D4.1 – FIRST REPORT ON PILOT REQUIREMENTS

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	31.12.2015
Actual date of submission	26.01.2016
Dissemination level	Public
Nature	Report
Version	2 (after internal review)
Work Package	4
Lead beneficiary	GCF
Responsible scientist/administrator	Sarah Wolf
Contributor(s)	Daniela Paolotti, Michele Tizzoni (ISI), Margaret Edwards (CoSMo), Steffen Fürst, Andreas Geiges, Alfred Ireland, Franziska Schütze, Gesine Steudle (GCF)
Internal reviewers	Miroslaw Kupczyk (PSNC), Cezar Ionescu (Chalmers), Jochen Buchholz (USTUTT)
Keywords	Health habits, green growth, global urbanisation, GSS, HPC, requirements
Total number of pages:	49

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

Version History

	Name	Partner	Date
From	**Sarah Wolf**	**GCF**	
First Version	**for internal review**		**Dec 4, 2015**
Second Version	**for submission**		**Dec 23, 2015**
Reviewed by	**Cezar Ionescu, Miroslaw Kupczyk, Jochen Buchholz**	**Chalmers PSNC USTUTT**	**Dec 6, 2015 Dec 7, 2015 Dec 14, 2015**
Approved by	**Coordinator**	**UP**	**Jan 26, 2015**

Abstract

This document presents requirements of the pilots to a Centre of Excellence for Global Systems Science in order to act as a steering document for the developments and operation in WP3 and WP5 and present the need of a Centre of Excellence in Global Systems Science from an economic viewpoint (WP2). It is the initial version of a living document with further iterations at months 16 and 26 that will identify potential changes in the requirements (or add new requirements) as they become apparent while the project work progresses.

The three pilots develop a synthetic information system for modelling and analysing global challenges in the fields of health habits, green growth, and global urbanization, respectively. This comes with requirements in the fields of high performance computing (HPC) as well as Global Systems Science (GSS) and requires the creation of a synergetic link between HPC and GSS. The present document details these requirements as far as they appear at this early stage in the new endeavour of joining the two fields.

Table of Contents

Abstract 2

Table of Contents 3

1 Introduction 4

2 Common requirements of all pilots 7

3 Task 4.1: Health Habits 15

4 Task 4.2: Green Growth 20

5 Task 4.3 Global Urbanization 29

6 Task 4.4 Future Applications 43

7 Conclusion 44

8 References 45

1 Introduction

1.1 The CoeGSS pilots

Three exemplary GSS challenges have been chosen as pilot studies for the Centre of Excellence for Global Systems Science:

Health Habits: a pilot study into the global diffusion of health-relevant habits such as smoking, overeating or physical exercise (Task 4.1)

Green Growth: a pilot study into the possibility of green growth, i.e. increased well-being in the economic, ecologic and social dimension, investigating the diffusion of initiatives such as feed-in tariffs, green business strategies, and individual lifestyle changes (Task 4.2)

Global Urbanisation: a pilot study into the systemic impact of infrastructure decisions on key urban performance indicators such as congestion, real estate prices and emissions (Task 4.3)

The common task of these pilots is to develop a flexible, large-scale synthetic information (SI) system, allowing access to the capabilities of HPC for tackling these challenges. The context for this work is set by the work package (WP4), which apart from the pilot tasks includes a "Future Applications" task that shall identify needs and opportunities for the future of HPC applications to complex global, anthropocentric systems.

1.2 Synthetic information systems

SI systems computationally integrate data from different sources in order to provide simulation tools for creating virtual scenarios that, for a given GSS question, reflect the relevant aspects of the global system under study with the appropriate granularity. CoeGSS develops such systems to expand the capability of policy makers and other stakeholders, such as businesses, to visualise the impact and trade-offs entailed by different courses of action. In particular, CoeGSS shall provide an HPC-based framework to generate customized synthetic populations for GSS applications as a tool to study global contagion dynamics for GSS topics as diverse as public health and markets.

The study of GSS problems presupposes dealing with large and complex anthropocentric systems. Therefore, very large, detailed agent based modelling (ABM) approaches have to be developed in addition to using more traditional mathematical modelling methods. Data sources vary from structured census records to unstructured raw and dynamic streams of information from the internet and mobile devices. Data-centric high performance computing (HPC) is a key to building and working with such computationally intensive synthetic information systems.

Synthetic populations are software representations of real populations that match statistical distributions of relevant aspects from reality. Mostly, synthetic populations model human populations, but other entities (such as firms, banks, cars) can also be represented.

For synthetic populations representing humans, synthetic individuals will come with a range of demographic attributes such as gender, age, income level etc. Synthetic households then consist of synthetic people with given relations between their ages. Personal and behavioural attributes can be added as needed and depending on the availability of data sources. Due to the synthetic nature, privacy of individuals is not compromised while statistical properties relevant to the problem under study are reflected by the virtual population.

The set of synthetic households can then be equipped with a set of activity patterns, including types and timing of activities, based on thousands of responses to an activity or time-use survey. Methods to replace traditional time-use surveys are an active area of research and can be incorporated in SI platforms as they appear. The activity pattern of a survey household is then attributed to synthetic households, where the matching between survey and synthetic households is done statistically, based for example on the number of workers in the household, or the number of children of various ages.

Geographic locations can then be assigned to households and their activities, based on land-use data and patterns, tax data, commercial information etc., and calibrated against travel time and traffic flow distributions. The population, with activity locations and itineraries is refined with the help of sophisticated techniques in combinatorial optimization, machine learning and agent-based modelling.

Activity-based social contact networks are another element of synthetic information systems. Interactions and dependencies in these networks will depend on questions under study, as people are involved in different social contact networks regarding different aspects of life (e.g., family, professional, religious networks, etc.).

Synthetic information cannot be collected by simple measurements or surveys. Rather it results from computationally mediated integration of data and the projection of information derived from diverse data sources onto formalized individual, computationally defined entities. The system of these informational entities are then analysed using further computational methods, for example, agent-based models.

1.3 About this document

This document presents requirements of the pilots to a Centre of Excellence for Global Systems Science in order to act as a steering document for the developments and operation in WP3 and WP5. It is the initial version of a living document with further iterations at months 16 and 26 that will identify changes in the requirements as they become apparent while the project work progresses.

SI projects require that research and innovation attention be directed to application problems, methods, architectures and platforms. Therefore, three GSS challenges have been chosen as pilot studies for CoeGSS – health habits, green growth, and global urbanization. The three fields are each very briefly described in the dedicated sections (3-5) below, to provide some background.

Initial pilot requirements to the centre of excellence are presented in Sections 4-7: Section 4 gives an overview of (more general) requirements that all three pilots have in common. Sections 5, 6 and 7 then each treat one pilot in particular, describing more specific requirements. Differences between the pilots arise not only from the different topics considered, but also from the different points of departure in terms of already existing tools and experience. The health habit task can draw on synthetic populations and models used in epidemiology, for the green growth task, experience in economic agent-based modelling can be built on, and for the global urbanization task, a suite of city models for various aspects (housing, transport, etc.) exists and can be combined with synthetic populations yet to be developed for creating more accurate simulations.

The initial requirements can act as a well-defined baseline for developments. However, a more complete and more precise set of requirements will arise from the project work, almost certainly revealing also some further changes to be made to the initial set. The living-document-nature of this deliverable accommodates this fact, and further releases will make the necessary adaptations. For project internal use, as input for example to WP3 and WP5, the relevant sections of the document may also be updated on shorter time scales than the deliverable schedule of CoeGSS foresees.

Section 8 adds a short input on the fourth task of the pilots work package. As future applications are not at the level of pilots that are being carried out, no concrete requirements can be stated at this point. However, the outlook on this task has been included in this document to provide a possible direction for future requirements.

Section 9 concludes.

2 Common requirements of all pilots

CoeGSS creates an intersection between the two as yet unrelated universes of HPC and GSS. During the starting phase of CoeGSS, it is therefore of utmost importance to bring together the two communities involved. What is needed for doing so may not be considered a requirement in a technical sense, but is nevertheless a sine qua non for the success of setting up the centre of excellence. Therefore, this section considers a few elements that are helpful or even necessary for bringing together the communities, before outlining general requirements that directly concern the task of building an HPC-GSS synthetic information system. More detailed and concrete requirements will emerge in the cooperation between HPC and GSS experts when working on this task. This work will be an iterative process of specifying synthetic agents and agent-based models, programming code, collecting and integrating data, running models at the HPC Centres, visualising and analysing simulation results, learning from them and revising or extending whatever necessary, and possibly further activities. In this process, HPC and GSS expertise will need to be merged for carrying out all the activities needed.

2.1 Interaction between HPC and GSS experts

So far, typical HPC applications concern quite homogeneous domains of discourse involving specialists sharing a common background, be it of chemical engineering, climatology, geosciences, etc. GSS, on the other hand, involves experts from very different fields, such as demography, economics and medicine, and on top of this, practitioners from business, policy-making and civil society. One of the most important developments in CoeGSS is thus the emergence of a common language between such groups, in particular between people from the HPC and the GSS sides within the project. Of course the emergence of a common language cannot be planned in a detailed manner, but networking activities can be undertaken to foster it. While some kind of interaction between project partners is needed in any project, it is mentioned here explicitly because close interaction is absolutely necessary for CoeGSS, and because it can be expected to be more difficult than when interacting within a discipline or a set of similar disciplines.

In order to learn enough of each other’s languages, a lot of communication is needed between HPC and GSS experts both in written and oral form, via documents, phone conferences and face-to-face meetings. Even at the project's early present stage, it has been seen that getting to know each other at the Kick-off meeting in Stuttgart in October 2015 was very helpful to initialise contact between the different groups – it matters for telephone conferences and emails, whether one has met the people one is communicating with or not. Deliverables are important documents for project internal exchange, as has been revealed by project partners' comments to an earlier draft of this deliverable and by learning from drafts of the initial deliverables of WP3 (conveying insight, e.g., on how different groups use the term "synthetic population" either as the set of synthetic individuals that can be used as agents in an agent-based model or as including the model itself) and WP5 (as an introduction to hardware, software, existing services of and access to the supercomputing centres in the project). These are just a few examples of the types of communication that are required, and, in particular at the start of the project will need to play a larger role than in projects rooted within an already established field.

