model, a single agent can be found in one of n different compartments and can move from one to the other with a given probability. For example, the susceptible-infected SI model reads:

$$S + I \xrightarrow{\beta} 2I, \tag{1}$$

where β is the rate of infection by which a susceptible individual gets infected by an infected one.

The model can be expressed as a set of differential equation describing the evolution of the system in time:

$$\frac{dS(t)}{dt} = -\beta S(t)\frac{I(t)}{N} \tag{2}$$

$$\frac{dI(t)}{dt} = +\beta S(t)\frac{I(t)}{N}, \tag{3}$$

where the I/N term accounts for the density of infected individuals, i.e., for the probability for a susceptible individual to encounter an I individual.

Of course, one can add other compartments to model epidemic processes involving more than just two states and define different transition rates to describe the system under investigation. In this version of the model we consider two kinds of transitions: spontaneous ones, in which the rates are just constants, and induced ones, where the rate of transition depends on other compartments (such as in the just outlined example).

1.1 Model definition

We define three compartments in our model [7, 4]:

- S (never smokers: agents that never smoked before and can start to smoke);
- I (current smokers: agents that are regular smokers);
- R (quitters: former smokers that temporarily quit the smoking habit).

Model without mortality

For simplicity, we initially assume the number of agents to be fixed in time, i.e., no individual enters or leaves the system. The dynamical equations of the model then read:

$$\begin{aligned} \frac{dS}{dt} &= -\beta S\frac{I}{N}, \\ \frac{dI}{dt} &= \beta S\frac{I}{N} - \gamma I + \delta R, \\ \frac{dR}{dt} &= \gamma I - \delta R, \end{aligned} \tag{4}$$

The transitions amongst the different compartments found in the system are:

- $S+I \xrightarrow{\beta} 2I$: infection process, in which a never smoker gets in contact with a current smoker and start smoking at rate β, which generally depends on many factors, e.g., the propensity of a given person to interact with others, the fact that an individual may become just occasional smoker rather than a regular one. We leave these details for next versions of the model. Just note that all of these mechanisms are accounted by the parameter β that sets the infection rate (that is, how much the epidemics is likely to spread when a never-smoker gets in contact with a regular smoker).

- $I \xrightarrow{\gamma} R$: quitting process, which occurs when a current smoker decides to quit smoking at rate γ.

- $R \xrightarrow{\delta} I$: relapse mechanism that goes in the opposite direction of the quitting process and occurs whenever a former smokers falls back to the smoking habit at rate δ.

Model with mortality

We can add mortality to the model by defining a mortality rate μ by which new agents are added to the system in the S compartment (natality) and later removed. The mortality rate μ is the same for all the agents in the system, regardless of their current compartment. From this perspective, an agent enters the system in the S (never-smoker) compartment and then leaves the system at rate μ, regardless of the compartment she is at that time. The three Eq. (4) then become:

$$\begin{aligned} \frac{dS}{dt} &= -\beta S\frac{I}{N} + \mu(N - S), \\ \frac{dI}{dt} &= \beta S\frac{I}{N} - \gamma D + \delta R - \mu I, \\ \frac{dR}{dt} &= \gamma D - \delta R - \mu R. \end{aligned} \tag{5}$$

Note that the mortality rate μ sets the number of individuals entering (and leaving as we are working in the constant population approximation) the system in a infinitesimal time interval dt, as $\mu N dt = |-\mu(S(t) + I(t) + R(t))|$ is the number of born (dead) people in the $(t, t + dt)$ time interval. For a finite time interval of Δt length, we have that $\mu N \Delta t = N_{\text{born}} = N_{\text{dead}}$ equals the number of born (dead) people as measured from census data, so that we can derive μ from real world data as we will show in Section 2.1.

1.2 Asymptotic behavior of the system

The observable of interest in epidemics systems is the basic reproduction number R_0 which, loosely speaking, estimates the number of second cases (smokers) generated by a single infected individual during his permanence in the I compartment, assuming that the system has a completely susceptible population [5].

The reproduction number R_0 corresponds to the largest eigenvalue of the Jacobian matrix J associated with the system in the asymptotic solution of no endemic state, i.e. when $I(t) = R(t) = 0$, that reads:

$$J = \begin{bmatrix} \frac{\beta}{\gamma+\mu} & \frac{\delta}{\gamma+\mu} \\ \frac{\gamma}{\gamma+\mu} & 0 \end{bmatrix}, \tag{6}$$

whose largest eigenvalue λ_L reads:

$$\lambda_L = R_0 = \frac{\beta}{2(\gamma+\mu)} \left[1 + \sqrt{1 + \frac{4\gamma\delta(\gamma+\mu)}{\beta^2(\delta+\mu)}} \right]. \tag{7}$$

The outbreak of the epidemic is set when $R_0 > 1$.

We can also obtain the asymptotic behavior of the system (i.e., the long time limit of the dynamics) in which we can solve the system equation by assuming the stationary state $dC(t)/dt = 0$ for all compartments $C = S$, I, R, finding:

$$\frac{S}{N} = \frac{\mu}{\beta\frac{I}{N} + \mu}, \tag{8}$$

$$\frac{I}{N} = \frac{\delta+\mu}{\delta+\mu+\gamma} - \frac{\mu}{\beta}, \tag{9}$$

$$\frac{R}{N} = \frac{\gamma}{\delta+\mu}\frac{I}{N}, \tag{10}$$

$$\tag{11}$$

2 Agent based model definition

We are now ready to define the discrete-time ABM model leveraging on the just outlined continuous-time model. For simplicity, let us assume that we are given a function *rate2prob* that converts the continuous time rates μ, β, γ, to their discrete time probabilities counterparts $m = \text{rate2prob}(\mu)$, $b = \text{rate2prob}(\beta I(t))$, $g = \text{rate2prob}(\gamma)$. We here focus on the model definition, while we will later show in Section 2.1 how to first measure the continuous rates and translate them to the corresponding discrete probabilities. Note that we did not translate the δ rate as we are going to implement the relapse mechanism in a truly agent-based way, i.e. each agent will store personal information on her quitting status. This mechanism may be translated in the continuous time model but one, or in a discrete time model with homogeneously mixed population, but at the price of introducing a single compartment for each possible quitting status of the agents. For example, in our formulation each currently smoking agent that tries to quit draws from a certain distribution $\mathcal{P}(m)$ the number of months m_q during which he will be a successful quitter. The agent then lowers this number for each step of the simulation until $m_q = 0$ meaning that the agent is relapsing. While the implementation of such a model is straightforward in the ABM framework, it would require the definition and creation of one quitting compartment for each possible

number of quitting months and the implementation of compartment-specific rates of relapse.

The parameters of the system are the natality (or mortality) probability m, the influence probability b, and spontaneous quitting probability g. In addition, we have the parameters $\mathbf{p}$ describing the distribution of the quitting times accounting. These parameters set the probability per single step for each agent to move from one compartment to the other. In addition, we have to define:

- N, the number of agents in the system, that will be a fraction $\phi \in (0, 1]$ of the total population P_{real} of the countries under investigation, i.e. $\phi = N/P_{\text{real}}$;
- C, the number of cells in the simulations, where each cell c corresponds to a particular area of the country under investigation and belongs to a specific geographical (or administrative) region r;
- $\mathcal{T}$, the time step of the ABM, i.e. each evolution step represent the evolution of the system for $\mathcal{T}$ units of time (could be days, months or years). As we will show in Section 2.1 this parameter sets the rule to translate the continuous time rates to their discrete counterparts;
- the number n_T of steps to reproduce, so that $T = n_T\mathcal{T}$ is the total period of time simulated;
- the $s(t = 0) = S(0)/N$, $i(0) = I(0)/N$, and $r(0) = R(0)/N$ initial conditions of the system, i.e. the fraction of the population that is found in every compartment at the beginning of the simulation; depending on the available data, these quantities may be defined on a single geographical region/local authority level, so that the generic cell c with real population $P_{c,\text{real}}$ belonging to the administrative region r will have $s_c(t = 0) = P_{c,\text{real}}\phi S_{r,\text{real}}(0)/P_{r,\text{real}}$, where $S_{r,\text{real}}(0)$ and $P_{r,\text{real}}$ are the measured number of never-smokers and the population of region r at the initial time of simulation. Note that we have implicitly assumed that a region r may (and in general will) contain more than one cell c.

With respect to the previous model implementation we let the γ rate (and thus also the probability g) to vary from one English region to the other, as measured by the National Health Association (NSH) on the quit/relapse ratios in England [8]. All the other rates and probabilities do not depend on the region r (and thus neither on the cell index c), e.g. $\beta_r = \beta$ for any region r. In other words, we set the parameters of the system at their finest available resolution [9, 10, 11, 12].

We also define the population of the system to be divided into C cells, defined accordingly to a selected partition of the geographical space, in this case we selected the SEDAC cell division of the entire globe in $1x1$ km cells [13]. In this model definition we assume that agents may interact only with other agents present in the same cell, leaving as a future development the inter-cell interactions of the agents.

Given this assumptions, and the parameters definition given above, we can define the model as follows:

(i) for each time step, cycle over the cells in the system;

(ii) for each cell compute the smoking prevalence $i_c(t)$ and, accordingly, the b probability of the smoking transmission from a current smoker to a never smoker;

(iii) cycle over the agents of that cell c and update their status accordingly to the probabilities:

$$\begin{aligned} S &\xrightarrow{b} I \\ I &\xrightarrow{g} R \end{aligned} \tag{12}$$

(iv) if the agents tries to quit at this step, she draws the number of months m that she will spend in the ex-smokers compartment from the time distribution $\mathcal{P}(m)$; if, instead, the agent is a former smoker, she will decrease the m counter by $\mathcal{T}$ months, being $\mathcal{T}$ the number of months per discrete simulation step; if, after this subtraction, $m \leq 0$ the agent relapse to the current smokers compartment;

(v) replace with probability m every agent in the system (mortality) with a never-smoker in the S compartment (natality), so as to keep the population constant in time.

The corresponding pseudo code of the model, in a `python`-like syntax, is as follows:

```
# for over the time steps of the model.
for t in range(tau):

    # for over the cells in the system.
    for cell in range(C):

        # The number of infected in the cell setting the probability
        # to get infected together with the rate $\beta$. Here we assume
        # that rate2prob is the function translating the rate to the
        # probability and that beta is a function returning the beta rate
        # for thespecific region that includes the cell.
        I_c_t = cell.n_infected
        rate_StoI= beta(cell) * I_c_t/cell.total_population
        prob_StoI= rate2prob(rate_StoI)

        # for over the agents in the cell.
        for agent in cell.iterate_agents():

            if agent.status == 'S':
                if random() < prob_StoI:
                    agent.set_status('I')
```

```
        elif agent.status == 'I':
            if random() < rate2prob( gamma(cell) ):
                agent.set_status('R')
                agent.quitting_time = quitDistribution.draw()

        elif agent.status == 'R':
            agent.quitting_time -= monthsPerStep
            if agentn.quitting_time <= 0:
                agent.set_status('I')

    # Here I can record the status of the system after the evolution
    # step and then proceed to the natality/mortaility step

    for agent in cell.iterate_agents():
        # For every agent let us try the mortality and let new agents enter
        # thesystem...
        if random() < rate2prob( mu(cell) ):
            #for each agent dying out I get a new agent in the 'S' state
            #soI simply update the status...
            agent.set_status('S')
```

In the code we indicate as `quitDistribution` the variance generator of the quitting distribution time which is sampled through the `draw()` method.

