EasyChair Preprint

№ 6468

Data Management in EpiGraph COVID-19 Epidemic Simulator

Miguel Guzmán-Merino, Christian Durán,
Maria-Cristina Marinescu, Concepción Delgado-Sanz,
Diana Gomez-Barroso, Jesus Carretero and David E. Singh

August 29, 2021

Data management in EpiGraph COVID-19 epidemic simulator*

Miguel Guzmán-Merino[1], Christian Durán[1], Maria-Cristina Marinescu[2], Concepción Delgado-Sanz[3,4], Diana Gomez-Barroso[3,4], Jesus Carretero[1], and David E. Singh[1]

[1] Department Computer Science, Universidad Carlos III de Madrid, Leganés, Spain.
[2] Barcelona Supercomputing Center, Barcelona, Spain.
[3] CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[4] National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain

Abstract. The transmission of COVID-19 through a population depends on many factors which model, incorporate, and integrate many heterogeneous data sources. The work we describe in this paper focuses on the data management aspect of EpiGraph, a scalable agent-based virus-propagation simulator. We describe the data acquisition and preprocessing tasks that are necessary to map the data to the different models implemented in EpiGraph in a way that is efficient and comprehensible. We also report on post-processing, analysis, and visualization of the outputs, tasks that are fundamental to make the simulation results useful for the final users. Our simulator captures complex interactions between social processes, virus characteristics, travel patterns, climate, vaccination, and non-pharmaceutical interventions. We end by demonstrating the entire pipeline with one evaluation for Spain for the third COVID wave starting on December 27th of 2020.

Keywords: Epidemiological simulation · COVID-19 · Heterogeneous data processing · Parallel tool

1 Introduction

The transmission of the COVID-19 virus through a population depends on many factors that reflect the makeup of the community, the characteristics and behaviours of the individuals, as well as the effect of the measures taken to curb its propagation. The larger the community, the more difficult it becomes to predict the outcomes. To tackle this problem, we have implemented EpiGraph, a scalable, parallel agent-based simulator. This paper centres on data management,

* This work has been supported by the Spanish Instituto de Salud Carlos III under the project grant 2020/00183/001, the project grant BCV-2021-1-0011, of the Spanish Supercomputing Network (RES) and the European Union's Horizon 2020 JTI-EuroHPC research and innovation program under grant agreement No 956748. We would like to thank to Diego Fernandez Olombrada for his support in the early collection of part of the data of this work.

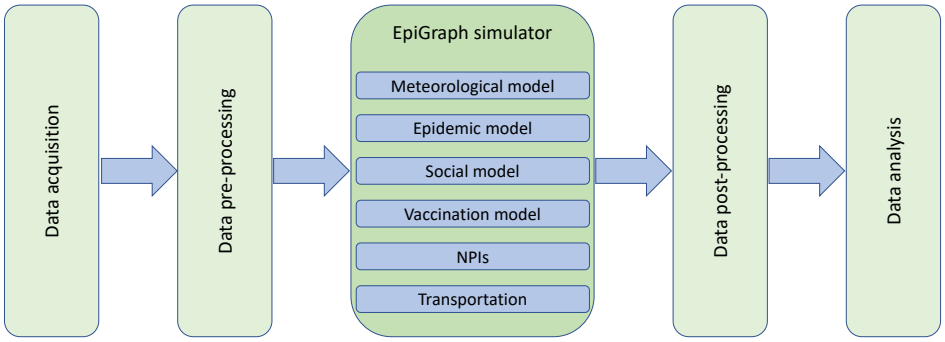

Fig. 1: Stages involved in EpiGraph simulations.

which turns out to be quite complex, given that EpiGraph implements several modules that reproduce the different aspects which have an impact on the virus propagation. The interplay of these modules simulates a complex phenomenon that takes many heterogeneous sources and data types as input by mapping them to the different parameters of the agent model. The aim of this work is two-fold, on one hand we aim to contribute to a better understanding on the modelling for the epidemic simulation to coronavirus pandemic, and on the other, we describe methodologies for an efficient integration of heterogeneous data.

Figure 1 shows the different stages involved in EpiGraph simulation. The input data is first obtained from multiple sources ranging from research papers, to public and private databases. These data are highly heterogeneous and have to be processed in a second stage using multiple technologies. Section 2 describes in detail these two stages.

The third stage of the figure corresponds to the simulation process. EpiGraph is an agent-based simulator that includes multiple models that realistically reproduce the environment where the infection spreads. The Meteorological model uses meteorological data to increase or reduce the disease's R0s values based on each particular weather condition. The Epidemic model is a compartmental stochastic extended SEIR model. Rather than the more common analytic models based on differential equations, Epigraph computes, for each infected individual, the duration of the different compartments and the transition probabilities.