2.2 Training

Training for both HPC and GSS groups, or intense interaction between the groups where each tries to learn from the other, is required to foster the development of a common language, as well as to provide GSS experts with the necessary skills for using the HPC infrastructure. The training task in CoeGSS (T6.3) should provide mutual training for project partners from the HPC and GSS communities. This should include introductions to both worlds, where efforts will have to be made not only by those who learn but also by those who teach, to present in a way that is accessible to non-experts of the field. More importantly, however, there should be a practical part of working on synthetic populations, models, simulations, and/or visualisation of at least one of the pilots. Feedback from this kind of training within the project will also be useful for setting up a training curriculum at the intersection between HPC and GSS.

On the one hand, HPC is a new tool to most partners involved in the project from the GSS side. On the other hand, for using the facilities of the supercomputing centres in the project, it must be ensured, of course, that people accessing the system have the necessary knowledge to use a supercomputer. Therefore, instructions or a more technical training, in some cases also introductory, are required for using the HPC systems and using them well. For example, runs on the system have to be well planned in advance. HPC experts should inform the pilots' GSS experts on the important points to consider when carrying out these runs.

2.3 Access and computing time

Two rather technical requirements are obvious:

- The project partners who run HPC simulations within each pilot need access to the systems at the supercomputing centres for doing so. Deliverable 5.1 describes how access is granted at HLRS and PSNC, so this is already taken care of. Further, D5.1 provides some introduction about how to use the systems, and some exchange supporting non-HPC-experts to carry an epidemiologic model over to a Stuttgart supercomputer for testing has taken place. While obvious, the point is listed here for completeness and as a reminder that many of the GSS partners have never used HPC before.

- Also obviously, computing time is needed for testing and calibrating models, and for producing simulations to explore (policy) scenarios. D5.1 also explains why the initial budget allocated to computing time within the CoeGSS project will probably not be sufficient for running all desired simulations and suggests how further required computing time may be allocated via the PRACE research infrastructure. Also federal funding of computing cycles is elaborated, e.g. the German research partners can apply for that directly..

2.4 Building a CoeGSS synthetic information system

Having roughly described synthetic information systems above, this section lists activities necessary for building and running HPC-GSS synthetic information systems. Each of these points to requirements that are further specified in the pilot specific sections or will be specified in the course of the project work.

Synthetic populations are already being used to explore scenarios in decision support, for example in the context of epidemiology. There are existing frameworks for developing synthetic populations and for running simulation models using these. Some of these frameworks are geared towards HPC. Also, there are some HPC frameworks for agent-based modelling, which may be useful for running pilot simulations using synthetic populations. It is a stated aim of CoeGSS to enhance and extend the capabilities of existing frameworks in the following directions:

1. increase the scale to global populations (billions of agents)
2. develop new methods for deriving relationships and activity patterns for agents
3. incorporate new data from social media sources in a running simulation
4. visualise the results of simulations and be able to interact with them in real time
5. test, verify, and validate the results

In the following list of activities that have to be carried out to build an SI system, these points are referred to where they apply. The activities will not generally be carried out simply in the order of the list. Rather an iterative process of stepping back and forth between them will occur. Activities are roughly grouped by points of reference: specification, data, models, and model output.

- **Defining the system:** depending on the problem to be studied, agents have to be defined with the relevant characteristics, environment, interaction networks, and activities.
- **Collecting data:** data must be collected that describe the real world population whose characteristics are to be statistically matched. Modelling global

populations (as aimed at with point 1.) implies global data requirements. For many types of data, the availability differs for different parts of the world, and for some parts of the world, data of interest may simply not be available, requiring methods to substitute for non-available data. However, this requirement is not the most urgent at the beginning of the project, since pilots can begin with smaller regions where good data is available.

Data collection concerns not only the agents' characteristics, but also the dynamics to be modelled. In epidemiology, the important micro-level dynamics is related to disease transmission between people. Relationships and activity patterns of agents in the respective synthetic populations are constructed in such a way as to faithfully represent these micro-dynamics. For example, an important point is to represent physical encounters between agents, where contagion can take place. For the global challenges under study in CoeGSS, micro-dynamics of "contagion" are yet to be studied, relating to aim number 2 above. The questions of which relationships and activity patterns of agents are the relevant ones, and how to represent these in models, requires the development of new methods, for example in the fields of data analytics or data collection. It has to be researched which kind of data (for example from social media) can be useful in designing the relevant model relationships and activity patterns, or how data collection, especially for individual data collected via apps or through other means, can be organized. On the HPC side, the data-centric approach of GSS-SI system may require new software technologies.

- **Building a shared pool of data:** all pilots should be able to benefit from data collection in the other pilots. With a view towards a framework for generating customized synthetic populations for addressing GSS challenges – that CoeGSS aims to provide – a common pool of data to build on is also required. Of course, each new problem to be addressed will require adding new data, but some data will be needed repeatedly, such as basic census data or gridded population data. A common pool of data that all pilots feed, and from which all pilots and future applications can use whatever is needed would therefore be very useful to have. To set up such a pool, a common data structure will be needed to access various data sets. Identifiers for each type of data (country specific data, geo-located data, agent-specific data, household data, census data, etc.) need to be defined for connecting the different sources. The data structure and chosen identifiers need to be designed to allow fast extraction of data, grouping and reorganizing. In particular, the data format used needs to support partial load and save functionality, as well as fast re-structuring. The pilots, that have already agreed to share data, but have not fixed any structures yet, should interact with data analytics experts from WP3 to develop a common data pool. Storing the data then is an issue that requires also the expertise and support of the HPC centres.

- **Pre-processing data:** GSS-simulations will require data to be globally available, independent of the source, quality and availability of the data. However, actual quality and availability of data will vary globally, but also within smaller regions to be modelled. To feed models, more or less homogeneous sets of data will be needed, so that pre-processing of data is required – it should be done automatically as far as possible. The pre-processing of data to fill gaps can be based on different approaches, like geo-statistical interpolation, inference from secondary data and eventually expert knowledge. Interpolation, finding correlated data sources for inferring missing data and handling different data structures is a common task in Big-data applications. Thus, methods from this field need to be included. It needs to be determined whether data interpolation remains a pre-processing task or needs to be included in the simulations. This decision mainly depends on the question whether it will be feasible to provide all data as one data set, or whether limited storage or access times favour for online-interpolation.
- **Ensuring and tracking data quality:** For each pilot sensitive data need to be identified, in which high data quality is required to ensure good results. For example, identifying the potential of electric mobility will require most accurate data in developed countries with high income, as well in fast-developing countries. Data in regions of very low income, on the other hand, may not contribute to the accuracy of the model. Quality of the data and the relevance for the specific models must be matched. This would ensure that important and relevant data are available at high quality. Also, when generating missing data by different methods of pre-processing of a dataset with gaps, it is necessary to track the quality of these estimates throughout all pre-processing steps for quality assurance.
- **Generation of synthetic populations:** software is required for creating the synthetic agents as entities on the computer according to the system definition and data. Some software tools for doing so are described in Deliverable D3.1. The exact choice of software tools that then become CoeGSS services will be made by the pilots through evaluation, and trying out the tool that seems to fit best, while being open to switch if necessary. Should no existing tool be adequate, or for license problems, for example, it might be necessary to program new software for generating the populations. This would then pose a new set of requirements: studying and evaluating methods available in the literature, and the whole cycle of software creation requirements (specify, program, test, etc) for the tool to be built.
- **Agent-based modelling:** for analysing the evolution of the system, an agent-based model in which the synthetic agents repeatedly interact needs to be developed. Sub-activities here are model specification, programming the model, documenting it, testing it at various stages. ABM frameworks can help

programming an agent-based model, and if used, become requirements (for example, the green growth task prefers to use such a framework, while the health habits task does not, see the respective sections).

- **Prototyping:** This is a further sub-activity of developing and programming models that is of particular importance at the start of the project – starting simple and in an iterative manner to be able to quickly check what works and what might not. During the prototyping phase, it is crucial to balance the fast development of running demonstrators and the required consideration for later HPC features. Including HPC features early on can be supported by early access to an HPC testing environment. This can be either the access on a testing node on the HLRS or PSNC infrastructure or a virtual environment, provided by HLRS or PSNC that matches the HPC system. Development access and the ability to test early prototypes would speed up the process to include and test HPC abilities.
- **Performing sensitivity analysis and calibrating models:** as ABMs with many agents quickly become complex, calibration is an important issue. Sensitivity analysis, to find out how much certain parameters influence the model output is a helpful first step. Finding sets of parameters that produce sensible system evolutions will require massively multiple model runs. Doing such runs in a structured way and interpreting the results requires software tools. Which ones are the tools of choice will yet have to be found out. Experience from HPC experts is welcome for helping the pilots find the most useful tools for their cases. Point 5 among the aims mentioned above relates to this activity: testing, verification, and validation of results are necessary in any modelling activity. However, given that CoeGSS pilots will use new data sources for large synthetic populations, there is a need to undertake these steps very carefully, and discover requirements (e.g. new methods) while entering this new field.
- **Installing code:** obviously, for running any simulations, software needs to be installed on the HPC systems. This includes the model code and libraries of ABMs, and possibly of frameworks used by the models; possibly also the software used for generating a synthetic population.
- **Optimizing code:** when using a pre-existing synthetic information system that shall be run on high performance computers, code optimization may be required to speed up simulations – see Section 3.7 for how an epidemiologic SI system shall be transferred to HPC for testing and as a basis upon which to build the system for health habits. When developing new synthetic information systems – as will be done, for example in the green growth pilot – code optimization might be carried out while writing the code by collaboration between modellers and HPC experts.
- **Running simulations:** this activity has already been mentioned e.g. for calibration, but it is also the core activity for decision support. In the latter sense, the CoeGSS aim of incorporating data from social media sources in a running

simulation (Point 4 above) relates to this activity. It requires the technical availability of this feature in the input structure of the models (integration of input at runtime) and in how the simulations are run (retrieving the data when needed).