2.1 Parameter estimation

Here we show how to estimate the parameters of the continuous-time model from real-world data.

We report here the data sources we uses in the parameter estimation:

- census data regarding population count, mortality and natality figures and per-age number of individuals in the United Kingdom [14];
- yearly smoking prevalence in the overall population and per age-bracket in the $2012-2015$ period [15]. Specifically, we use the $I(y)$ incidence of smoking in the whole population per year y and the $I_a(y)$ prevalence per age bracket a and year y;
- the statistics about quit attempts and relapse rate, i.e., the fraction f_{quit} of the smoking population that tries to quit in a given year and the fraction f_{relapse} of unsuccessful tries (relapses) amongst them. The former is measured as the *number of people setting a quit date per 100,000 smokers* every year [15]. The relapse rate is instead measured four weeks after the quitting date, while supplementary studies measure it six months and one year after the quit attempt [16].

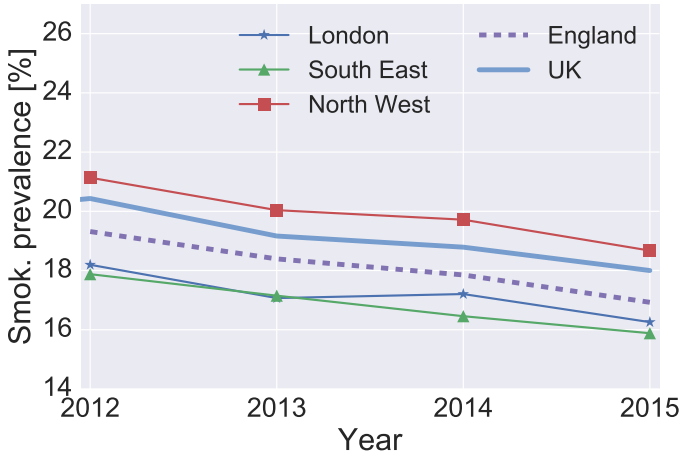


Fig. 1: The temporal evolution of smoking prevalence for the last 5 years in the London, South East, and North West regions in England and at a national level for England and UK (see legend for markers, data from [11, 12]).

Mortality rate

As mentioned after Eq. (5), the term $\mu N \Delta t$ sets the number of agents born in a finite time interval Δt. Thus, to measure μ we have to get the number of dead (born) people in a reference time interval Δt (usually one year) in a given country. As we are working in the static population approximation, we set the mortality rate to be $\mu = \phi M = \phi \langle D_{\text{real}}, B_{\text{real}} \rangle$, where the total mortality count $M = \langle \dots \rangle$ is the arithmetic average of the measured number of deaths D_{real} and births B_{real}. For instance, in England and Wales together there has been $D_{\text{real}} = 529655$ and $B_{\text{real}} = 697852$ in 2015 over a population of $P_{\text{real}} \simeq 5.7 \cdot 10^7$, so that we can set $M = 6.1 \cdot 10^5$ [14].

Cessation rate

The available datasets normally provide the number (or the fraction of the population) of never, current and former smokers measured at Δt intervals (usually every year) for a specific country. Usually, national authorities also provide data on the number of current smokers that try to quit in a given year (thus setting a fraction f_{quit} of smokers that tried to quit) and the failure rate of these attempts at different time intervals afterwards (number of former smokers f_{relapse} that relapse to the I compartment after having tried to quit) [8, 16].

One has to convert these numbers into the rates of the continuous time model and later calculate the corresponding probabilities for the ABM discrete time modelling framework. In the process one has also to take into account the interplay of different terms in determining the value of such parameters, e.g. the fraction of the population in the I compartment concurs in the determination of the transition rate from $S \to I$.

To this extent, one possible approach is to focus on the terms subtracting agents from one compartment C to another C' and approximate the value of the other compartments $C'' \neq C$ as constant during the evolution step whose length is Δt (usually one year).

We now present the procedure for the $I \to R$ transition. The term subtracting agents from $I \to R$ is:

$$\frac{dI(t)}{dt} = -\gamma I(t), \tag{13}$$

so that we can write the generic solution of Eq. (13) at time t for $I'(t) = I'_0 e^{-\gamma t}$, where $I'_0 = I'(t = 0)$ is a constant and $I'(t)$ is the dynamical solution for the current smoker compartment accounting only for the quitting mechanism (and it has not to be confused with the overall $I(t)$). Then, we can write the relative variation of $I'(t)$ between time 0 and t, so that

$$\frac{I'(t)}{I'(0)} = e^{-\gamma t} \;\to\; \left(1 - \frac{I'(t)}{I'(0)}\right) = 1 - e^{-\gamma t} \tag{14}$$

is the fraction of current smokers that tried to quit smoking during the time interval $[0, t]$. By getting this fraction from empirical data referring to the $[0, \Delta t]$ period we can set the value of γ in the $[0, \Delta t]$ interval to be:

$$\gamma_{[0,\Delta t]} = -\frac{1}{\Delta t} \ln\left(\frac{I'(\Delta t)}{I'(0)}\right) = -\frac{1}{\Delta t} \ln\left(1 - f_{\text{quit}}\right), \tag{15}$$

where f_{quit} is the fraction of smokers trying to quit in a given time interval.
For example, in England about the $f_{\text{quit}} = 5.1\%$ of smokers actually set a quit date every year. Amongst them, about $1 - f_{\text{relapse}} = 50\%$ is still a successful quitter after 4 weeks [15]. This figure drops to 32% after six months and to 14.6% after one year [16]. Using $f_{\text{quit}} = 5\%$ we find:

$$\gamma \simeq 7.3 \cdot 10^{-3} \text{ years}^{-1}. \tag{16}$$

In the current model implementation we let γ to vary from one region to the other as we implement the quit and relapse data of Table (1), so that $\gamma_R = -1/\Delta t \ln\left(1 - f_{\text{quit,R}}\right)$, where $f_{\text{quit,R}}$ is the fraction of smokers that tries to quit in a given region R.

Relapse mechanism

The relapse mechanism is implemented by letting the agents who is quitting to draw a relapse time m measured in months from a given time distribution $\mathcal{P}(m)$. The latter is derived from real world data regarding the quitting and relapse behavior of people in England [8, 16].

Starting from the national health data we have, besides the already outlined data on smokers quitting, the fraction r of these quitters that relapse after one month for each one of the nine regions of England. This fraction is measured both as individuals self-reporting the quitting status after one month as well for the actual number of individuals who have been found negative to the carbon monoxide test (thus being CO-validated). Data are reported in Table (1)

Indicator Region	Quitters	Successful	Successful CO-validated
East Midlands region	7232	4056	2421
East of England region	7557	4112	2918
London region	7401	3844	2694
North East region	9532	4393	3519
North West region	8572	3714	1935
South East region	6030	3299	2432
South West region	6647	3444	2676
West Midlands region	8379	4403	3525
Yorkshire and the Humber region	5558	2994	2260

Tab. 1: The number of smokers per 100,000 smokers setting a quit date (Quitters). The number of these individuals still being successful quitters after one month (Successful, self reported) and the subset of those who where also validated via carbon-monoxide test (CO-validated). Data from [8].

It is reasonable to expect that the reported quitters' success rate gives an over estimation of the real figures, as one this numbers are obviously bound to drop as time passes from the quitting date. That is why we incorporated in the relapse analysis the data of a study following the individuals trying to quit smoking for one year [16]. We report the findings of this study in Fig. 2. Specifically, in we show that the fraction $q(m)$ of successful quitters m months after their quitting trial lowers at $\sim 20\%$ after one year.

We model this relapse mechanism by means of the hazard function $h(t)$, i.e., the conditional probability to fail (relapse) at time t given no previous failure, and survival (or reliability) function $s(t)$, the probability of no failure before time t. The hazard function $h(t)$ is then defined as:

$$h(t) = \frac{r(t) - r(t + \Delta t)}{\Delta t\, r(t)} = \frac{f(t)}{r(t)}, \tag{17}$$

where $f(t)$ is the time-to-first-failure distribution. Eq. (17) can be thought as a conditional probability for an agent to fail during a discrete time interval $[t, t + \Delta t]$ and it is measured as the number of failures observed in such time interval divided by the number $r(t)$ of individuals that never failed up to time t. The latter can be expressed in terms of the complementary cumulative distribution (CCDF) of $f(t)$, i.e., $r(t) = CCDF(f(t)) = 1 - CDF(f(t)) = 1 - F(t)$, where $F(t)$ is the cumulative of $f(t)$.

We find that the best fit to the empirical survival probability function $r(t) = q(m)$ (where $q(m)$ is the fraction of quitters still successful and without failures up to month m) is described by a $\chi^2(t)$ time to first failure distribution with $k = 0.55$ degrees of freedom, as shown in Fig. 2. We also tested that this assumption fitting the empirical data with a q-exponential distribution that returned $q = 1.64$, i.e., one χ^2 distribution is sufficient to

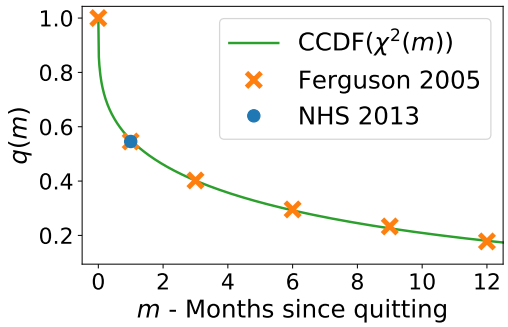


Fig. 2: The fraction $q(m)$ of successful quitters after m months since their quitting date as reported for $m = 1$ by the 2013 National Health Survey [8] (blue dot), $m \in [1, 12]$ by the study in [16] (orange crosses), and the fitted survival function $q(m) = r(m)$ when using a $\chi^2(m)$ time to first failure distribution with $k = 0.67$ degrees of freedom.

describe the variance of the times to failure [17].

Hence, we will implement the relapse mechanism by letting a quitting agent draw a relapse time m from the χ^2 distribution whose parameters are set by the previous analysis. Then, after m months are simulated, the agent will move back to the current smoker compartment. We will also assume that if an agent draws a relapse period $m > 12$ months we will assume her as a permanent quitter, so that she will never relapse. This assumption is due to the limited time span of the quitters follow-up and may be refined when longer follow-up studies will be available.