The social model reproduces the social habits of four different main group types: students, workers, stay-at-home people, and elders. A group can represents a certain number of individuals that interact during work hours - for instance, groups are the students belonging to the same classroom, workers of the same company. Note that an interaction involves the co-location in time at a distance that is small enough to make disease transmission possible. EpiGraph model interaction during work hours, school-time, family time, and leisure, including multiple professions. See [24, 16] for additional details. The risk of infection, given

by the specific R_0 value of the infected individual, also depends on two factors that reduce the transmission risk: the vaccination of the susceptible individual that is in contact with the infected one, and the use of non-pharmaceutical intervention (NPIs), like the use of face masks. These factors are modelled by the Vaccination and NPI models included in the simulator. Finally, the transportation model computes the number of travellers between the urban areas that are being simulated based on the geographical distance between them.

There is a large amount of output data produced during the simulation (we have simulated up to 200M individuals and 650 cities). In order to provide a comprehensive analysis of the results, a post processing stage (fourth stage) is carried out. Then, the fifth stage uses this information to generate different statistical data that summarize the simulation output and graphically displaying the results. Section 2.3 describes more in detail the analysis and visualization stage.

2 Epigraph data management

This section provides a description of the data used and produced by the simulator, as well as how these data are processed. Figure 2 shows the different data sources involved in a simulation and how they interact with the different software components. Epigraph consists of two main software pieces that are used in combination with several auxiliary programs. The first component is the Scenario generation in which the different urban areas used in the simulation are created. Note that these urban areas only contain information about the characteristics of the individual in the population and the way they interact with other individuals. The input data sources (upper part in the figure) are geolocation provided by web applications that are used to identify the geographic coordinates of each city, its related NUTS code, as well as the distances between each pair of cities. The second data source are the Eurostat, and Spanish-equivalent INE, that provide the demographic data used by the simulator. This information, depicted in Section 2.1, includes among other, the population pyramid and the distribution of employment related to each city. Two different social-network graphs are used for generating the contact patterns of each individual. Finally, contact matrices, extracted from public surveys, are used to provide statistical information of the average number of contacts between individuals of certain age ranges.

The Epigraph generator is an MPI program written in C that creates, for each urban area, the characteristics (age, profession, etc.) of all individuals belonging to the same urban area as well as the related contact patterns[5]. We call this *social fabric*, and we store it as sparse matrices, in which each node is an individual and each edge is a time-dependent interaction with other person.

The social fabric created in the scenario generation stage is used as input in the scenario simulation (lower part of Figure 2). Note that this information can be reused among multiple simulations when the social fabric (i.e. the simulated

[5] Note the EpiGraph employs static and dynamic contact patterns, and in this section we are referring to the static one.

urban areas) do not change. Regarding the data sources shown in the figure, the COVID-19 model parameters were taken from the existing literature. The non-pharmaceutical interventions (NPIs) applied in each region and the coronavirus incidence were processed using Excel and Bash scripts. The vaccination data is a combination of the different parameters used in vaccine efficacy models and data for the existing doses administrated in each region. This information was taken from the existing literature and government databases, respectively. Finally, the meteorological data consists of a collection of meteorological measurements (pressure, temperature, etc.) of each urban area that was processed using Matlab.

All the previous data is used by EpiGraph to perform the scenario simulation. The simulator output is a collection of trace files with the state of each individual for each simulated time step in each urban area. Note that this information is very rich in contents, because it includes, in combination with the individual characteristics (health, age, occupation, etc.), the actions taken or applied to the person (vaccination, use of NPIs, travel, etc) for each time step. The following sections provide details about the data sources involved in the simulations and how these data were processed.

2.1 Scenario generation

Geographical data. EpiGraph simulations comprises one or multiple urban areas, that are identified by names, coordinates and NUTS codes. The first two parameters are used by the transport model for calculating the distance between the cities, i.e., the number of individuals in a population that travel between the cities. The latter parameters is used to carry out database search.

- The city's latitude and longitude were obtained from online databases, based on these coordinates, and using Google Maps services. By means of these services it was possible to obtain the distances of each city with the other ones -which is needed by EpiGraph's transport model-.
- NUTS codes represent a three-levels division of the European territory [14]. We use the happyGISCO tool [9] (which is an interface to Eurostat Gisco web services) in combination with Python scripts to obtain the city NUTS codes from their coordinates. The red arrow in Figure 2 highlights that the city NUTS codes are used by shell scripts for selecting the related demographic data of each city.

Demographic data. This set of data defines the characteristics of the simulated population. For Spain, the data was taken from the Spanish National Statistics Institute (INE) [19] with an aggregation level of province. For the rest of European countries, the source of data was EuroStat [10] with an aggregation level of country.