- **Analysing output:** once simulations have been run, output data need to be analysed in order to understand the results, both in a calibration phase and when using the model in decision support. As the modelled systems are not deterministic, in most cases, many realizations will have to be analysed together, e.g. using statistical tools, to extract results from the sets of simulations. The exact tools required will be clear only at a later stage in the project.
- **Visualising output and results:** output data from simulations or from the just mentioned analysis of the output data need to be visualised, in order to facilitate understanding of and communication about the results. In particular, for providing decision support for stakeholders, results have to be made accessible and presented in easily understandable ways. A basic element in the visualisation of synthetic populations is a map of the area the population is located in. Simpler results can be visualised using colours or the height of spikes on such a map, but visualising more complex results requires new ideas. In particular, visualisation of uncertainty will be important for all pilots, in that the synthetic information systems are to be employed to explore possible futures. This may contain not only statistical measures over large sets of simulations carried out, but given the complexity of the fields under study it may also be of interest to visualise subjective probabilities representing expert judgements, different levels of the quality of input data for different parts of the world simulated, etc. Here the pilots will work together with the visualisation experts from WP3.
- **Creating a user-friendly interface for non-experts:** some form of a user interface is necessary to interact with any model (set parameters, define (initial) data input, define which data to include in the model output etc.), a graphical user interface is often nice to have for easier interaction. As CoeGSS is to provide decision support, and stakeholders will usually be experts of the field under consideration but generally non-experts in dealing with computational models, let alone HPC-GSS-SI systems, a graphical user interface, that is moreover easy to use, will be necessary as soon as the stakeholders shall be enabled to use the system themselves. It will have to be found out by means of demonstrators and close interaction with stakeholders, in how far this is feasible and desired.

A remark about the use of existing software tools – both for generating synthetic populations and for agent-based modelling frameworks – is in order: it has to be clarified how these tools can be used in terms of licensing.

An overview of existing software is given in deliverable D3.1; here, we would like only to point to a lesson learned from the ILUMASS (Integrated Land-Use Modelling and Transportation System Simulation) project.

Involving seven research institutions in Germany, ILUMASS aimed to develop a completely microscopic model of land use, transport and environmental impact in urban regions. The project failed to achieve its goal of simulating policy scenarios. The model consists of three separate modules for land-use, transport and the environment, which interact with one another. It was instantiated for the urban region of Dortmund, with a population of 2.6 million people. Agents that made up the land-use part of the model included individual members of the population, firms, and residential buildings. The transport model involved computing a weekly activity plan for the synthetic population living in the study area based on the socio-demographic data for each person. The environment module models the environmental impact of traffic forecasts generated in the transport module. The data for the model was managed with a Geographic Information System (GIS), and the study area was divided into a grid of 352000 cells of size 100m by 100m, and the coordinates of each cell were used as the locations of agents. Reasons for the failure of the project include underestimation of time required to set up Application Programming Interfaces required for the simulation; inefficient testing procedures (running long simulations on the whole study area only to recognise a small but critical error at the end); use of a sub-optimal, file-based inter-module data exchange system; and more generally, the extreme computing time requirements of some parts of the simulation (see http://www.spiekermann-wegener.de/pro/ilumass_e.htm, and references provided there, especially Wagner and Wegener, 2007). This underscores some of the points made above: testing procedures need to be accurately developed, and data-centric modelling as well as data inclusion in simulations are critical points.

3 Task 4.1: Health Habits

3.1 Short problem description

Many health conditions are caused by risk behaviours, such as problem drinking, substance use, smoking, overeating, or unprotected sexual intercourse (Schwarzer, 2008).

Monitoring, forecasting and controlling the spread of such behaviours in the general population represents a challenge for policy makers. Indeed, the future of healthcare in Europe will also depend on how much the demographic change and the increasing elderly population will see a corresponding increase in conditions associated to risk behaviours. Recent projections indicate that "if most of the future gains in life expectancy are spent in good health and free of disability, this could offset more than a half of the projected increases in spending due to an ageing population" (Centola, 2011).

From a theoretical and modelling perspective, the challenge of describing how individuals modify their habits has been tackled in the field of health psychology at the individual level, for example by Schwarzer (2008). However, how behavioural changes occur at the population level is not completely understood and it represents an active research field. Recent works on the subject, for example by Przywara (2010), have shown how social networks and interactions among them play a crucial role in the spread of health habits. Moreover, to describe the dynamics of behaviour changes one needs to take into account the notion of social contagion (Christakis and Fowler 2007).

The main goal of this pilot is to create a Synthetic Information system, integrating large and heterogeneous data sources, hat will allow to describe and model the spread of relevant health habits at the population level in Europe.

3.2 Initial example 1: smoking

Around 700.000 people die every year in Europe from tobacco-related diseases and about 13 million people suffer from smoking-related diseases in the EU, with a significant impact on the healthcare system and the economy of the continent. It has been estimated that the annual cost of tobacco to the European economy is of more than half a trillion euros, or about 4.6% of EU GDP (Jarvis 2012).

These figures put the tobacco epidemic at the top priority in a list of health related habits that have devastating consequences on the European society and the population's well-being. Most European countries have implemented strict regulations on tobacco consumption and such policies have been successful in reducing the prevalence of smokers in Europe. However, the declining trends of smokers in Europe have seen a slowdown in the past years with notable variations between countries.

Currently, the rate of smokers varies from the lowest in Sweden (11%) and Finland (19%) to the highest in Greece (38%) and Bulgaria (35%) (European Commission 2015).

Responding effectively to the tobacco epidemic requires understanding and modelling how smoking behaviour is transmitted. This makes the smoking habit an excellent target for the synthetic information system to be developed in T4.1. Mathematical models that describe smoking dynamics are available in the literature (Sharomi and Gumel 2008, Levy et al 2006). The synthetic information system will generate a synthetic population of agents, with interaction rules based on the current knowledge of smoking transmission mechanisms and their mathematical formulation. This will allow to monitor and forecast future trends in smoking prevalence across Europe.

3.3 Initial example 2: obesity

The worldwide prevalence of obesity nearly doubled between 1980 and 2008. According to country estimates for 2008, over 50% of both men and women in the WHO European Region were overweight, and roughly 23% of women and 20% of men were obese (WHO 2015).

As reported by the WHO (2015), based on the latest estimates in European Union countries, overweight affects 30-70% and obesity affects 10-30% of adults. Estimates of the number of overweight children in the WHO European Region increased from 1990 to 2008. Over 60% of children who are overweight before puberty will be overweight in early adulthood. Childhood obesity is strongly associated with risk factors for cardiovascular disease, type 2 diabetes, orthopaedic problems, mental disorders, underachievement in school and lower self-esteem. These statistics put obesity on top of the health conditions that can have serious consequence in the near future for the European healthcare system and are driven by detrimental behaviours, such as the lack of physical activity.

Understanding the mechanisms that are driving the spread of obesity is an important challenge for social science, psychology and medicine (Hill and Peters 1998). Current knowledge identifies a number of risk factors such as physical activity, diet, the individual social network, and environmental factors that concur to the global obesity prevalence (Przywara 2010, Hill and Peters 1998). However, the quest for introducing effective policies against the obesity epidemic remains an open public health challenge (Davey 2004).

Due to its relevance, the obesity epidemic represents an ideal test bed for the synthetic information system to be developed by CoeGSS. By integrating large scale demographic information, official population statistics, and models of social contagion, we aim at developing a large SIS to monitor and explore trends in the prevalence of obesity in Europe. Modelling will consent to explore different future epidemic scenarios and evaluate their impact on the European public health system.

3.4 Initial synthetic population and workflow

The synthetic population to be developed will be initially structured into agents representing the population of a single country, chosen among the following ones: United Kingdom, Germany, Italy, or France. The components will be extended to other European countries in further developments of the platform.

The basic elements of the synthetic population will be initially based on previous experience in the modelling of infectious diseases. More specifically, among others, one leading example will be the GLEAMviz simulator tool (Broeck et al 2011), which simulates the spread of emerging human-to-human infectious diseases across the world. The idea is to replicate some basic ingredients of GLEAMviz, after the necessary adaptation to the topic of interest, i.e. the diffusion of health habits.

In particular, the synthetic population will be based on a high-resolution population database that is at the core of the GLEAMviz tool: the Gridded Population of the World project by the Socio-Economic Data and Applications Center (SEDAC) (http://sedac.ciesin.columbia.edu/data/collection/gpw-v3) that estimates population with a granularity given by a lattice of cells covering the whole planet at a resolution of 15 × 15 minutes of arc.

The second step will be to integrate such population distribution with additional information on relevant health statistics, in order to reproduce a statistically correct representation of health conditions and health habits in the population under study. Following the initial examples, the synthetic population will reproduce the prevalence of smoking or obesity in the country of interest, by age groups and by region (depending on the granularity at which data on these health habits will be available).

As outlined in Section 4, agents will then be assigned to relevant social clusters, which will be defined depending on the contagion dynamics of interest. For instance, in the case of smoking, a possible aggregation of individuals will be in households, schools, workplaces and public spaces. In all of these settings, agents will experience different chances to be exposed to tobacco consumption and also to become a smoker.

Finally, models of interactions between agents, based on the notions of social contagion or complex contagion, will be developed to add a dynamic layer to the synthetic population. Such dynamic layer will describe possible scenarios of contagion processes, and will allow to assess the impact of future trends in the prevalence of health habits. The definition of such models will require an extensive research effort to identify relevant parameters, calibrate them on real data, and perform numerical simulations.

3.5 Data requirements

Data requirements will arise from the tasks described above. Here, we list a few data sources that may be relevant for defining the synthetic population of interest. The list is not intended to be comprehensive at this stage and further relevant sources can be added in the future:

- Population database of the Gridded Population of the World project by the Socio-Economic Data and Applications Center (SEDAC) available at:

  http://sedac.ciesin.columbia.edu/data/collection/gpw-v3

- Official country statistics related to health habits provided by Eurostat and available at: http://www.europeandataportal.eu/

3.6 Software requirements

Initial software requirements are based on agent based modelling of infectious disease dynamics. Models that are state-of-the-art in the literature are generally coded in C/C++ (Chao et al 2010, Merler and Ajelli 2010).

Each simulated individual is usually represented by a C/C++ structure that includes unique identifiers for the person and for each of the attributes that the individual carries, such as age, job status, health conditions (obese or not-obese), habits (smoker or not-smoker), and similar.

Dealing with large populations of up to a hundred millions individuals (Chao et al 2010) requires the parallelization of the code, typically done using OpenMPI (http://www.open-mpi.org/) or similar software.

In terms of performance, we expect to reach results that are in line with the epidemic simulation platforms. In the case of FluTE, a large scale agent-based simulation model for influenza epidemics (Chao et al 2010), simulating an epidemic covering the continental United States (population of 280 million) takes about 192 hours of total CPU time, on a high performance cluster with 32 CPUs. If the code is scalable, such numbers could be greatly improved with the availability of a large super-computing infrastructure.