Initiation rate

The last parameter we have to estimate is the infection rate β. This is the most complicated computation as we have to account for the prevalence of smoking in the population (i.e. the $I(t)/N$ fraction). In order to evaluate β we have to measure from real data the $S \to I$ transition, and thus have the estimation of the number of *new* smokers that enter the I compartment per period of time. As these data are usually not available in national statistics, we have to apply some workarounds, by choosing one (or more) of the following strategies and compare their performances:

- **direct evaluation from data**: as at this stage we have defined the γ, δ, and μ parameters we can directly measure β by solving the second of Eq. (5) with respect to β:

$$\Delta I = [I(\Delta t) - I(0)]\,\Delta t = \left(\beta S(0)\frac{I(0)}{N} - \gamma D(0) + \delta R(0) + \mu I(0)\right)\Delta t. \tag{18}$$

  However, the value of $R(t)$ (i.e. the amount of former smokers in a population) is not easily found in most of the countries. Moreover, we are setting β to a combination of the already measured parameters and collected data. We are then superimposing

that the whole system behaves precisely as the theoretical model, and this can lead to wrong estimates of the β parameter;

- **approximate the number of new smokers from data**: as it is recognized that the all the smoking epidemic dynamics takes place when the agent is in the $12 \lesssim y \lesssim 24$ age bracket [15, 7, 4], we can apply the approximation in which we do not observe smoking cessation during this period. Thus, in the countries where data on smoking prevalence $I_a(t)$ divided per age brackets a in the $11 \lesssim y \lesssim 25$ are available, we can infer the number of new smoker per year by looking at the variation of smoking incidence between age brackets (weighted by the P_a population falling in the age bracket a) in the time interval $11 \leq y \leq 25$, for instance

$$N_{\text{new smok.}}(t) = \sum_{a=11}^{25-1} P_{a+1} I_{a+1}(t) - P_a I_a(t) \tag{19}$$

  We now use the fact that the $S \to I$ transition yields $dI'(t) = -dS'(t) = \beta I'(t)/N S'(t) dt = N_{\text{new smok.}}(t)$ is the number of new smokers due to initiation, giving us the β parameter;

- **asymptotic evaluation of** β: in this approach we define β as the value that reproduces (starting from heterogeneous initial conditions) the asymptotic prevalence $I(t \to \infty)$ in better agreement with the empirical measure $I_{\text{real}}(t)$. The requirement on the diverse initial conditions enforces the prerequisite of stability of the endemic solution of the system in the asymptotic limit. Indeed, if two representations of the system starting from diverse initial conditions are returning different asymptotic states of the system it means that we are exploring an unstable region of the parameter space. We have then to limit ourself in the region of the parameters corresponding to an asymptotic stable state, as we are approximating our system to be at the equilibrium.

Translate continuous rates to discrete time ABM models with steps of arbitrary length

In order to simulate an ABM, the last step is to translate the continuous time rates into discrete transition probabilities intended to be used in an ABM featuring time steps of arbitrary length $\mathcal{T} = \Delta t/k$. We recall that Δt is the time resolution of empirical data (usually $\Delta t = 1$ year) and $k \in \mathbb{N}$ sets the ratio $\Delta t/\mathcal{T}$. Note that, in general, $\mathcal{T}$ may be either a fraction or a multiple of the of the reference time interval Δt, but we will here focus on the $k \geq 1$ case.

In the previous section we evaluated the rates γ, μ, β, and δ at which agents leave one compartment in favor of another. Let us focus, for instance, on the $I \to R$ transition which happens at rate γ. In the approximation scheme where we track the flow of agents from I to R only we find that $I(t) = I(0)e^{-\gamma t}$. Now to translate γ to the probability g to leave in a discrete time $\mathcal{T}$ the compartment $I \to R$. Given that there are k time step of length $\mathcal{T}$

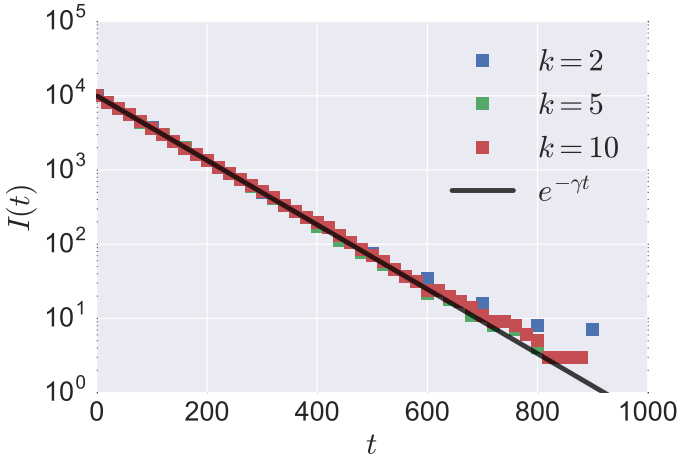


Fig. 3: The discrete version of the model compared to the continuous time solution as found in Eq. (20) with $\gamma = 0.01$, $I(0) = 10^5$ and $\Delta t = 1$ simulated with $k = 2$, 5, 10.

in the Δt interval we can write the population $I(t = \Delta t)$ as

$$I(t = \Delta t) = \underbrace{I(0)e^{-\gamma \Delta t}}_{\text{continuous time}} = \underbrace{I(0)(1-g)^k}_{\text{discrete time}} \rightarrow g = 1 - e^{-\gamma \Delta t/k}. \tag{20}$$

In Fig. 2.1 we show the result of a single compartment model in which we only have an exit rate $\gamma = 0.001$ that we compare to the corresponding discrete time model with $k \in [2, 5, 10]$. As one can see the continuous time analytical solution $I(t) \propto e^{-\gamma t}$ is indistinguishable from the discrete solutions at different time steps resolution k.

Regarding our smoking example, if we want to simulate our system at steps of 3 months (i.e. with $k = 4$), we would have to convert $\gamma = 7.3 \cdot 10^{-}3\text{years}^{-1}$ into g as:

$$g = 1 - e^{-\gamma/4} = 1.83 \cdot 10^{-3}. \tag{21}$$

The multi-transitions case

Particular attention has to be devoted to the case in which we have two (or more) transitions leading from a compartment to others. For instance, the relapse and mortality mechanisms set two possible ways out from the R compartment lead by the δ and μ transition rates, respectively.

In such a case we can apply the same reasoning outlined in the previous section but taking care of accounting for both the two transitions. Specifically, let us suppose that we have an SI model with two distinct diseases for which we create two separate compartments, i.e. I_1 and I_2. Now suppose that the transitions are spontaneous (i.e. they do not depend on the I/N disease prevalence) and that their rates are β for the $S \rightarrow I_1$ and γ for the $S \rightarrow I_2$ channel.

We can write the dynamical equations of the system as

$$\begin{aligned} \frac{dS(t)}{dt} &= -(\beta+\gamma)S \\ \frac{dI_1(t)}{dt} &= +\beta S \\ \frac{dI_2(t)}{dt} &= +\gamma S. \end{aligned} \tag{22}$$

Now given the general initial conditions $S(t=0) = S_i$, $I_1(t=0) = I_{1i}$, $I_2(t=0) = I_{2i}$ and the conservation condition $N = S(t) + I_1(t) + I_2(t)$ for any t we eventually find:

$$\begin{aligned} S(t) &= S_i e^{-(\beta+\gamma)t} \\ I_1(t) &= I_{1i} + S_i \frac{\beta}{\beta+\gamma}\left(1 - e^{-(\beta+\gamma)t}\right) \\ I_2(t) &= I_{2i} + S_i \frac{\gamma}{\beta+\gamma}\left(1 - e^{-(\beta+\gamma)t}\right). \end{aligned} \tag{23}$$

We now have to translate the β and γ rates to the corresponding probabilities b and g for a discrete time evolution that considers k evolution steps in a unit time. To this end, note that the number of people ΔI_1 moving from $S \to I_1$ in a unit time (and thus in k discrete time steps) can be written as:

$$\Delta I_1 = I_1(t+1) - I_1(t) = S(t)\left[1 - (1-b)^k\right] = S_i \frac{\beta}{\beta+\gamma}\left(e^{-(\beta+\gamma)(t+1)} - e^{-(\beta+\gamma)t}\right), \tag{24}$$

Where we have used the fact that $(1-b)^k$ is the probability to never select the $S \to I_1$ transition in k steps, so that $1-(1-b)^k$ is the probability for an individual to end up in I_1 within k steps. We can repeat the same procedure for the $S \to I_2$ transition by substituting $\beta \to \gamma$ and b with g, then resolving Eq. (24) with respect to b and g finding that

$$\begin{aligned} b &= 1 - \left[1 - \frac{\beta}{\beta+\gamma}\left(1 - e^{-(\beta+\gamma)}\right)\right]^{1/k} \\ g &= 1 - \left[1 - \frac{\gamma}{\beta+\gamma}\left(1 - e^{-(\beta+\gamma)}\right)\right]^{1/k}. \end{aligned} \tag{25}$$

To test the prediction of Eq. (25) we run a set of 20 simulations of $N = 10^4$ individuals with $\beta = 10^{-4}$ and $\gamma = 5 \cdot \beta$ evolving for $t = 10^4$ time units with $k = [1, 2, 4, 8]$ and $S_i = 0.9N$, $I_{1i} = 0.075N$ and $I_{2i} = 0.025$.

We show the results of such simulations in Fig. 4. We show that, for different value of k and by using the rate to probability rule of Eq. (25), we correctly recover the theoretical predictions of Eq. (24) with the overall relative error for each time step which is constantly below the 1% for all the evolution time (it is growing only for the S compartment at long evolution times as $S \to 0$).

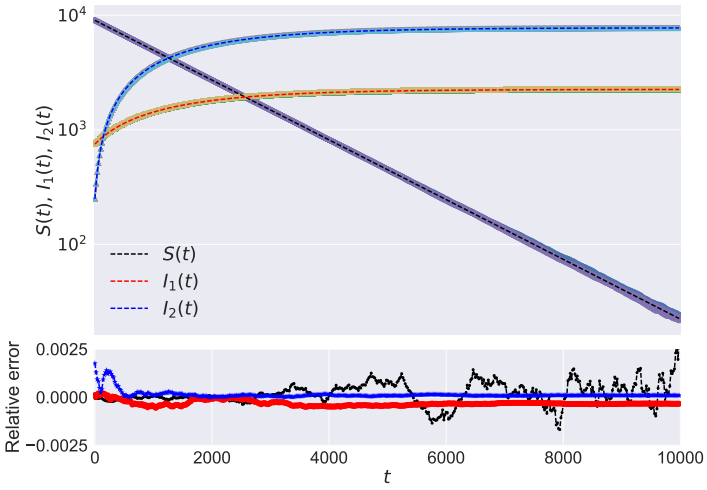


Fig. 4: (Top) The theoretical prediction for $S(t)$ (black dashed line), $I_1(t)$ (red dashed), and $I_2(t)$ (blue dashed) compared to the numerical results over 20 runs for different values of the discrete steps number $k \in [1, 2, 4, 8]$ (symbols that are not distinguishable below the analytical predictions). (Bottom) The relative error between the simulated number of individuals in a given compartment $C(t)$ and the theoretical prediction $T(t)$ computed as $(C(t) - T(t))/T(t)$ for the S (black), I_1 (red) and $I_2(blue)$ compartments.

3 Simulations

Data pre-processing

The data pre-processing comprehends the importing procedure of data from health and census agencies, their matching with geographical vector layers (shape-files), the computation of the model parameters and, lastly, the creation of the rasters to be given as input to the simulation.

We refer to the updated document we circulated on the SVN[1] for details, while we report here just the workflow we implemented.

Smoking and census data

We use data from [11, 15, 14] to get the number of people N_{LAD} living in each **L**ocal **A**uthority **D**istrict (LAD), as well as the number of never S_{LAD}, current I_{LAD}, and ex R_{LAD} smokers for each LAD.