3.7 Gathering experience with the CoeGSS infrastructure

As an initial test of the integration between a GSS simulation tool and the HPC infrastructure available in CoeGSS, we will try to deploy the GLEAMviz simulator at the high-performance computing centre HLRS. The GLEAMviz simulator is made of a server unit and a client application. This initial test should be focused on the adaptation of the server unit to the HPC infrastructure. It should fulfill a twofold purpose:

- The GLEAMviz simulator tool represents a type of Synthetic Information system with a similar architecture of the agent-based models to be developed in T4.1. The test should inform the development of the initial synthetic population outlined above, so that the availability of a large HPC infrastructure is put to the best possible use, for example via the communication and data model implemented.

- The test is a common task that starts the necessary interaction of HPC and GSS experts for this pilot.

In particular, goals of the initial test should be:

- to identify and resolve the intellectual property issues arising
- to define HPC software requirements for the SI system;
- to identify any code compatibility issue between the systems;
- to identify the best options for parallelization of the GLEAMviz server unit;
- to assess the gain of performance to be expected by running the SI system on a HPC.

Results of the test will be reported in the next version of the current deliverable.

4 Task 4.2: Green Growth

4.1 Short problem description

Climate change is an example of the kind of global challenge addressed by Global Systems Science: greenhouse gas (GHG) emissions produced by a large number of heterogeneous agents accumulate in the global atmosphere with long term impact on the climate system, without individual agents observing direct consequences of their emissions. For decades, reducing emissions has been perceived to come at an economic cost in terms of growth, jobs or welfare – as portrayed by the majority of economic assessments of climate change mitigation policy and summarized by the IPCC (2014). The idea of green growth, as promoted for example by the OECD (2011), contests the existence of a tradeoff between present economic development and avoiding future climate change: recent studies find that it is possible to increase wellbeing in the economic, ecologic, and social spheres simultaneously (e.g., the UNEP (2011) Green Economy Report). This pilot investigates the diffusion of green growth initiatives, such as policy measures (feed-in tariffs, building regulations), business strategies (brands for electric vehicles, developing apps for healthy habits), or lifestyle changes (switch to vegetarian diets or cycling) by means of a synthetic information system that represents large populations of heterogeneous individuals – contrasting the standard economic modelling practice of considering representative agents.

Emissions are distributed unevenly throughout the economy; the most relevant sectors include energy, land use change, and agriculture. Within the energy related sectors, electricity and heat, transport, and industry play important roles, as illustrated by Figure 1.

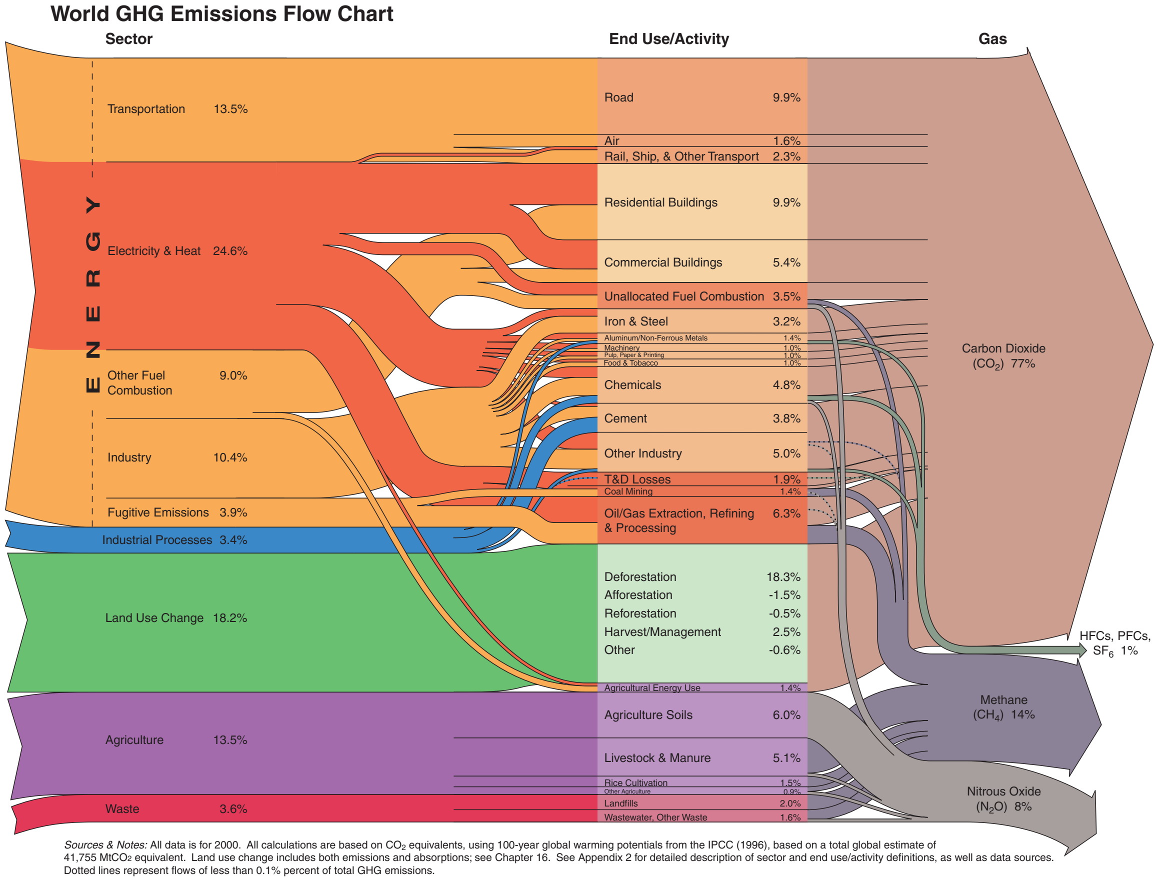

Figure 1 – Source: http://www.wri.org/sites/default/files/resources/world_greenhouse_gas_emissions_flowchart.pdf

Potential measures for reducing emissions differ substantially between and even within sectors, for example between measures for reducing the emission intensity of energy (meaning less emissions per unit of energy) and those for increasing energy efficiency (meaning that less energy is used). The green growth pilot study will consider the transport sector as the initial example to work on.

4.2 Initial example: transport

According to the IPCC, transport emissions account for around 23% of total energy-related CO2 emissions worldwide (Sims et al. 2014). Largely differing estimations exist about the development of transport emissions to come. For example, Creutzig et al. (2015) have recently stated that the total CO2 emissions of the transport sector could be halved by 2050, using policy incentives and existing technology alone, while referring to estimates that see transport emissions double over the same time frame. Also, in the EU, transport sector emissions have increased (by about 20%) between 1990 and 2011, while emissions from the other relevant sectors decreased (e.g., industry -34%, electricity and heat about -15%) over the same period (European Commission 2013).

Within the transport sector about 70% of emissions are due to road transportation (of which again 70% are due to passenger cars) (Eurostat 2015) – indicating that the development of the global fleet of cars plays an important role for tackling the challenge of reversing the trend of increasing emissions from transport. There are aggregate models about the development of the global car fleet (e.g., Dargay et al. 2007, and references therein), however, the CoeGSS green growth pilot's synthetic information system will allow to study this development at a new level of detail.

4.3 Initial synthetic population

To begin with, the green growth pilot will develop a global synthetic population for analysing scenarios of the development of emissions from cars. Given the global scope of the automotive industry, and the importance of the development of the car market in emerging economies, one would miss important parts of the picture if one restricted the analysis on cars driven in Europe.

Agents for the populations are people, grouped into households, who own cars. New cars bought and cars scrapped will influence the evolution of the "car population", activities and networks that influence these decisions are to be modelled. The idea is, in the first instance, to keep the model as simple as possible.

Relevant variables for households, as well as for cars, need to be specified, and will contain:

- People: basic demographic information like gender, age, etc
- Households: household size, age structure, income level, location
- Cars: age, fuel-type (petrol, diesel, hybrid-electric, battery-electric), fuel efficiency class (for conventional fuels), manufacturer

Further information about people or households is needed for specifying driving habits:

- Do people need to drive or do they have alternatives to driving? Proximity to workplace, public transport availability, ability to cycle, etc.
- Willingness to be environmentally friendly

Cars have to be distributed among households to match available statistics on numbers of cars per capita.

Networks between households have to be specified, however, it is not clear yet, which networks will be the most important ones. The question of who encounters whom may play a less important role than for disease transmission, while social networks constituted by peer groups may play a larger role. A literature search for mechanisms of "contagion" for consumer decisions will determine which networks to focus on first.

Environment characteristics that need to be taken into account will include the transportation infrastructure density (road, rail), the density of fuel stations, the network of charging stations for electrical cars and possibly others.

4.4 Preliminary, or training model

In order to start as simple as possible, a preliminary model is being built: a synthetic "population" of only cars of a few different types that have different emission intensities, initially, even just cars (i.e., a single type). These are distributed on a global map, on as fine a grid as possible, to gain familiarity with both the process of synthesising datasets (car types per country, car density per country, population density and thus cars distributed within countries) and with (computational representations of) the map underlying a population. Visualisation can begin at this basic a level and shall be undertaken, again for learning purposes.

Figure 2 illustrates a rough and preliminary picture of a global map of cars.

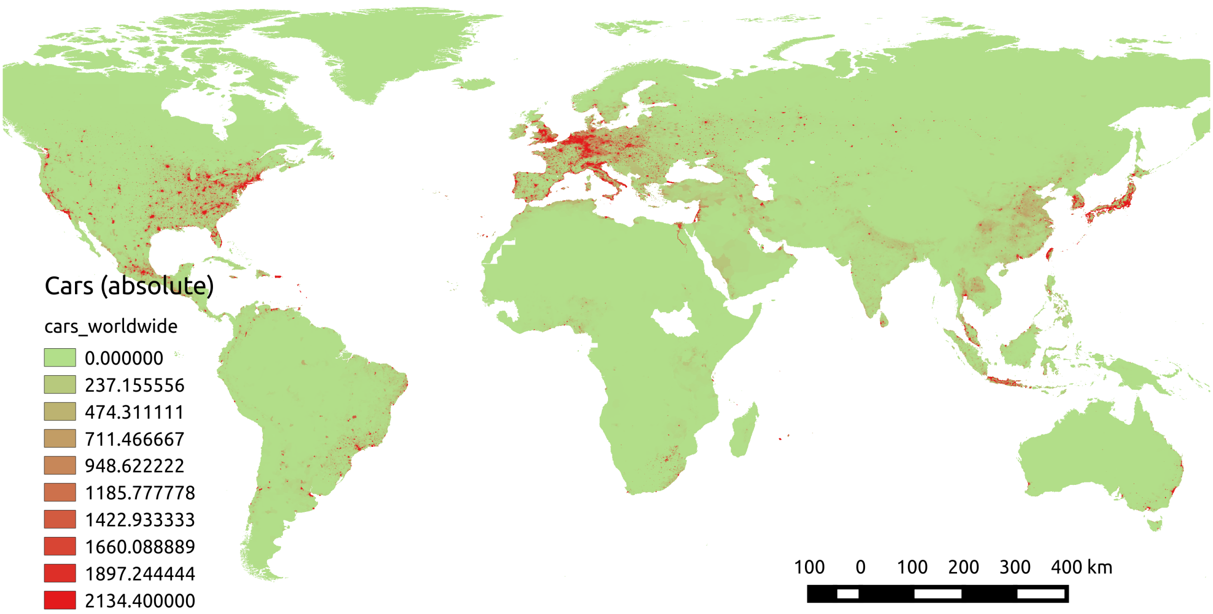

Figure 2 Rough sketch of the global car fleet

The ingredients to this picture are a gridded population data set and numbers of cars per 1000 inhabitants by country, with some pre-processing. The picture is preliminary in that (apart from the fact that a few countries are missing) numbers of cars used are not all taken in the same year, and there is, so far, no difference made between areas of different population densities within countries. That is, at this point the fact that people in densely populated areas may own relatively less cars, is disregarded. Further steps will be made to produce a more reliable map of the global car fleet.

A very simple comparison of this sketch of the global car population with two other maps, however, immediately displays the green growth challenge for the particular case of transport. Figure 3 plots the population data used to create the car map: while regions with a high density of cars concentrate in Europe, North America, Japan and few other places, the population map has various other areas of high density, particularly in China, India, and some areas of Africa. A very rough picture of GDP per capita (just by country, see Figure 4) shows the very different averaged economic possibilities of these people. Green growth is about increasing well-being in the economic, ecologic and social dimension. If the many people in so far poor regions had the possibility to afford cars and would buy cars, it is obvious that without structural change towards "green cars" it would be impossible to reverse the trend of increasing emissions from cars.

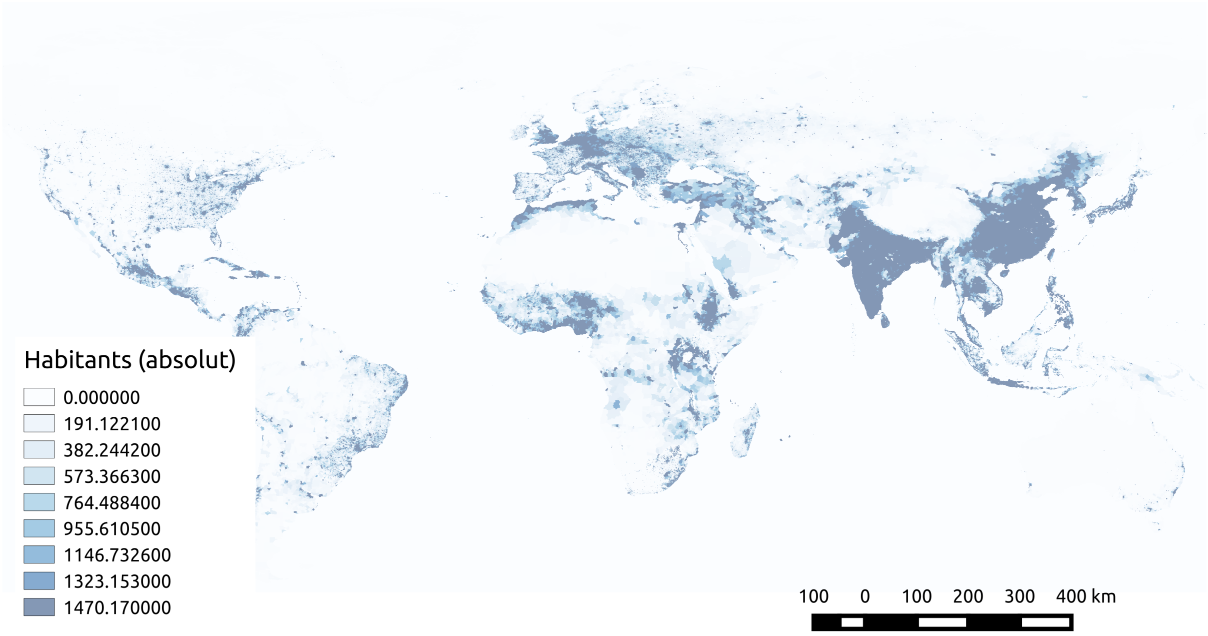

Figure 3 World population

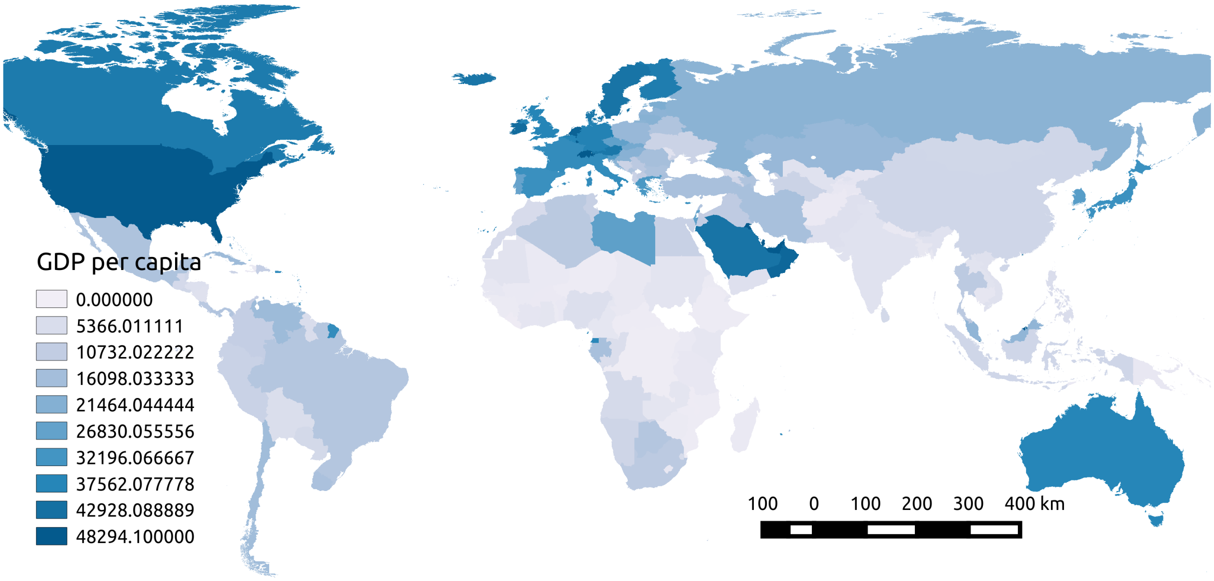

Figure 4 GDP per capita per country

The static picture presented in Figure 2 will be complemented with simple dynamics for each grid cell, computing the next number of cars from the given number, a trend, the neighbouring numbers, a saturation level that correlates with the number of the cell's inhabitants, and/or other characteristics. GCF's experience in agent-based modelling shows that it is very easy to build a model for which it is very hard to understand what it actually does. Thus, for the dynamics of the preliminary simulations, it is an explicit aim to start as simple as possible. "Contagion" dynamics for environmentally friendly car types driven by aggregate probability distributions will be a first step. As mentioned in Section 4, including HPC abilities as early as possible will depend on easy access on a testing environment (virtual or not).

Further steps will emerge from what is learnt while doing the first steps.

4.5 Data requirements

Data is needed for generating the synthetic agents of the initial synthetic population, determining the networks they interact in and the activity patterns they use, as well as for calibrating the model that describes the system's evolution. Exact data requirements will arise for each of these purposes while doing the modelling work. More generally, it can already be stated that data for all the above-mentioned characteristics of the initial synthetic population (from variables for people, households, and cars to environment characteristics) will be required.

Basic demographic data for people and census data for households will be an important building block of any synthetic population. If however, synthetic people can be used from some pre-existing synthetic population – for example a population developed by the CoeGSS pilots in common, which each pilot can then further refine – this type of data may not need to be collected. If it has to be collected, data availability is good for European countries, for example via the Eurostat databases.

Possible starting points for finding data on cars, car usage, and especially data for the dynamics of the car "population" include:

- Cars per 1000 people: https://en.wikipedia.org/wiki/

  List_of_countries_by_vehicles_per_capita
- cars by different types, for Europe: Eurostat (2015a)
- cars' emissions: studies such as ICCT (2014)
- fuel consumption, km driven etc.: depending on the availability of the data from the app owners, FuelLog (https://play.google.com/store/apps/details?id=ch.simonmorgenthaler.fuellog&hl=en) or similar apps might be sources of individual level data
- dynamical part of the model (when is a new car bought and which one?):
  - networks of influence between agents: insights from network research in social media (e.g., European Commission 2010, and references therein), or in computer games (e.g. Corominas-Murta and Thurner 2015) may be helpful
  - how concerned are people about the environment or climate: individual data collection, for example, via apps could be of interest, as well as surveys such as Carle (2015)

Beginning at these starting points, data requirements will crystalise in an iterative process of specifying synthetic agents together with a model, searching for the data needed according to this specification, and, if necessary, revising the specification. At some points, it will even make more sense to choose some elements of the specification only after having done some preliminary checks on data availability. For example, wanting to model a number of different types of cars with respect to their emission intensity, it seems to make most sense to choose this number with an eye on how many different types are distinguished in the datasets one can draw on.

Concerning the preliminary model, currently, geo data is gridded on a 3432x8640 raster and encoded as geotiff. For now, this includes also inactive cells (ocean, lakes). Thus, data points are only saved for active cells (about 9 Mio data points). Data that is not referenced to location is currently stored in csv-tables. As soon as the common data pool structure for all pilots (see Section 4) has been set up, this practice needs to be reviewed and probably revised accordingly.

4.6 Software requirements

As stated in Section 2, software tools for pre-processing data, for generating the synthetic agents and ABM software, possibly supported by an ABM or HPC framework, is needed for then running simulations of the model with the synthetic agents. Further, software tools for easily running the model with many sets of parameters are needed for sensitivity analysis and calibration.

For the preliminary model, data from different sources are pre-processed within Python that allows for fast prototyping. The package "pandas" allows for the efficient combination of multiple data sources as well as for the creation and handling of complex data structures. Pandas supports common data formats used in Big-Data-applications and thus supports easy transfer of data and interfacing to other software.