We also retrieved the statistic on smoke quitting and relapse. In this case the data are available at a regional level (a superset of the LAD administrative division) so that a LAD↔Region lookup is needed to associate to each LAD the percentage of smokers who

[1] See the documents in the `WP04/tools/Pilot41_data_pre-processing/geoData` SVN repository.

tried quitting in one year and the percentage of success of such tries. We then merged these data by using the LADs unique codes as the merging key of the two tables. The resulting tables looks like:

LADcode	LADname	population	indicator	period	value	Frac. Quitter	F. Success 1month	gamma
E06000047	County Durham	517773	current smokers	2013	0.2209	0.0953	0.461	0.1002

Parameters estimation

The parameter that can be estimated per LAD is the quit rate γ together with the initial conditions on S, I and R. At a national level we can instead estimate the parameters of the the time to first relapse distribution $\chi^2(m)$ and the natality (mortality) rate μ. The transmissivity β is left free as the parameter to optimize during model calibration.

The evaluation of γ_{LAD} for each LAD is done by using the fraction $q_{\text{LAD}}(t)$ of current smokers that tries to quit (the column Frac. Quitter in the table above). By applying for each row (LAD) Eq. (15) we populate the *gamma* column with the corresponding γ_{LAD}.

The β parameter could be evaluated at a national level, by computing the number of new smokers for a given year, by approximating the number of new smokers from data, as we described above Eq. (19). Specifically, one can import from [11] the prevalence of smoking per each age bracket in the $[11, 24]$ age range, together with the population pyramid data from [14] and compute the number of smokers in each age bracket. By doing so, the number of new smokers is $N_{\text{new smok.}}(\text{year} = 2013)$ and β is set by the $dI'(t) = -dS'(t) = \beta I'(t)/NS'(t)dt = N_{\text{new smok.}}(t)$ relation, giving $\beta = 0.3067$ yr^{-1}. However, this estimation is quite coarse and indeed we will calibrate the model on β, both at the national scale and at the LAD level.

Finally, the $\mu = 0.0127$ yr^{-1} and the time to first relapse distribution $\chi^2(m)$ parameter $k = 0.55$ are derived nation-wide as discussed in Section 2.1.

The last step of data pre-processing is the creation of the rasters aligned with the SEDAC population raster [13]. We create a raster for each LAD-dependent observable and parameter, i.e., for $S(t)$, $I(t)$, $R(t)$ and γ. We set the global parameters μ, β, and δ as well as the number k of discrete steps to simulate per unit time in the configuration file instead.

Model implementation

The implementation of the model in `Pandora` is being made by importing the generated rasters that are used to track the number of individuals in a specific compartment in a given geographical cell. The model is the one outlined in Section 2.

We run simulations for $T = 96$ time steps with $k = 12$ on a yearly basis (i.e., each evolution step accounts for 1 month so that we simulate a total of 8-years evolution time from the beginning of 2013 to the end of 2020) and we serialize the $S(t)$, $I(t)$ and $R(t)$ rasters

every 6 time steps (i.e., every 6 months). We simulate the entire population of England that is $N \sim 5.7 \cdot 10^7$ individuals. The results of a single run are reported in Fig. 5(b) where we show the evolution of the smoking prevalence on a map for each LAD.

During each simulation step we evaluate the probabilities to pass from one compartment to the other accordingly to the respective rate. In particular, when simulating the $I \to R, S$ transitions we use Eq. (25) to translate the γ and μ rates to the corresponding probabilities, while we simulate the $S \to I$ transition rate for each cell c by computing $\tilde{\beta}_c = \beta I_c / N_c$ and then translating it to the transition probability.

Data post-processing and preliminary results

The output of the simulation consists in the rasters containing the information on the number of individuals within each cell thath belong to every compartment at each time step.

The implemented post-processing workflow consists in the following workflow:

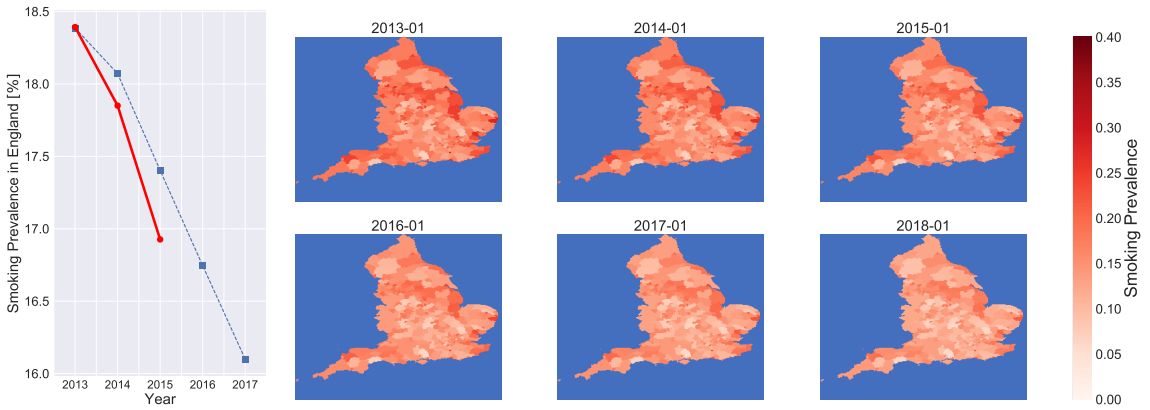
- import as a Python array the raster of each step and for each smoking indicator the corresponding entry in the `hdf5` file (the output of the simulations);
- query the geo-database for all the vectorial boundaries intersecting with the raster extension box;
- for every boundary, get all the cells that overlap with this boundary and aggregate over them the counter of agents in each compartment;
- divide the boundary's counter of each compartment (for example the current smokers $I_b(t)$ in administrative area b) by the population N_b of the same boundary to get the smoking prevalence $i_b(t) = I_b(t)/N_b$.

  In other words, by evaluating $I_b(t) = \sum_{\text{cell} \in b} I_{\text{cell}}(t)$, i.e., by summing the number of agents of in the I compartment at time t and dividing it by the constant total population N_b of the administrative area.

Geographical databse

We also developed the data pre- and post-processing so as to include the usage of the geographical mongodb database in the workflow. The latter allows us to easily perform different tasks:

- systematically store the census, economic, social, and health data in a consistent and persistent way and later query the database to retrieve data of specific geographical locations;
- retrieve census/health data at an arbitrary aggregation level (e.g., at the LAD, regional or national level) with the corresponding record the database;
- quick lookup of rasters cells intersecting with a given boundary;



(a) (b)

Fig. 5: (a) The national smoking prevalence in England as found in real world data [11] (red circles, period 2013-2015) and from preliminary simulations (blue squares, period 2013-2017). (b) The mapped prevalence per LAD in England during the simulation period.

- thanks to hierarchical structure of the database entries we can perform aggregation (e.g., aggregate regional to the national level) and disaggregation (downstream propagation of indicators from a national level to local quantities) from potentially any source of data that stores a matching key with the database entries (csv, pandas data-frames, shape-files, rasters, etc.);

- thanks to the built-in spatial search tools, we can easily retrieve and/or match data both in raster and shape-file form.

TODO: More details on the Geo-DataBase to come.

3.1 Future developments of the model

The model is a simple implementation of an epidemic process on top of the ABM structure. Though simple, it accounts for social pressure in the initiation mechanism and for a complete *initiation-quitting-relapse* behavioral circle. Nevertheless the model and the analytical framework allows for further development as other mechanisms may be included in the model as new data become available. Amongst them we point out the most important ones:

- include the social pressure mechanism also in the relapse process, thus adding a $I(t)/N$ term also in the $R \to I$ transition, by modifying the expectation value of the first time to relapse distribution $\chi^2(m)$;

- implement the synthetic population approach, i.e., assigning to the agents age, workplace/school location, income and other socio-economical indicators that affect their

propensity toward smoking initiation/cessation/relapse; in this sense, we are collaborating with GCF, IMT and Chalmers to enrich the synthetic population structure with a social network reproducing the real-world features of the individuals net of social contacts;

- add, when possible, compartment-dependent mortality rates so as to model the relative mortality risk depending on both their health behavior and the time passed as smokers and/or since having quit the habit;

- include policies and advertisement campaigns in the model by changing the model parameters depending on the supposed impact of such interventions, e.g., by modifying the transmission rate β or the quitting rate γ.

References

[1] John C. Lang, Daniel M. Abrams, and Hans De Sterck. The influence of societal individualism on a century of tobacco use: modelling the prevalence of smoking. *BMC Public Health*, 15(1):1–13, 2015.

[2] Rahmatollah Beheshti and Gita Sukthankar. A normative agent-based model for predicting smoking cessation trends. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 557–564, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

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

[4] D T Levy, K M Cummings, and A Hyland. A simulation of the effects of youth initiation policies on overall cigarette use. *American Journal of Public Health*, 90(8):1311–1314, 08 2000.

[5] Alessandro Vespignani Alain Barrat, Marc Barthelemy. *Dynamical Processes on Complex Networks*. 1 edition, 2008.

[6] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, May 2009.

[7] Carlos Castillo-Garsow, Guarionex Jordán-Salivia, and Ariel Rodriguez-Herrera. Mathematical models for the dynamics of tobacco use, recovery, and relapse. *Public Health*, 84(4):543–547, 1997.

[8] Local Tobacco Control Profiles for England. Smoking Quitters. `http://www.tobaccoprofiles.info/search/quit`, 2016. [Online; accessed 2-December-2016].

[9] Data of the French government. `https://www.data.gouv.fr/fr/`, 2016. [Online; accessed 10-November-2016].

[10] E. U. Commission. Eurostat, your key to European statistics. http://ec.europa.eu/eurostat, 2016. [Online; accessed 20-October-2016].

[11] United Kingdom National Health Service. Data and information. https://www.england.nhs.uk, 2016. [Online; accessed 12-December-2016].

[12] Data Gov UK. Opening uo government. https://data.gov.uk, 2016. [Online; accessed 12-December-2016].

[13] NY: NASA Socioeconomic Data Center for International Earth Science Information Network CIESIN Columbia University, Palisades and Applications Center (SEDAC). Gridded population of the world, version 4 (gpwv4): Population count adjusted to match 2015 revision of un wpp country totals. http://dx.doi.org/10.7927/H4SF2T42, 2016. Accessed 21 11 2016.

[14] Office for National Statistics. UK government. https://www.ons.gov.uk/peoplepopulationandcommunity, 2016. [Online; accessed 5-December-2016].

[15] Local Tobacco Control Profiles for England. Public Health England. http://www.tobaccoprofiles.info, 2016. [Online; accessed 8-December-2016].

[16] Janet Ferguson, Linda Bauld, John Chesterman, and Ken Judge. The english smoking treatment services: one-year outcomes. *Addiction*, 100:59–69, 2005.

[17] Keith Briggs and Christian Beck. Modelling train delays with q-exponential functions. *Physica A: Statistical Mechanics and its Applications*, 378(2):498 – 504, 2007.

EVS30 Symposium
Stuttgart, Germany, October 9 - 11, 2017

Electric mobility in view of Green Growth

Sarah Wolf[1], Steffen Fürst[1], Andreas Geiges[1], Gesine A. Steudle[1],
Jette von Postel[1], Carlo C. Jaeger[1,2]
[1]*Global Climate Forum, Berlin; Sarah Wolf (corresponding author) sarah.wolf@globalclimateforum.org*
[2]*Arizona State University*

Executive Summary

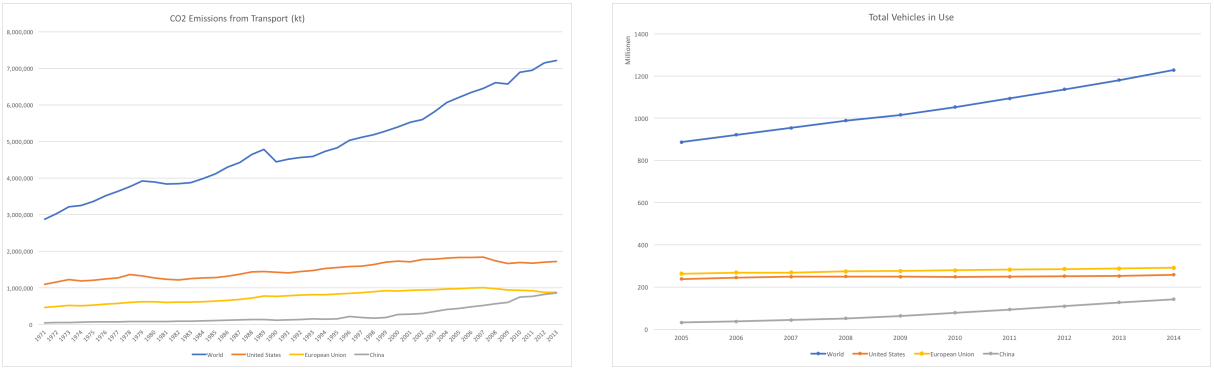
A transition of the global car market towards electric mobility can play a part in turning the risk of climate change into an opportunity of green growth, that is, in increasing environmental, economic, and social well-being. This paper presents work in progress on a global-scope high-resolution simulation model for analysing potential evolutions of the gobal car fleet, and in particular the diffusion of electric vehicles within it. Grounded in Global Systems Science, the approach takes a systemic perspective, draws on the framework of extended evolution, and uses agent-based modelling. Our aim is to engage with, and invite feedback from, the experts on various aspects of electric mobility gathering at this symposium.

Keywords: modeling, car, EV (electric vehicle), consumers

1 Introduction

"Green growth" is described by the OECD as a "twin challenge: expanding economic opportunities for all in the context of a growing global population; and addressing environmental pressures that, if left unaddressed, could undermine our ability to seize these opportunities."[1]
The global car fleet, counting more than 1.2 billion vehicles and growing, is expected to reach 2 billion by 2030 [2]. It contributes to environmental pressures, in particular, in terms of air pollution and CO_2 emissions; in the US, the transport sector has surpassed the power sector to become the number one emitter [3]. Currently, about 2 million cars, that is, much less than one percent of the global car fleet, are electric vehicles [4].
With increasing pressure to reduce CO_2 emissions of the transport sector, following from policies for avoiding climate change, with millenials not buying as many cars as previous generations did [5], and with the "3 revolutions" of vehicle electrification, automation and shared mobility [6] in sight, the global car market seems to be facing a transition. Traditional car manufacturers acknowledge this: Mary Barra, CEO of General Motors, believes "the auto industry will change more in the next five to 10 years than it has in the last 50" [7] and Volkswagen's CEO Matthias Müller has announced the company's deepest transformation since its foundation in the wake of its diesel emissions scandal [8]. The fact that Tesla has recently exceeded both Ford and GM in terms of market capitalisation, while producing a fraction of a percent both of the number of cars and of revenue in comparison with either of the traditional manufacturers, points to a strong belief of investors in the future of electric mobility [9].
Our research question on electric mobility in view of green growth is how a transition of the global car market can be achieved in such a way as to realise economic and environmental benefits at the same time, that is, to turn the risk of climate change into an opportunity of green growth. Therefore, we work on analysing potential future evolutions of the car centered global system. Rooted in Global Systems Science, our approach takes a systemic perspective, draws on the framework of extended evolution, uses agent-based modelling, and considers stakeholder engagement essential. In particular, we here present work in progress on a global-scope high-resolution simulation model for analysing the diffusion of electric vehicles in order to engage with and invite feedback from the experts on all kinds of aspects of electric mobility gathering at this symposium.
The remainder of this paper is organized as follows: Section 2 briefly introduces concepts and methods employed. Section 3 specifies the car centered global system, and Section 4 presents the modelling

Figure 1: Trends in CO_2 emissions from transport (left) and numbers of total vehicles in use (right)



work in progress with first results. Section 5 sketches two directions of further work, before Section 6 concludes.

2 Concepts and methods

This section briefly discusses the concept of green growth in general and with reference to the car centered global system. It sketches the field of Global Systems Science (GSS) in which this work is rooted, and some helpful concepts from the field of extended evolution. Finally, the tool of synthetic information systems is briefly introduced.

2.1 Green growth

The concept "green growth" comes with many definitions in the literature (see, e.g., [10, 11] for a collection and a review, respectively); however, this paper is not the place to go into these. Here, we define green growth with respect to "business as usual" (BAU) or "brown" growth. For simplicity, given a plausible growth path of the world economy in the 21st century, we may consider "green growth" those growth paths where at any point in time GDP growth is greater and greenhouse gas emissions are lower than along the brown path. Similarly, such a "relative" definition can be given in terms of other or more indicators for environmental, economic, and social characteristics of the world's development path. "Inclusive green growth" [12] can be defined by adding that inequality should also be lower than in the reference BAU path at any point in time, etc.

This paper is to be seen in the context of previous work [13, 14, 15] which has shown that climate policy, together with a set of other policies, has the potential to trigger a shift to green growth via the following mechanisms: given the current, fossil fuel based economy, a serious decarbonisation requires large investments. Large investments entail growth, jobs, and technical progress. However, no single economic actor has the potential to provide such large investments alone, and incentives for investing into a green economy are small as long as there is not a coordinated move towards it. Strict climate policy, combined with a credible investment impulse, can be the signal needed to re-coordinate investors' expectations towards green growth. Once triggered, the virtuous circle of investors' expectations, investments, technical progress and growth can keep the economy on a green growth path with larger investments, lower unemployment, higher growth, and lower emissions than in the BAU case.

Moving from the macro-economic view to a certain sector (transport) or activity (mobility), as is the case here, definitions cannot be adapted by simply replacing the economy by this sector in a one-to-one manner. While emissions reductions achieved within a sector can be considered separately, the accompanying transition in economic terms may go beyond this sector. For example, jobs lost in a "brown" sector may be replaced by jobs in other sectors rather than in a "green" counterpart of the original sector. A macro-economic view is therefore still necessary to consider green growth opportunities arising from a certain sector.

Considering green growth from a mobility perspective, emissions from the transport sector (about 70% of which are produced by road transport) are of particular interest. In contrast to important other sectors, the global trend is increasing ([16], and see Figure 1 (left)). In a plausible BAU growth path for the 21st century, this trend is likely to continue, not least because global numbers of cars are likely to keep increasing (see Figure 1 (right)). This trend reflects increasing wealth of the population in large parts of the world; at the same time, increasing income generally comes with increasing mobility needs [17],

which constitutes a feedback effect that stabilises the trend in global car numbers. Reversing the trend in transport emissions without curbing benefits (e.g., in terms of growth or employment) that relate to the existing trend in car numbers would lead to a green growth path with respect to mobility; with the above relative definition, slowing down the emissions trend compared with the BAU scenario is a sufficient criterion for "green".
However, one exact definition is not the point in this work. The global systems perspective and modelling tools used here (described below) allow for observing a large number of indicators in model runs. For example, in a model with high spatial resolution, the numbers of cars with an internal combustion engine can be aggregated to the level of cities, municipalities, etc. This allows to estimate, for example, not only greenhouse gas emissions, but also levels of air pollution in some areas of interest such as megacities. "Green growth" can then be defined on a case to case basis in terms of those indicators most relevant for a given study.

2.2 Global Systems Science

Global Systems Science (GSS) is an emerging research field that combines data-driven computer simulation modelling with engagement of stakeholders and citizens to support decision makers faced with global challenges (see, e.g., [18]). Green growth is such a global challenge due to the long-term and global-scope effects of (local) greenhouse gas emissions from the activities of up to 7 bn people.
On global challenges, a systemic perspective needs to be taken to develop evidence and understanding on the underlying global system and its potential future evolutions, in particular for looking at potential effects of alternative decisions.
GSS has a policy informatics side – it describes global systems with the help of computational tools (see Section 4 below) – and an engagement side: ongoing dialogues between modellers and decision makers help shape a simulation model in the most useful directions, so that it can address the questions decision makers have. Further, and more importantly, many of the details in addressing a global challenge involve value judgements and human behaviour. Understanding a global system and evaluating policy options includes engaging citizens in the policy-making and policy evaluation process at an early stage.
Global systems are complex systems, made up of a multitude of heterogeneous actors and other elements interacting in complex networks at multiple scales, giving rise, for example, to feedbacks in and path-dependency of the system evolution. This evolution is non-deterministic; the open future not only results from the systems' complexity which prohibits knowing all the details that would be necessary to describe such a system exactly, but also from its reflexivity: actors in the system can react to predictions made about the system, thus potentially invalidating these.
A first step in analysing a global challenge is therefore to identify the global system to be studied. The car centered global system considered here will be sketched in Section 3.

2.3 The framework of extended evolution

In conceptualising the open future of the car centered global system, the framework of extended evolution [19] provides a useful anchor. This framework considers not only random mutations of genes and natural selection according to fitness, but also regulatory networks (which determine the sequence and intensity in which genes are activated) and niches (the environment an organism lives in, that has often been shaped by earlier generations of the same kind of organisms) as important factors in the evolution of species.
Analogies between biological and technological (or, more generally, cultural) evolution have been studied (see, e.g., [20]). We think it is helpful to add the ideas of regulatory networks and niches when considering the car centered global system. Together, these restrain the space of possibilities that the system may realize in its further evolution, while keeping its dynamics in a non-deterministic mode with an open future.

2.4 Synthetic information systems

Having identified and defined the global system under consideration, one then represents a (usually much simplified) version of this system on the computer to run simulations, which allow one to explore, as in a virtual laboratory, possible scenarios of the system's future evolution and related uncertainties.
An agent-based model (ABM, see, e.g. [21]) represents many heterogeneous actors in this system as agents, their environment, and the complex networks in which they interact in model code. A model simulation run then carries out such interactions repeatedly, giving rise to a trajectory displaying potential overall system dynamics. Generally, many runs are carried out to account for uncertainty.
We speak of a synthetic information system when the ABM's agents are initialised by a synthetic population – a set of virtual agents that, for relevant characteristics, statistically match the corresponding

distributions found in the real-world population (see, e.g., [22] and references therein). Analysis and interactive visualisation of simulation results of the synthetic information system allow to analyse the system and various alternative decisions for gaining a deeper understanding and a better overview.
In particular, the work presented here is carried out in the context of enhancing GSS modelling through High Performance Computing (HPC) and Data Analytics (HPDA), for example by enabling the use of high-resolution data sets, by allowing models to grow in complexity and grow towards global scales, and by facilitating deeper analysis of larger sets of output data from model simulation runs (see coegss.eu). On the potential of agent-based models in addressing grand challenges of global scope see also [23].

3 The car centered global system

This section describes the car centered global system, drawing system boundaries and categorising elements as seen best fit for our overall research question. While the modelling work on this system (Section 4) does not represent all points introduced here, a structured description of the system (in terms of agents, environment, and interaction networks, but also thinking about regulatory networks and niches) is a fist step to analysing this – or really any – system.