Spatial data is most efficiently handled using geo information systems (GIS). The library libgdal provides tools for spatial interpolation, raster vectorised data, geo-transformation and compression of geodata. It can be used within Python, qgis and Pandora (see later), thus is important as an interface between the different frameworks.

The initial synthetic population will most probably be generated with the help of the R-package simPop (https://cran.r-project.org/web/packages/simPop/index.html). The generator of the SimTRAVEL Research Initiative (http://urbanmodel.asu.edu/popgen/trainingmaterials.html) seemed interesting because of detailed training material provided, however, installation could not be completed due to outdated links, and there seems to be no further support for the software.

Concerning an ABM-HPC framework, different options are being evaluated, but it is likely that the model will be implemented in C++, using Pandora (http://xrubio.github.io/pandora/), an agent based modelling framework for HPC environments or the parallel programming system Charm++. Other options looked into were HPC-Repast and Flame, which however seemed to be less user-friendly concerning HPC. Also, Pandora, being less comprehensive, seems easier to completely understand in order to add extensions if necessary. GCF has in the meantime established contact with the team developing Pandora, based at the Barcelona Supercomputing Centre.

Pandora has the following external dependencies: Libtinyxml for xml support, libgdal1 for GIS support, and mpich2 for MPI support. As build tool, SCons is used, which is implemented as a Python script. Therefore, a Python installation with version 2.x with x >= 4 is required. HDF5 is used as data format.

As for software tools for multiple runs, sensitivity analysis and calibration, experience from the HPC-community within the project is required to help find those tools that are most useful to this pilot.

4.7 Visualisation requirements

Pandora comes with a GUI tool called Cassandra, which also allows the user to visualize the simulation. However, some first tests using a simulation with about 150.000 agents and 100 time steps (the epidemic example of Pandora) on a Dual Xeon E5-2630 Workstation with 32 GB RAM and an SSD based storage system created the impression that Cassandra is not optimized for big data sets. For example, Cassandra reads all the results at the beginning of a session, which causes an unnecessary limitation to the size of the data sets. Its single threaded implementation probably explains why it keeps unresponsive also afterwards.

While Cassandra could be useful to explore the behavior of the model using scaled down simulations, for the visualization of full-blown runs, another tool that allows the visualization of GIS data and time-series of statistical figures is needed. The tool should allow the user to easily compare multiple runs of the model, and would ideally allow to brush subsets of data points like e.g. ggobi (http://www.ggobi.org/). Again, HPC-expertise from the project partners is needed to identify such a tool.

5 Task 4.3 Global Urbanization

5.1 Short problem description

Optimizing city development choices is an essential challenge for the future, driven by the growth of world population increasingly living in cities and the opportunities of more overall and complete approaches assisted by always more global and intelligent technological innovations. It is consequently essential to evaluate possible options to choose the optimal ones (Hak et al. 2016, Hassan and Lee 2015, Hayek et al. 2015).

Cities are complexly defined (Allen 1997, Batty 2006 and 2013, Portugali 2000, Pumain 2000) by the interaction of processes as different as real estate, transportation, economy, society and politics. Insightful decisions prove therefore essential to determine city development from qualitative and quantitative points of view and its harmonious integration in its environment. While being an essential stake for the world of tomorrow, finding insightful decisions for the qualitative and quantitative complex city requires specific modelling and simulation tools.

Every feature can be modelled at various scales (Portugali and Benenson 1996, Shen 2002, Wissen et al. 2015, Krugman 1996, Weidlich 2000, Jordan et al. 2014, Filatova 2014, Lerman and Omer 2016), following the available data and the modelling purpose. For instance, transport offer and traffic can be simply reflected by static aggregate accessibility indicators (possibly simply based on existing data) when evaluating business real estate prices. It will however have to be finely modelled (Nagel and Paczuski 1996) to evaluate precise travel time.

Therefore, precisely a complex systems approach proves essential to conceptualize in a clearer way and predict more realistically the complexity of city evolution, based on synthetic populations, and for the simulation of which high performance computation will prove precious. Indeed, modelling city and population dynamics rests on individual's multi-fold characteristics, particularly here geographic (concerning housing or transportation), which will influence their evolution. Therefore, the possibility of creating sets of synthetic populations with realistically statistically distributed characteristics' values will prove precious for any simulation. It can permit furthermore to clarify influences between very different elements of the city and point to possible or more efficient levers to improve cities' everyday life. For instance, it will allow exploring the impact of development choices, the precise two-way relation between price mechanisms and infrastructure decisions, but it might also open the way to study the effect of overall opinion and behaviour dynamics and assess the benefit of information campaigns or other public incentives.

5.2 Workflow quick overview

A city model shall be developed using a progressive and cyclic implementation. Our purpose is to implement with the CoSMo platform city models including various sub models and allowing for various scales, starting with the most aggregate (or coarse) and refining them, following available data and observed simulation time.

The purpose of this incremental approach is three-fold:

- to allow (if appropriate) to propose more complete modelling solutions at different scales
  - choose the most appropriate following the needs
  - particularly to optimize precision and simulation time
- optimize interactions with other workflows
  - get first results quickly
  - facilitate consequently first stage coordination with
    - synthetic populations
    - and HPC, HPDA
- help assess over the comparison of the results at different scales, the benefit of both
  - complex system approach with greater detail
  - and therefore of linked HPC/HPDA.

Synthetic populations will be implemented for use within the model, awaiting possible links to those of the project (WP3), either by importing provided initial state, or if possible, libraries and providing at least an interface with the rest of the pilot.

To the ends of model and synthetic population development, it will be necessary to find data as detailed as possible: indeed, if coarse grained data (overall travelling mode preferences for instance) is often free and easy to find, more detailed and specific data can become more difficultly accessible (precisely localized travelling mode preferences).

5.3 City pilot and synthetic populations' details

The synthetic human population should describe demographic features (age, gender, family status (single, married, children)), education level, profession, income, housing, transport (transport mode preferences and travel needs, particularly commuting), economic budget for housing, transport, and other features of agents.

Non-human synthetic "populations" (i.e. organized data sets) might concern transport (infrastructures (roads, etc), vehicles (cars, buses), transport offer (bus lines), and so on), real estate offer (buildings, dwelling units and the like).

The global urbanisation pilot is to integrate consequently a set of sub-models:

- **Real estate model**, allowing to evaluate the evolution of real estate offer and price particularly based on
  - Initial offer and pricing
  - Location demand
- **Transport model,** allowing to calculate (multi-modal) traffic flows and travel times based on
  - Travel demand
  - Travellers' modal preferences
  - Transport infrastructure and transport service offers
- **Pollution model**, allowing to calculate
  - peak emissions based on
    - vehicle fleet
      - for every vehicle in this fleet: vehicles' technical characteristics, particularly concerning emissions following speed
  - pollutants dispersal based on
    - meteorological data
      - wind
      - temperature, …
    - (city topography if available?)
- **Population model** allowing to evaluate behaviours (particularly concerning housing and transport) based on
  - Initial behaviours
  - Evolution rules if accessible?
- **Economic model** allowing to evaluate the evolution of goods, particularly real estate and transport prices based on
  - Initial prices
  - Offer
  - Demand
    - Depending on profiles' (or agents') preferences and constraints

These interdependent sub-models are to be linked in the global urbanisation pilot. Dependencies can be of different kinds:

- Implicitly or explicitly over static data
- Explicitly over dynamic coupling: this can lead to parametrized or various kinds of coupling.

Hereafter we list a first set of possible synthetic populations and elements of the model.

5.3.1 Transportation model

5.3.1.1 *Purpose*

The purpose of the transportation model is to evaluate city multi-modal transportation flows in a city following existing infrastructure and needs. It should therefore allow to assess effective benefits such as accessibility of various parts of the city, but also drawbacks such as noise and pollution: therefore, this model should also allow, once linked to a pollution model, to evaluate GHG emissions.

5.3.1.2 *Model elements*

Multimodal transportation network comprising

1. Physical network
2. Transport service offer network
3. Transport operators
4. Travellers
5. Vehicles

1. ***Physical network*** should provide a description of the physical transport network for the different transportation modes (road, bus lines, tramway, subway, etc.)
    a. Roads (linked to intersections)
        i. Identification of extreme nodes
        ii. Length
        iii. Possible transportation modes
        iv. Maximum number of vehicles
        v. Recommended speed
    b. Intersections
        i. Location
        ii. Possible modes
        iii. (Linked roads)

    Similarly to the two previous points

    c. Railways (linked to rail junctions)
    d. Rail junctions
2. ***Transport service offer***
    a. Travel line sections
    b. Stops (in a travel line)
    c. Connections (between two travel lines)
3. ***Transport operators***
    a. Company or owner name
    b. Transport mode

   c. Fleet size
   d. List of transport service lines, each one defined by
      i. Type of vehicle (see hereafter)
      ii. Itinerary
      iii. Effective runs
         1. Frequency or
         2. Schedule
   e. (Linked to economic model: pricing policies)

4. ***Travellers***
   a. List of travel demands (particularly commuting) every one defined by
      i. Travel origin / house
      ii. Travel destination / office, school or university
   b. Transportation mode possibilities
      i. Owning or not a private vehicle
      ii. Public transport subscription
      iii. (Linked to economic model: Transport budget)
   c. Transportation mode preferences
   d. (Linked to economic model: transport budget)

5. ***Vehicles***
   a. Vehicle
      i. Type of vehicle
      ii. Maximum speed
      iii. Maximum number of passengers
      iv. (Linked to pollution model: pollutants emission)
      v. (Linked to economic model
         1. Investment cost
         2. Maintenance cost
         3. Operating cost)

5.3.2 Pollution model

5.3.2.1 Purpose

The purpose of this model is to evaluate pollution (emission and dispersal) linked to transport, particularly GHG emissions. It depends consequently on the traffic calculated by the transport model.

5.3.2.2 Model elements

- For pollutant emissions
  - (From the transport model: vehicle travels)
  - For every traveling vehicle, characteristics defining
    - Emission of various pollutants (COPERT) following

        - Vehicle model
        - Speed
        - (Road slope)
- For pollutant dispersal
    - Meteorological data
        - Wind
        - Temperature, …
    - (City topography if available?)

5.3.3 Real estate model

5.3.3.1 Purpose

The purpose of this model is to provide the real estate availability and pricing in the city and its evolution. It is linked to the economic and population models. It is also linked to the transport model since firstly the transport demand depends on the location of housing and businesses and secondly accessibility (defined by the transport infrastructure) influences the real estate value and price.