3.1 A description of the system's elements

Viewing electric mobility from a systemic perspective requires the consideration of a large number of heterogeneous, spatially distributed agents.
The system first of all includes a global population of actual and potential consumers; we consider households as potential car buyers. Properties of households that are relevant here include the number of people in a household and their ages, the household's location, its mobility needs, its income, the number and properties of cars owned, and many more. In buying (or also in re-selling) a car, a household's decisions may be influenced by various factors. One of these is the local environment, that can include infrastructure (charging stations, public transport, etc.), or regulation at the level of their city or state. For example, some Chinese megacities have a quota policy to control car ownership growth: in Beijing the right to buy a conventional car has to be won in a lottery, in Shanghai such rights are auctioned [24]. Another factor influencing decisions is the local interaction with other agents, for example in terms of congestion and accidents. Last but not least, social influence can shape decisions, in networks between agents that can go beyond local interactions, and that are typically characterised by hubs, clusters, assortativity, as well as community and hierarchical structures [25].
Another type of agents in our system are firms in the car industry, with car manufacturers operating in a global market (as just one example, among Volkswagen sales in 2016, Europe accounted for about 41%, closely followed by about 39% in China [26]) and suppliers of car components. From a green growth perspective, the large numbers of people employed by the car industry, especially if the suppliers of car components are included, also play a role. For example, Volkswagen has about 600'000 employees worldwide, nearly half of which in Germany, where the company is the single largest private employer.
Key public authorities, in particular those of Germany, the U.S., China and Japan could be considered as yet another type of agents. However, it seems more useful here to think of the regulations they pass as part of the environment that households and firms interact in. Hence, this environment includes an administrative layer with laws and procedures governing the admission of cars on the road, the insurance of cars, and standards they have to meet (including the clean air standards violated in the recent diesel scandal). It also includes a geographically anchored layer with rural and urban areas and transport infrastructure like roads and gas or charging stations. The thus defined environment in the car centered global system co-evolves with the global car fleet, but changes at much longer time scales than that of car sales.
The car centered global system is an open system whose evolution is in turn influenced by what surrounds it, such as the global oil industry or geopolitical arrangements in climate policy. Like for the question of whether to include regulator agents or a regulation layer in the environment, the system boundaries are up to definition, and it may be useful to draw them differently when other questions are asked.

3.2 An extended evolution perspective on this system

Attention to regulatory networks and niches helps to understand the remarkable inertia of the car centered global system. Laws and regulations stabilise the car as a fundamental element of mobility in contemporary society. A "regulatory network" shapes the evolution of the car centered global system: regulations passed by different administrations are the nodes, or vertices, in this network. Car manufacturers, that have to consider the regulations in all places where they want to be in the market, constitute the links, or edges, between these nodes. As regulation can be passed at different and nested levels (e.g., the EU or the US, countries or states, and cities), this can be considered a multiscale network. Configurations

in this network may influence the evolution of the global car fleet (by specifying which manufacturer needs to meet which regulations) and may at the same time co-evolve with it (e.g., if some manufacturer withdraws from some market).
Prices can also be viewed as part of the regulatory network; they can play an important role in the process by which a given innovation survives and spreads, or shares the fate of most innovations: to disappear. Carbon prices on the one hand and battery prices on the other will shape the space of possibilities for a transition to electric mobility.
At the same time, the global oil industry is part of the niche that maintains a central role for the internal combustion engine in today's world society. Spatial regions with fully developed charging infrastructures can be niches from which electric mobility may eventually spread.

4 Simulating potential evolutions of the global car fleet

To start simple and add complexity step by step, we have not defined an agent-based model including all of the above system elements from the start, but focus on households and the demand side first. Also, we first defined and implemented a spatially explicit innovation diffusion model on a global scale.

4.1 A spatially explicit innovation diffusion model

The initial innovation diffusion model operates on grid cells on a global map, with a resolution of 2.5 arc-minutes, corresponding to about 5km by 5km at the equator. Scenarios for the evolution of total car numbers are provided as an exogenous input. The model considers only two classes of cars, "brown" and "green" ones. For now, we consider green cars to correspond to battery electric vehicles due to data availability; however, the model structure allows to easily replace this assumption with others, for which the required data is available. A diffusion of green cars takes place within the ranges set by total car sales per time step. This model can be considered a geographic cellular automaton: spatially differentiated input data and a basic transition rule determine the number of green cars in each next step taking into account the neighbouring cells. We run model simulations for a 2009–2025 timeframe.

4.1.1 Scenarios of the total number of cars

The input data required by the model consists of maps of cars bought per cell per time-step. These have been prepared by combining gridded population data and scenarios from [28] with a rate of car scrappage and numbers of cars per 1000 people per country. For the latter, we used data from OICA [29] for the first decade of the timeframe under consideration. To obtain scenarios for the second decade, car ownership scenarios were computed, on the one hand, by extrapolating current trends and, on the other hand, with the help of a model by Dargay and colleagues [27]. This model estimates numbers of cars per 1000 people based on a country's GDP per capita, population density, level of urbanisation, and a country specific saturation level. The necessary input data and scenarios were obtained from standard data sources [30, 31, 32, 33, 34]. Details can be found in [35].
Figure 2 shows resulting projections of total car numbers by continent, Figure 3 shows the same numbers in spatially explit manner for two points in time. While in 2009, cars were primarily concentrated in three world regions – North America, Europe and Japan with South Korea – fast growth in the numbers of

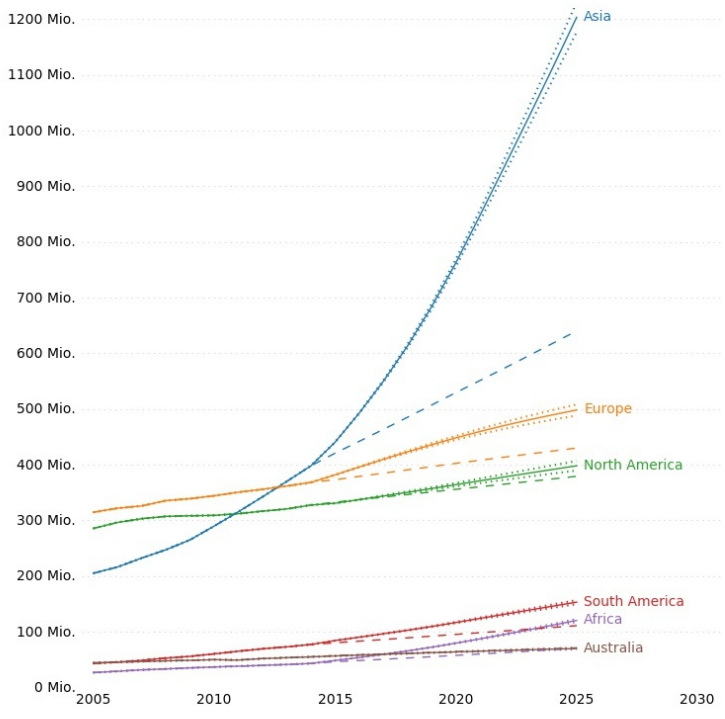


Figure 2: Scenarios for total numbers of cars by continent: Linear trends, and the model by Dargay et al with high, medium, and low population estimates [27]

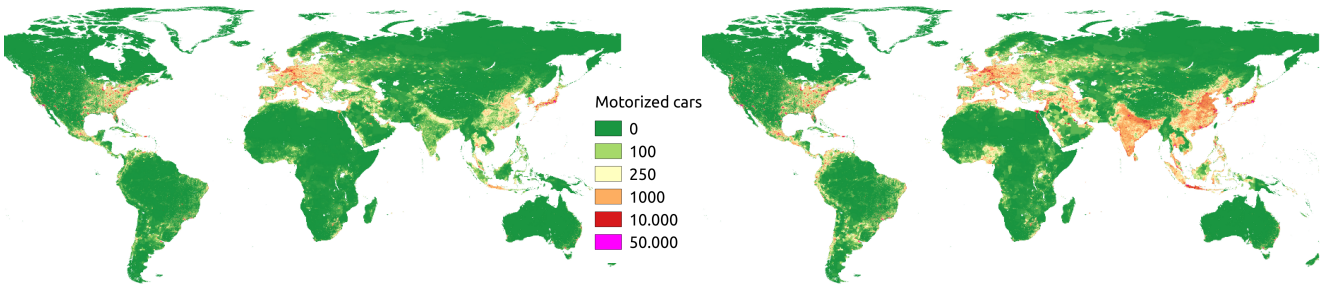


Figure 3: Total number of cars 2009 (left) and Dargay et al model scenario for 2025 (right)

cars happens primarily in China, India and Indonesia. Therefore, the future of the car industry will depend to a large extent on what will happen in this part of the world.

4.1.2 The diffusion mechanism

Given the spatially explicit dynamics for the total number of cars, each car bought in a model simulation can be a green or a brown car. The diffusion dynamics for green cars has an innovation and an imitation component.

- Innovation: as electric cars are already on the market, from time to time somebody will buy such a car for a variety of reasons, modelled as a random variable. However, to buy an electric car people need a certain income. Other factors being equal, the higher the relative GDP in the cell's country, the higher the probability that they will do so. Also, countries differ quite extensively in the level of policy support provided for electric mobility, in the form of subsidies, privileged access to lanes or parking spots, and many more. Therefore, the innovation component includes a policy factor, determined by calibrating model output to electric vehicle sales data.

- Imitation: the more electric cars are present in a given neighbourhood, the higher the probability that a consumer chooses one. This represents observations by this consumer and takes the number of green cars already present in this neighbourhood as an indicator of the existence of an electric-car-friendly infrastructure. In the current model version, the neighbourhood consists of 2 rings of cells around a given cell, and the respective numbers are weighted with inverse distance between the cells.

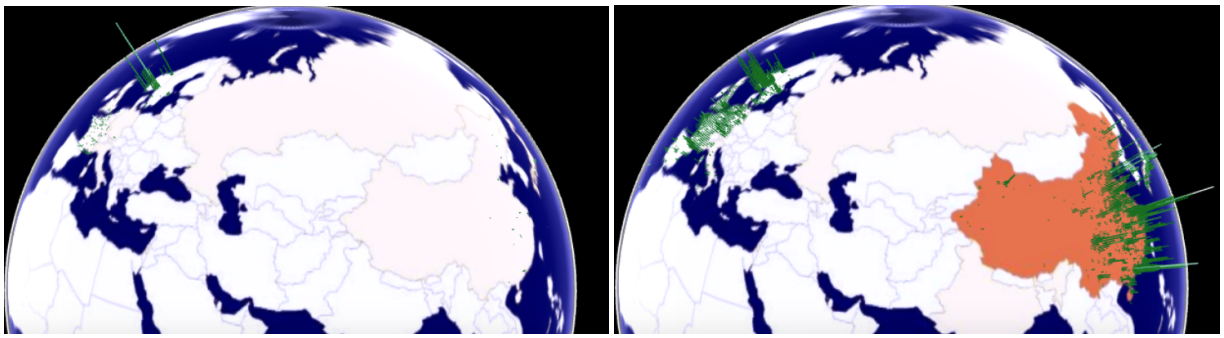
4.2 First results

The presented work also served to explore the use of HPC tools for agent-based modelling. Therefore, the model was programmed using Pandora [36] and run on the fine spatial grid of the population data with global scope. The country-specific policy parameter for the innovation component has been calibrated for the 15 countries listed in Table 1, according to data availability. It allows to compare country-specific incentives on the basis of a common model which takes the countries' GDP and spatial population structure into account. When analysing those policies that exist in each of these countries (see [37, 38]), together with the respective dates when these policies were enacted, the range of values spanned by this parameter, and the relative value for one country as compared to another can help structure the study of the effectiveness of certain policy mixes in certain economic, cultural and societal backgrounds. A step of model refinement which would calibrate the model to smaller sub-areas, for example in the US, could then be confronted with study results as in [39] to gain a better understanding of the diffusion of electric vehicles in various circumstances.