5.3.3.2 Model elements

Real-estate model defines different parts

1. Building
2. Dwelling unit
3. Inhabitant

1. Building

a. Geographical position

b. Quantitative characteristics
    i. Total surface
    ii. Number of dwelling units

c. Qualitative characteristics
    i. Accessibility (linked to transport model)
    ii. Environment
        1. Level of noise (linked to transport model)
        2. Level of pollution (linked to pollution model)
        3. …
    iii. Neighbouring facilities
        1. For housing, for instance:
            a. Schools
            b. Museums, libraries
            c. Retail stores
        2. For businesses, for instance:

a. Number of neighbouring restaurants
b. Proximity of railway station or airport

2. ***Dwelling unit***
   a. (Building)
   b. Purpose
      i. Nightly occupancy
         1. Housing
      ii. Daily occupancy
         1. Businesses
         2. Education (pre-school day care, schools, high schools, universities, …)
         3. Administration / State
         4. Retail stores
         5. Parks
         6. Sports
         7. Transportation linked places
         8. …
   c. Quantitative characteristics
      i. Surface
      ii. Number of units (rooms)
      iii. Linked to economic model: Price (unit price / total price)
   d. Qualitative characteristics
3. ***Inhabitant***
   a. List of linked dwelling units with a
      i. Location
      ii. Occupancy period
   b. Housing budget (linked to the economic model)

5.3.4 Population model

5.3.4.1 Purpose

The population model is central in the city dynamics, interacting with the economic, housing and transport models.

5.3.4.2 Model elements

1. Association
2. Household
3. Individual

Specialized particularly into

4. Inhabitant (see Real estate model)
5. Traveller (see Transport model)

6. Consumer (see Economic model)
7. Interaction network

1. ***Association***
   a. List of households or individuals
2. ***Household***
   a. List of individuals
   b. (Linked to real estate model: residence)
   c. (Linked to economic model: budget particularly concerning housing, transport)
3. ***Individual***
   a. Gender
   b. Age
   c. Daily and nightly occupation and daily location ('inhabitant', linked to real estate model)
   d. Transport preferences ('traveller', linked to transport model)
   e. (Linked to economic model: personal budget, particularly concerning transport)

   Further details

   f. Level of education
   g. Profession
   h. Income

7. ***Interaction network***
   a. Type of network from different points of view
      i. Semantic (ex. information, economy, travel)
      ii. Operational (ex. flows of public news or private messages, of money, of vehicles)
   b. List of nodes (associations, households or individuals)
   c. List of single or bi-directional links between pairs of nodes

5.3.5 Economic model

5.3.5.1 Purpose

The purpose of the economic model is not only to allow assessing economically various simulation scenarios but to integrate economy-based dynamics in the city model.

5.3.5.2 Model elements

1. Market
2. Good

3. Economic agent

Specialized particularly into

4. Investor
5. Producer
6. Retailer
7. Consumer

1. ***Market***
   a. Good
   b. List of economic agents
2. ***Good (for instance real estate or vehicle)***
   a. Investment price
      i. Basic production value
      ii. Additional market value (following offer and demand)
      iii. Taxes
   b. Rental or leasing price (if applicable)
   c. Operating cost (per unit of use) (if applicable)
   d. Maintenance cost (per period) (if applicable)
   e. Refurbishment cost (if applicable)
3. ***Economic agent*** (for instance city council, retail stores, businesses, real estate investors, …)
   a. Budget
   b. List of targeted goods with preference criteria

5.3.6 Synthetic population refinements

Refining basic synthetic populations for use with the above described models, shared requirements or perspectives for future applications arise in the fields of population complexity, network complexity, and spatial analytics, as listed in the following.

5.3.6.1 Population complexity

Characteristics interdependence:

- Purpose is to generate synthetic population with realistic distributions not only of single characteristics but taken as a set, together i.e. reflecting correlations.
- Scientific stakes
  - Simulate real populations more precisely
  - Master biases
  - Clarify result analysis
  - Improve predictions

  - Consequently improve and contribute to better put into light the benefit of synthetic populations and precise simulations coming with HPC / HPDA
- Possible future applications in this pilot
  - Population characteristics
    - Household size is linked to age.
    - Income is linked to education, profession and experience.
    - Daily occupation and location is linked to age and profession (particularly concerning length of studies)
  - Housing description
    - Price is linked to intrinsic value but also location and neighbouring facilities.

Hierarchised (multi-scale) characterisation:

- Purpose: allow for a multiple level population characterization
  - For instance socio-professional categories at different grains:
    - Primary / Secondary / tertiary sectors
      - For tertiary can split into merchant or non-merchant services
        - Non merchant services can split into cultural offer (public libraries), education (public offer), health, …
        - Merchant services can split into education (private offer), transport, housing, retail stores, cultural offer (bookshops, …), miscellaneous proximity services and others…
- Scientific stakes
  - Allow to study the influence of the data refinement on the overall dynamics
      - Consequently the importance (or not) of
        - Precise characterization data
        - Required HPC/HPDA for the corresponding calculations
- Possible future applications in this pilot
  - Population characteristics
  - Housing description
  - Transportation details

5.3.6.2 Network complexity

Multiplex networks:

- Various purposes can be foreseen:
  - Allow to generate different kinds of (random) networks for instance

            - Various degrees of connectivity distribution
                - Degree preserving randomization?
            - Various tendencies to cluster, in various numbers of groups and with various clustering degrees
            - Scale free networks?
                - Small-world networks
    - If possible: specifying different possible characteristics of the connections
            - Reciprocity (possibility to specify asymmetrical relationship (A influences B but not the contrary))
            - Open question concerning the relevance of transitivity (the friend of my friend will become my friend) might be handled at the model level rather than by generating specific vertex.
            - Propinquity (tendency to link with persons geographically close)
                - This might vary also following the location
                    - Quantitatively: airports foster less new links between persons than neighbouring residence
                    - Qualitatively: neighbourhood relationships differ from the ones between colleagues
            - Homophily (tendency to bound with similar profiles vs dissimilar ones)
                - This might influence the propagation of new behaviours across different cultural and social groups by limiting its effect to a priori characterization
            - Multiplexity (differentiate different kinds of links)
                - For instance carpooling might occur with colleagues rather than friends, the latter being more liable to influence week-end green behaviours (preferring the bike for instance)
- Scientific stakes
    - Allow to study the influence of the network details on the overall dynamics
            - Consequently the importance (or not) of
            - Precise network data
            - Required HPC/HPDA for the corresponding calculations
- Possible future applications in this pilot
    - Refined social network (for information and behaviour propagation)
    - Associating individuals to different places, point of interest, activities…
    - Refined travel calculation

5.3.6.3 Spatial analytics

Simulating population spatially:

- Purpose: generate spatial distributions of people or activities.
- Scientific stakes
  - Allow to simulate spatially located synthetic populations
  - Individuals' locations are here not independent but interdependent since globally constrained by limited shared space (for instance for housing, every dwelling can hold one and only one household, while integrating further economic constraints since depending on household income and price gradient).
  - Provide various spatial distributions to more thoroughly study the influence of initial spatial distribution on the overall dynamics
    - Consequently the importance (or not) of
      - Precise network data
      - Required HPC/HPDA for the corresponding calculations
- Possible future applications in this pilot
  - Simulate activity locations (education, business, residential) in the city
    - For instance to study the influence of the concentration of retail stores or spread of businesses on
      - Economic activity
      - Pollution
  - Simulate spatially located population
    - For example, to study the influence of population initial location on real estate price evolution?

5.4 Data requirements

Corresponding data is required first on one, then on a few other European cities if possible (the first targeted city is Paris), as complete and detailed as possible.

The main challenge for this pilot is to get precise geographically localized data, and data which reflect distributions' correlations following the different characteristics (for instance, the housing is linked to the income, itself linked to the age, education and profession). Additionally, to be able to simulate the city at different scales, it would be interesting to get data at different levels of detail (for instance profession can be split into primary, secondary or tertiary sector, but also detailed down to specific professions such as researcher, or any intermediate level; some data providers offer different levels of aggregation but aggregate versions can also be calculated from detailed data sets, with however the difficulty of keeping realistic correlations between the characteristics' values.

Data is needed for the features of the human and non-human populations mentioned above, i.e. demographic data (age, gender, family status (single, married, children)), education level, profession, income, housing, transport data, respectively data on transport infrastructures, real estate offer and so on.

Typical study cities will count a few hundreds of thousands to millions inhabitants. However due to the complexity of the topic, simulations will count equivalent number of dwelling units, vehicles, economic calculations, etc…

Additionally, calculation challenges might be raised by algorithmic complexity (for instance shortest path finding for the transport model).

5.5 Software requirements

This model is to be designed and implemented with the CoSMo modelling software in version 4.1.0 or higher (supported platforms are Winodws 8.1 64 bits, Debian 7 64 bits, transition to Debian 8 underway, Ubuntu 14.04 LTS 64 bits, OS X 10.10 64 bits), leading to a C++ simulation engine runnable with HPC / HPDA facilities. CoSMo provides a simulator API with which to write batch simulation scripts in C++ (or Python) allowing to define further simulations based on first simulation results (for instance to find optimal parameter values with an optimization algorithm launching various simulations), assuming HPC allows for dynamic definition and allocation of new tasks by a launched process. Depending on HPC requirements and possibilities, further tools such as Charm++, MPI or specifically written scripts might be required to distribute simulations and centralize their results.

Concerning the link to independent synthetic populations the CoSMo software allows to load automatically initialization values (for instance provided as text files or from databases (SQLite and Spatialite supported); furthermore, it can integrate synthetic populations as addins, if these are implemented with C++ and provided for instance as a library. This requires minor ad hoc C++ implementation for the integration in the existing model, defining interactions with the existing classes.

Expected value of HPC for this pilot is especially to help launch a great number of different simulations allowing to explore model results following different simulation inputs (defined by parameters values) and eventually optimize them. Indeed, proving how more complete and finer grained models succeed in simulating reality more appropriately not only from a quantitative, but also a qualitative point of view, will help prove the necessity of such complex models, requiring HPC.

Visualization concerns in the first place two dimensional maps of cities, however if possible unfolding different features (population, traffic, prices, pollution, etc.). Visualization challenges concern also more generally finding insights concerning the study of model dynamics, reflected by many indicators (spatially localized prices, traffic flows, pollution, time localized budgets and expenses, and more) and depending on many parameters (particularly the initial state of the model concerning prices, pricing policy, population characteristics, transport offer, and the like).

6 Task 4.4 Future Applications

This task is not in itself a pilot, however, identifying needs and opportunities for the future of HPC applications to complex global, anthropocentric systems, it can point out directions for future requirements to the centre of excellence. At the state of writing this deliverable, pilots are taking up speed, first interaction between GSS, HPC, and computer science project partners has supported coming up with some common use of technical terms and has prepared the grounds for first GSS test runs on the HPC systems. The focus for starting CoeGSS has necessarily been on the "current applications" – the pilots – rather than on possible future applications. Nevertheless, one topic, for the study of which a symbiosis of HPC and GSS could prove valuable, has been identified: financial stability. In fact, "financial contagion" and its consequences could be investigated with the help of a synthetic information system including financial actors as well as firms in the real economy, to better understand the links between the financial and the real economy, and to explore the implications of different sets of policies aiming at stabilising the financial system. As illustrated for example by Battiston et al (2012), systemic risk in the financial system depends on the network of financial exposures among institutions. This network, determining how after some shock financial distress amplifies or can be contained, can be represented in a synthetic information system to analyse the transmission of shocks between the real and the financial sector.

Relevant data about financial and economic enterprises, such as equity, the network arising from an agent having another agent's assets on its balance sheet, and others is often not publicly available; thus the fact that synthetic information systems do not compromise privacy makes these an interesting tool. In fact, a specific task for CoeGSS might arise due to privacy reasons: building a synthetic information system "dummy" according to a given specification to then hand over to stakeholders (say, a central bank) who would then themselves have to fill it with the data that cannot be disclosed.

With IMT as a project partner, the CoeGSS consortium includes expertise and contact networks that may make this an interesting future application.

7 Conclusion

Initial pilot requirements to a Centre of Excellence for Global Systems Science are collected in this document. These center around HPC-based synthetic information systems for GSS applications, which the pilots are developing. The word "initial" can be interpreted in two ways here: some requirements are important for starting the work at the intersection between HPC and GSS, and may become less important, or may simply be resolved throughout the project duration. In another sense, the requirements are "only initial" in the sense that they will have to be adapted and revised throughout the project.

In the initial phase of this endeavour, the overarching requirement is that of creating a synergetic link between the two communities involved, for example by fostering the emergence of a common language shared by GSS and HPC experts. This document can in itself underline the necessity of this: being written by GSS partners of CoeGSS, the ideas about HPC expressed in this document may not all make perfect sense to the HPC partners, while some of the more descriptive GSS focused texts may not be clear to the latter. By inviting scientific exchange, this document (among others) may support the interaction between the two communities.

On a more technical side, initial requirements (e.g. in terms of software or data) have been collected as a starting point. Some general requirements for building HPC-based synthetic information systems are shared by the three pilots. However, as the pilots consider different global challenges, and as they start from rather different points of departure, specific requirements differ between them.

Project partners working on the health habits pilot already have a synthetic information system geared to analysing epidemics. They can base the work on an existing population, adapting the model to what is needed for analysing health habits.

Project partners working on the global urbanisation pilot have a sophisticated set of models, so they need to focus on a synthetic population that is adapted to these models.

In the green growth pilot, project partners can draw on experience with agent-based modelling, however, both the synthetic population and the model for running simulations are developed simultaneously.

8 References

8.1 Task 4.1

Broeck, Wouter V., et al. "The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale." BMC Infectious Diseases 11.1 (2011): 37.

Centola, Damon. "An experimental study of homophily in the adoption of health behavior." Science 334.6060 (2011): 1269-1272.

Chao, Dennis L., et al. "FluTE, a publicly available stochastic influenza epidemic simulation model." PLoS Comput Biol 6.1 (2010): e1000656.

Christakis, Nicholas A., and James H. Fowler. "The spread of obesity in a large social network over 32 years." New England Journal of Medicine 357.4 (2007): 370-379.

Davey, R. C. "The obesity epidemic: too much food for thought?" British Journal of Sports Medicine 38.3 (2004): 360-363.

European Commission Press Release: http://europa.eu/rapid/press-release_IP-15-5028_en.htm

Hill, James O., and John C. Peters. "Environmental contributions to the obesity epidemic." Science 280.5368 (1998): 1371-1374.

Jarvis, Andrew. "A study on liability and the Heath costs of smoking." DG SANCO (2008/C6/046) Updated Final Report. GHK, University of Exeter and the Public Health Advocacy Institute (2012).

Levy, David T., Joseph E. Bauer, and Hye-ryeon Lee. "Simulation modeling and tobacco control: creating more robust public health policies." American Journal of Public Health 96.3 (2006): 494.

Merler, Stefano, and Marco Ajelli. "The role of population heterogeneity and human mobility in the spread of pandemic influenza." Proceedings of the Royal Society of London B: Biological Sciences 277.1681 (2010): 557-565.

Przywara, Bartosz. Projecting future health care expenditure at European level: drivers, methodology and main results. No. 417. Directorate General Economic and Monetary Affairs (DG ECFIN), European Commission, 2010.

Schwarzer, Ralf. "Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors." Applied Psychology 57.1 (2008): 1-29.

Sharomi, O., and A. B. Gumel. "Curtailing smoking dynamics: a mathematical modeling approach." Applied Mathematics and Computation 195.2 (2008): 475-499.

WHO Euro Region URL: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/obesity/data-and-statistics. Visited December 2015.

8.2 Task 4.2

Carle, J. (2015) Climate Change Seen as Top Global Threat. http://www.pewglobal.org/2015/07/14/ climate-change-seen-as-top-global-threat/?beta=true&utm_expid=53098246-2.Lly4CFSVQG2lphsg-KopIg.1

Corominas-Murtra, B., S. Thurner (2015) The weak core and the structure of elites in social multiplex networks. *Interconnected Networks*

Creutzig, F., P. Jochem, O.Y. Edelenbosch, L. Mattauch, D.P. van Vuuren, D. McCollum, J. Minx (2015) Transport: A roadblock to climate change mitigation? *Science* (2015), 20: 911-912.

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

EEA (2013) Annual European Union greenhouse gas inventory 1990-2011 and inventory report 2013. European Environmental Agency.

European Commission (2013): EU Transport in Figures. Statistical Pocketbook 2013, http://ec.europa.eu/transport/facts-fundings/statistics/pocketbook-2013_en.htm

Eurostat (2015) Passenger cars, by motor energy. http://ec.europa.eu/eurostat/en/web/products-datasets/-/ROAD_EQS_CARMOT

ICCT (2014) CO2 emissions from new passenger cars in the EU: Car manufacturers' performance in 2013. ICCT Briefing. http://www.theicct.org/co2-new-cars-eu-manufacturer-performance-2014

IPCC (2014). Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014. http://mitigation2014.org/report/publication/.

Jaeger, C.C. (2012) Scarcity and coordination in the global commons. In: C.C. Jaeger, K. Hasselmann, G. Leipold, D. Mangalagiu, and J. D. Tábara (eds.), Reframing the Problem of Climate Change: From Zero Sum Game to Win-Win Solutions pp. 85–101. Routledge.

OECD (2011) Towards green growth. OECD http://www.oecd.org/greengrowth/48224539.pdf.

Sims R., R. Schaeffer, F. Creutzig, X. Cruz-Núñez, M. D'Agosto, D. Dimitriu, M.J. Figueroa Meza, L. Fulton, S. Kobayashi, O. Lah, A. McKinnon, P. Newman, M. Ouyang, J.J. Schauer, D. Sperling, and G. Tiwari (2014): Transport. In: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA

European Commission (2010) Social Networks Overview: Current Trends and Research Challenges. Luxembourg: Publications Office of the European Union

Wagner, P., M. Wegener (2007) Urban land use, transport and environment models: experiences with an integrated mircroscopic approach. *disP* 170 (3): 45-56.

8.3 Task 4.3

Allen, P.M. (1997) Cities and regions as self-organizing systems: modls of complexity. Taylor and Francis, London.

Batty, M. (2006) Cities and complexity. Understanding cities with cellular automata, agent-based models, and fractals. The MIT Press.

Batty, M. (2013) The new science of cities. The MIT Press.

Filatova, T. (2014) Empirical agent-based land market: Integrating adaptive economic behavior in urban land-use models.

Hák T. et al. (2016) Sustainable Development Goals: A need for relevant indicators, Ecological Indicators 60 (2016) 565–573

Hassan, A.M., Lee, H. (2015) Toward the sustainable development of urban areas: An overview of global trends in trials and policies, Land Use Policy 48 (2015) 199–212

Hayek, U W, Efthymioub, D Farooqc, B. Wirthd,T. von Teiche, M. Neuenschwandera, N. Grêt-Regameya, A. (2015) Quality of urban patterns: Spatially explicit evidence for multiple scales. Landscape and Urban Planning 142 (2015) 47–62

Jordan, R., Birkin, M., Evans, A. (2014) An agent-based model of residential mobility, Assessing the impacts of urban regeneration policy in the EASEL district, Computers, Environment and Urban Systems 48 (2014) 49–63

Krugman, P R (1996) The self-organizing economy. Blackwell, Cambridge, Mass.

Lerman, Y., Omer, I. (2016) Urban area types and spatial distribution of pedestrians: Lessons from Tel Aviv, Computers, Environment and Urban Systems 55 (2016) 11–23

Nagel, K and Paczuski, M (1995). Emergent traffic jams. Physical review E, 51, 2909-2918.

Portugali, J (2000) Self organization and the city. Springer-Verlag, Berlin.

Portugali, J and Benenson, I (1996) Human agents between local and global forces in a self organizing city. In F Schweitzer (ed) Self organization of complex structures: from individual to collective dynamics. Gordon and Breach, London, 537-545

Pumain, D (2000) An evolutionary model of urban systems. In I Ianos, D Pumain

Shen, G (2002) Fractal dimension and fractal growth of urbanized areas. International journal of geographic information science, 16, 419-437

Weidlich, W (2000). Sociodynamics: a systemic approach to mathematical modelling in the social sciences.Harwood Academic Publishers, Amsterdam.

Wissen Hayek, U., Efthymiou, D., Farooq, B., von Wirth, T., Teich, M., Neuenschwander, N., Gret-Regamey, A. (2015) Quality of urban patterns: Spatially explicit evidence for multiple scales, Landscape and Urban Planning 142 (2015) 47–62

8.4 Task 4.4

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012). “DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk” Scientific Reports, 2, 541, doi:10.1038/srep00541.