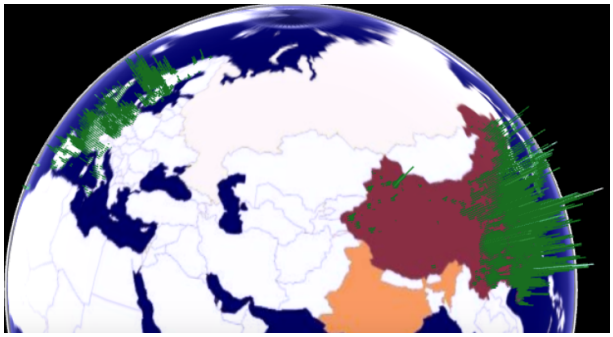
Table 1: Policy factors calibrated

factor	country
0.06667	Korea
0.2	Italy, Spain
0.26667	Canada
0.4	Japan
0.53333	United Kingdom, Portugal, South Aftrica
0.66667	United States, Germany, Netherlands
1.66667	China
2	Sweden, France
10	Norway

Figure 4: Shares (spike height) and total numbers of BEV (spike colour), and total number of cars (country colour)



2009 2017



2025

Model results can be visualised in global maps. Figure 4 shows screenshots on the results from these 15 countries obtained with an interactive visualisation tool that is based on R and shiny [40]. In the maps, the height of spikes depicts the share, and colour the total numbers of electric vehicles. In addition, total numbers of cars in the country are displayed by colouring the underlying country map.
In simulation results for 2009, Norway stands out as the lonely innovator for electric vehicles. The picture for 2017 shows that total numbers of cars in China have increased with respect to other countries, where electric vehicle shares are growing and the largest total amounts (white spikes) occur. Finally, the simulation for 2025 shows the largest EV shares and total numbers in China, and we see that total numbers of cars have massively increased in China, and are increasing also in India. To take a closer look at electric vehicle shares, which are picking up in urban areas both in China and in several places in Europe, the tool allows to zoom into certain areas and to move the globe.
In two dimensions, the map in Figure 5 shows that the Chinese coastline is likely to play an important role in the diffusion of electric vehicles. Three megalopolis, defined by the Chinese leadership around three key cities, Beijing, Shanghai and Shenzhen (next to Hong Kong), have the potential to become territorial niches for the initial stage in the evolution of individual mobility based on electric cars. Against this background, one can begin to evaluate strategic choices by relevant actors: for example, the Volkswagen decision to engage with leading Chinese IT firms like Huawei in view of the digitalisation of car traffic, rather than partnering with American companies like Apple or Google, makes perfect sense [41]. So does the recent announcement by Volvo (owned by the Chinese Geely Automobile Holdings) to introduce only hybrids or battery electric vehicles as new models from 2019 on [42], especially when viewed in the Chinese context.
A similar map for Northwestern Europe, Figure 6, shows a large urban chain from England, through the Benelux countries, Germany, Switzerland and France, connecting Manchester with Marseille and the French Mediterranean coast, with the Rhine valley as its backbone. If one wants to establish a powerful regional niche facilitating the transition to e-mobility in the midst of Europe,

Figure 5: Detail from simulation output for China: share of electric vehicles in 2025.

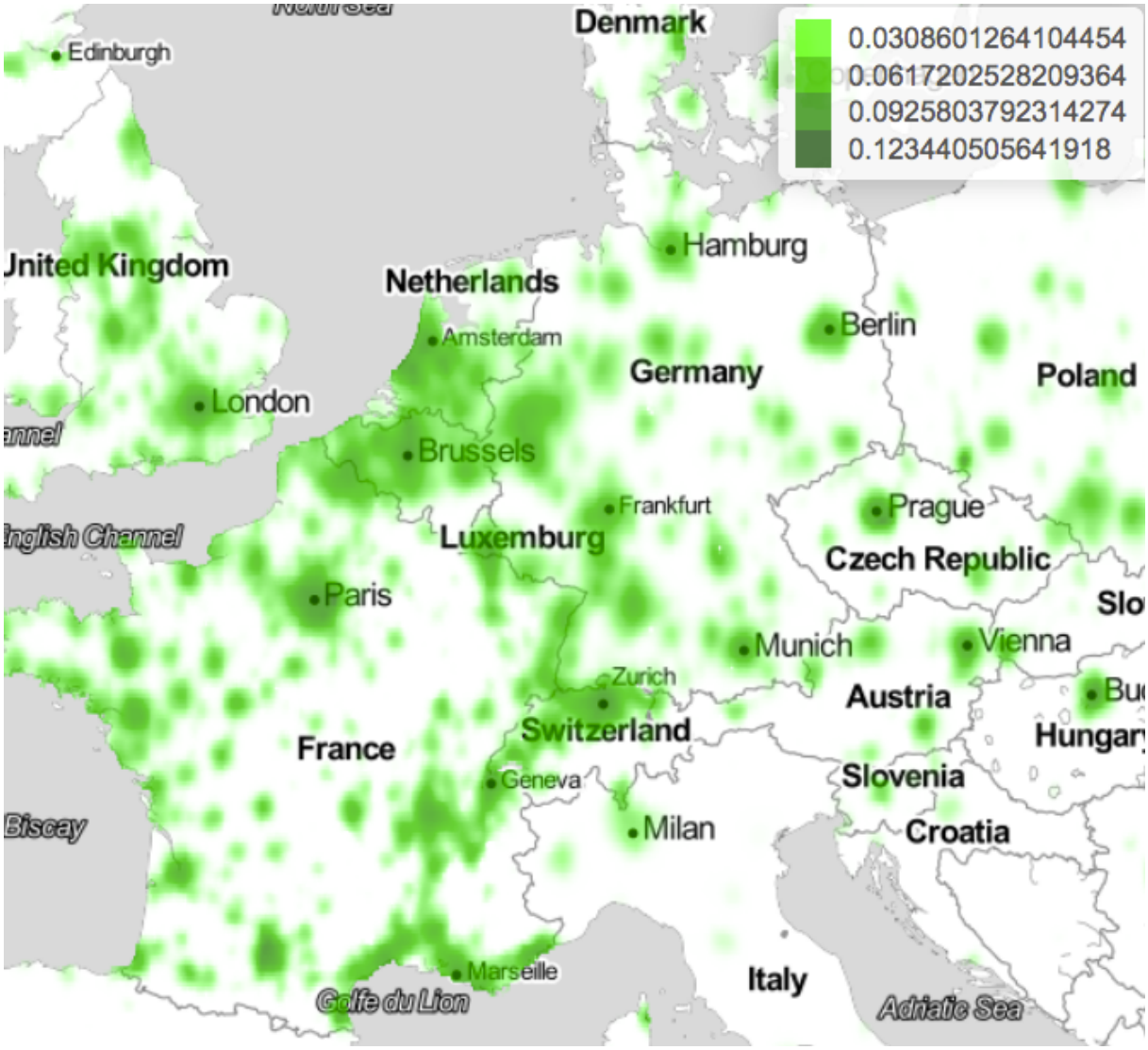


Figure 6: Detail from simulation output for Northwestern Europe: share of electric vehicles in 2025.

this corridor deserves special attention by policy makers. In addition, the ambitious project of the energy transition in Germany – with the explicit goal of shutting down nuclear energy and bringing greenhouse gas emissions close to zero – sets a context in which electric mobility can offer a possibility of using connected electric cars not only as mobility devices, but also as media of energy storage. Returning to the framework of extended evolution, this corresponds to an enlargement of functional characteristics, which can open up favourable spaces for an innovation in a fitness landscape that initially was not well suited for this particular innovation.

5 Further work in progress

To refine the model from the previous section in the spirit of Section 2.4, currently two complementary strands of work are in progress.

5.1 Basic agent dynamics with a refined synthetic population

Innovation diffusion models with a discrete time-step can be be disaggregated into agents (see, e.g., [43]). This is a next step we are undertaking with the model presented above. Without going into complex decision making procedures of the agents in a first step, the decisions will still be based on innovation and imitation probabilities. However, feeding the model with an existing detailed synthetic population, we can begin to differentiate agents by their properties, such as household size, income,

mobility needs, and an indicator for environmental concern. This shall be done in cooperation with a group that is specialised in producing synthetic populations. Also, the network between the agents can then be taken into account, which vanishes in the aggregate version. Network structures can make important differences in the diffusion of innovations (see, e.g., [44]), and social networks in the real world show quite a few particularities, such as a heavy tailed degree distribution (many nodes have very many connections), a high clustering coefficient (i.e., a large number of closed patterns: the friend of my friend is likely to also be my friend), assortativity (agents link predominantly to similar agents), community structures (clusters of agents much more linked to other agents within the group than to those outside) and hierarchical structures (subgroups of groups) [25]. Therefore, as one of the next steps in work with this model, networks shall be explicitly considered. This will allow us to analyse commonalities and differences with the cell-wise aggregate model, in order to explore at which points an explicit agent-based model can provide information on the underlying global system that a more aggregate model is unable to provide.

5.2 A basic synthetic population with refined agent dynamics

At the same time, we are working on a model that does take into account more complex decision making processes of the agents. In particular, we consider agents that take their decisions by optimizing expected utility, using subjective probabilities that they update based on information from other agents within their network. To initialise this model version, we updated and extended an existing synthetic population. Details, first results on how the mechanisms in this model play out in terms of overall system dynamics, and a discussion on data analytics tools for model analysis are presented in [45] and are beyond the scope of this text.

6 Conclusions and outlook

Obtaining green growth is one of the grand or global challenges society is currently facing. A transition to sustainable mobility is part of what needs to be achieved for this. The car centered global system, and the diffusion of electric mobility within this system, are elements that can play a crucial role. Vice versa, the idea of green growth can play a role in achieving a transition to sustainable mobility: in a world where cars largely resemble welfare, freedom, status, etc., and where conventional cars come with internal combustion engines, an alternative narrative, that may help switch to another convention, is wanting. Green growth puts the focus on benefits to be obtained from reducing emissions (by using electric vehicles and renewable transport, such as bikes, by sharing cars and rides, and so on) such as reduced air pollution or urban space freed up for other uses than traffic and many more. In this vein, this paper has presented an approach to and tools (under construction) for analysing electric mobility in view of green growth.

An important part of constructing these tools is the development of agent-based simulation models. This work can benefit from interaction with experts on various aspects of electric mobility: first, to shape the model into the most relevant directions, it is of interest to us which questions potential users of such a tool would like to see answered. Second, to fill in details into the current, still basic, model, expertise from the field is required. The present paper is intended to initiate dialogues with experts in relation to the Electric Vehicle Symposium.

While work in progress, the above examples of first results indicate how a simulation model and its (visualised) output can foster structured thinking about a global challenge, and enhance one's understanding of a given system by pointing out potential evolutions. Making oneself aware of potential futures in a given system is a first step in shaping its future evolution by then evaluating which of these futures are desirable, or which should be avoided, and which measures help steer the system in the direction of the former.

Acknowledgments

We thank two anonymous reviewers for helpful comments. Funding from the EU under grant 676547 (Centre of Excellence for Global Systems Science) is gratefully acknowledged.

References

[1] OECD, "What is green growth and how can it help deliver sustainable development?" 2017. [Online]. http://www.oecd.org/greengrowth/whatisgreengrowthandhowcanithelpdeliversustainabledevelopment.htm

[2] D. Sperling and D. Gordon, *Two billion cars: driving toward sustainability*. Oxford University Press, 2009.

[3] J. DeCicco, "Tailpipes top smokestacks as nation's largest CO2 emitters," http://www.carsclimate.com/2016/09/tailpipes-top-smokestacks.html, 2016.

[4] N. Lutsey, "The rise of electric vehicles: The second million," ICCT blog, published 2017.01.31, 2017. [Online]. http://www.theicct.org/blogs/staff/second-million-electric-vehicles

[5] T. Dutzik, J. Inglis, and P. Baxandall, "Millennials in motion, changing travel habits of young americans and the implications for public policy," U.S. PIRG Education Fund and Frontier Group, 2014. http://www.uspirg.org/sites/pirg/files/reports/Millennials%20in%20Motion%20USPIRG.pdf

[6] L. Fulton, J. Mason, and D. Meroux, "Three Revolutions in Urban Transportation," May 2017. https://www.itdp.org/wp-content/uploads/2017/04/UCD-ITDP-3R-Report-FINAL.pdf

[7] M. Barra, "GM 2015 Sustainability Report Chairman & CEO Message: To Our Stakeholders," https://media.gm.com/dld/content/dam/Media/images/US/Release_Images/2016/05-2016/Sustainability/GM-Sustainability-MaryBarra-Letter.pdf, 2016.

[8] Volkswagen, "Matthias Müller: We have launched the biggest change process in Volkswagen's history," Volkswagen press release, https://www.volkswagen-media-services.com/en/detailpage/-/detail/Matthias-Mller-We-have-launched-the-biggest-change-process-in-Volkswagens-history/view/3710903/7a5bbec13158edd433c6630f5ac445da?p_p_auth=NcZ61zbL, 2016.

[9] J. B. Stewart and W. Brangham, "Numbers Behind Tesla, GM, Ford, the "New Big Three"," 2017. [Online]. http://www.supplychain247.com/article/numbers_behind_tesla_gm_ford_the_new_big_three

[10] UNDESA, "A Guidebook to the Green Economy: Issue 1: Green Economy, Green Growth, and Low Carbon Development - history, definitions and a guide to recent publications," United Nations Division for Sustainable Development, 2012.

[11] M.-C. Rische, A. Roehlig, and J. Stoever, "Green, greener, grey. Disentangeling different types of green growth," *HWWI Research Paper*, no. 160, 2014. http://www.hwwi.org/fileadmin/_migrated/tx_wilpubdb/HWWI_Research_Paper_160.pdf

[12] The World Bank, "Inclusive Green Growth. The Pathway to Sustainable Development," The World Bank, 2012. http://siteresources.worldbank.org/EXTSDNET/Resources/Inclusive_Green_Growth_May_2012.pdf

[13] C. C. Jaeger, L. Paroussos, D. Mangalagiu, R. Kupers, A. Mandel, and J. D. Tábara, "A New Growth Path for Europe – Generating Prosperity and Jobs in the Low-Carbon Economy," Synthesis Report, A study commissioned by the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, Berlin, 2011. http://www.globalclimateforum.org/fileadmin/ecf-documents/publications/reports/A_New_Growth_Path_for_Europe_-_Synthesis_Report.pdf

[14] C. C. Jaeger, F. Schütze, S. Fürst, D. Mangalagiu, F. Meißner, J. Mielke, G. A. Steudle, and S. Wolf, "Investment-Oriented Climate Policy: An opportunity for Europe," A study commissioned by the German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety, Berlin, 2015. http://www.globalclimateforum.org/fileadmin/ecf-documents/news-events/2016_0217_Investment-oriented_climate_policy/Investment-oriented_climate_policy-_An_opportunity_for_Europe.pdf

[15] C. C. Jaeger, K. Hasselmann, G. Leipold, D. Mangalagiu, and J. D. Tábara, *Reframing the Problem of Climate Change: From Zero Sum Game to Win-Win Solutions*. Earthscan, 2012.

[16] International Energy Agency, "CO_2 Emissions from Fuel Combustion. Highlights," 2016. http://www.iea.org/publications/freepublications/publication/CO2EmissionsfromFuelCombustion_Highlights_2016.pdf

[17] D. Kalinowska and U. Kunert, "Ageing and Mobility in Germany: Are Women Taking the Fast Lane?" DIW Berlin Discussion Paper No. 892, May 2009. http://www.diw.de/documents/publikationen/73/diw_01.c.98162.de/dp892.pdf

[18] R. Dum and J. Johnson, "Global Systems Science and Policy" in: *Non-Equilibrium Social Science and Policy*. J. Johnson, A. Nowak, P. Ormerod, B. Rosewell, YC Zhang, Eds. Springer, 2017.

[19] M. D. Laubichler and J. Renn, "Extended Evolution: A conceptual framework for integrating regulatory networks and niche construction," *Journal of Experimental Zoology (Part B: Molecular and Developmental Evolution)*, vol. 9999, pp. 1–13, 2015. http://pubman.mpiwg-berlin.mpg.de/pubman/item/escidoc:723342:2/component/escidoc:723343/Laubichler_et_al-2015-Journal_of_Experimental_Zoology_Part_B_Molecular_and_Developmental_Evolution.pdf

[20] J. Ziman, *Technological Innovation as an Evolutionary Process*. Cambridge University Press, 2000, ch. Evolutionary models for technological change, pp. 1–11. http://assets.cambridge.org/97805216/23612/sample/9780521623612wsc00.pdf

[21] L. Tesfatsion and K. Judd, *Agent-Based Computational Economics*, ser. Handbook of Computational Economics. North-Holland: Elsevier, 2006, no. 2.

[22] F. Gargiulo, S. Ternes, S. Huet, and G. Deffuant, "An Iterative Approach for Generating Statistically Realistic Populations of Households," *PLoS ONE*, vol. 5, p. e8828, January 2010. https://doi.org/10.1371/journal.pone.0008828

[23] A. J. Heppenstall, A. T. Crooks, M. Batty, and L. M. See, "Reflections and conclusions: Geographical models to address grand challenges," in *Agent-based models of geographical systems*, A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, Eds. Springer, 2012, pp. 739–747.

[24] F. Suwei and L. Qiang, "Car Ownership Control in Chinese Mega Cities: Shanghai, Beijing and Guangzhou," 2013. [Online]. Available: https://www.lta.gov.sg/ltaacademy/doc/13Sep040-Feng_CarOwnershipControl.pdf

[25] M. S. Granovetter, "The Strength of Weak Ties", *Americal Journal of Sociology*, vol. 78, no. 6, 1973.

[26] Volkswagen AG, "Volkswagen Konzern liefert 10,3 Millionen Fahrzeuge im Jahr 2016 aus," Press Release, January 2017. [Online]. https://www.volkswagen-media-services.com/detailpage/-/detail/Volkswagen-Konzern-liefert-103-Millionen-Fahrzeuge-im-Jahr-2016-aus/view/4447022/657bd2c10865b3dee40ed4dc685c90fb?p_p_auth=AyVxrcS0

[27] J. Dargay, D. Gately, and M. Sommer, "Vehicle Ownership and Income Growth, Worldwide: 1960-2030," *The Energy Journal*, vol. 28, no. 4, pp. 143–170, 2007.

[28] Socioeconomic Data and Applications Center (SEDAC), "Gridded Population of the World, v3," http://sedac.ciesin.columbia.edu/data/collection/gpw-v3/maps/services, 2015, visited December 2015.

[29] OICA, "Vehicles in use," http://www.oica.net/category/vehicles-in-use/, 2015, visited December 2015.

[30] International Monetary Fund, "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, 2016, visited July 2016.

[31] U.S. Energy Information Administration, "International Energy Outlook 2016, World GDP per capita by region expressed in purchasing power parity," http://www.eia.gov/forecasts/aeo/data/browser/#/?id=47-IEO2016&sourcekey=0, 2016, visited July 2016.

[32] United Nations Department of Economic and Social Affairs, "World Population Prospects, the 2015 Revision," https://esa.un.org/unpd/wpp/Download/Standard/Population/, 2015, visited July 2015.

[33] World Bank, "Land area (sq. km)," http://data.worldbank.org/indicator/AG.LND.TOTL.K2, 2015, visited July 2016.

[34] United Nations Department of Economic and Social Affairs, "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," https://esa.un.org/unpd/wup/CD-ROM/, 2014, visited July 2016.

[35] S. Wolf, M. Dreyer, M. Edwards, S. Fürst, A. Geiges, J. Hilpert, J. von Postel, F. Saracco, M. Tizzoni, and E. Ubaldi, "CoeGSS Deliverable D4.4: First Status Report of the Pilots," 2016. http://coegss.eu/wp-content/uploads/2016/02/D4.4.pdf

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

[37] EAFO (European Alternative Fuels Observatory), "Battery electic vehicles for European countries," http://www.eafo.eu, 2017, visited December 2017.

[38] IEA, "Global EV Outlook 2016," https://www.iea.org/publications/freepublications/publication/Global_EV_Outlook_2016.pdf, 2016, visited December 2016.

[39] Z. McDonald, "Comparing the Top 10 Cities for Electric Vehicle Adoption," Posted August 11, 2016. [Online]. http://www.fleetcarma.com/top-cities-electric-vehicle-sales

[40] "shiny," 2017. [Online]. https://shiny.rstudio.com

[41] Huawei Technologies Co., Ltd., "Huawei and Volkswagen Collaborate to Connect Automobiles and Smartphones," Press Release, May 2015. [Online]. https://www.huawei.eu/media-centre/press-releases/huawei-and-volkswagen-collaborate-connect-automobiles-and-smartphones

[42] J. Ewing, "Volvo, Betting on Electric, Moves to Phase Out Conventional Engines," The New York Times, July 2017. [Online]. https://www.nytimes.com/2017/07/05/business/energy-environment/volvo-hybrid-electric-car.html?emc=edit_nn_20170706&nl=morning-briefing&nlid=77854954&ref=business&te=1

[43] E. Kiesling, M. Guenther, C. Stummer, and L. M. Wakolbinger, "Agent-based Simulation of Innovation Diffusion: A Review," *Central European Journal of Operations Research*, vol. 20, pp. 183–230, June 2012.

[44] M. Edwards, S. Huet, F. Goreaud, and G. Deffuant, "Comparing an individual-based model of behaviour diffusion with its mean field aggregate approximation," *Journal of Artificial Societies and Social Simulation*, vol. 6, no. 4, 2003.

[45] A. Geiges, O. Allerbo, S. Fuerst, and S. Wolf, "Modeling consumer- and innovation-driven transitions of the car market: Modeling and analysis concepts," 2017, prepared for the 2nd Workshop on Agent-based modelling at ESCP Europe Berlin. Available upon request.