Demographic data include the following attributes related to the social activity: percentage of the people for the collectives of students, elderly people,

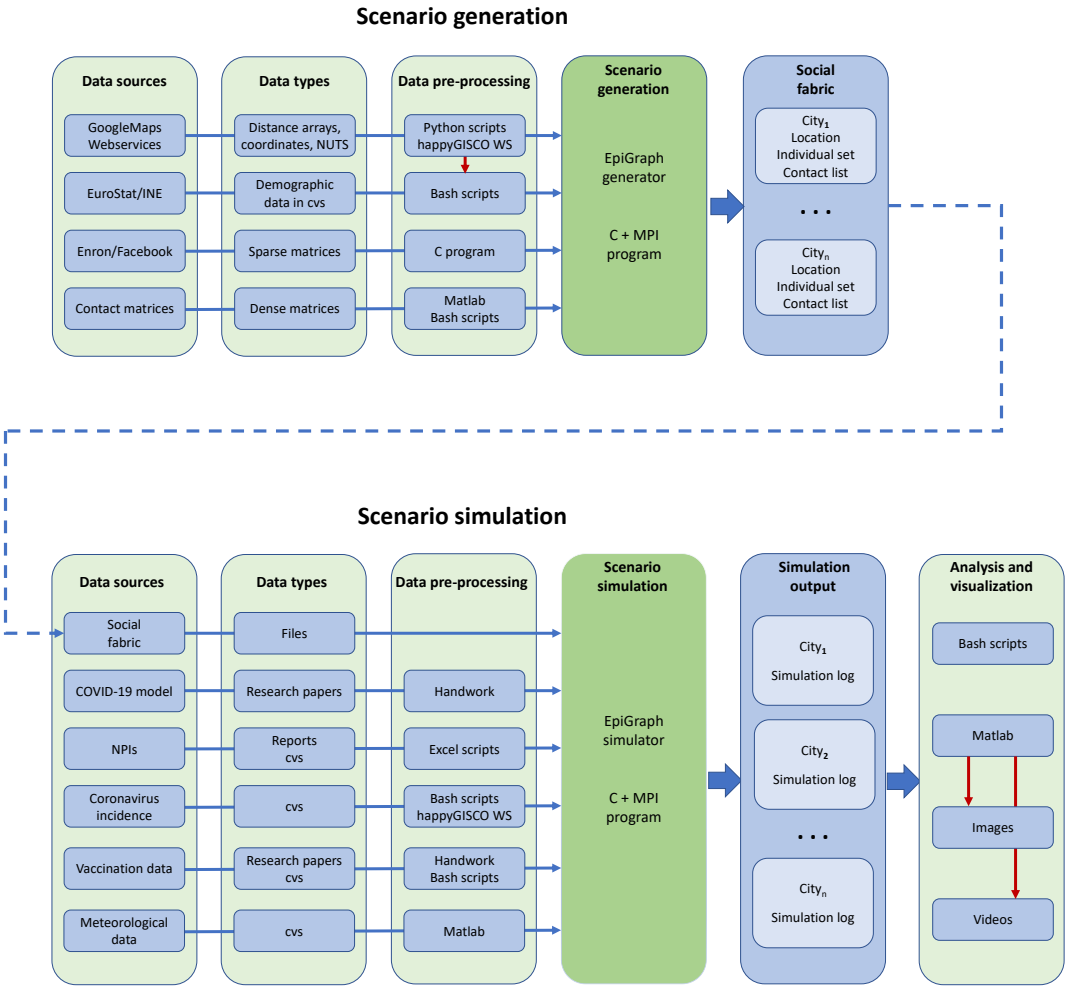

Fig. 2: Overview of the data flow related to EpiGraph simulation. In the scenario generation stage (upper part figure) the the social fabric is generated and stored in files. In a second stage this information is used in combination with other sources of data (lower part figure) to perform the simulation.

workers and unemployed; Regarding the elderly people we distinguish the sub-collectives that live in nursing homes or at home. The worker collectives are broken down by the following sub-collectives: industry, building, catering, services, security forces, education, front-line-health, non-front-line-health, social-health and transport. Note that each collective and profession has a different social

pattern. The household size is percentage of homes with one to five members that is used to model the family contacts.

Demographic data also includes: the population and population pyramid related to each simulated city; The percentage of essential workers; and the collective group sizes that includes a normal distribution comprising the minimum and maximum number of people involved in the same collective. For instance, the number of students in the same classroom or the number of workers in a company[6].

Population-mixing data. This data is used to generate the social pattern, i.e. the social interactions, between the simulated individuals. In epidemic simulations, population mixing is a crucial factor that determines the realism and accuracy of the simulations. The sources of information for the social model are described below:

- Social network graphs: we have employed the Enron Email Corpus graph [7] (70,578 nodes and 312,620 edges) for generating the work, elderly and informal meetup contacts, and a Facebook graph [2] (250,000 edges and 3,239,137 edges) for the school contacts. We have developed in [16] a variation of the Random Walk algorithm [17] which generates scaled sub-graphs from the Enron and Facebook data-sets, with an specific average connectivity $< k >$ provided as input argument. This value of connectivity is obtained from the contract matrices that are explained next.
- Contact matrices [6] are dense matrices in which each element $A_{i,j}$ represents average number of daily interactions between individuals of ages i and j. The contact matrix repository includes data for various countries and some regions of these countries, including sub-contact matrices for school, work, and community contacts. These contact matrices were processed using Matlab and Bash scripts. The school contact matrix was used to generate the student's school contacts (using the Facebook graph), with a number of interactions per age specified by the matrix. In a similar way, the work contact matrix was used to define the age distribution of the work, stay-at-home and elders interactions. Finally, the community contact matrix was used to generate the connectivity related to the leisure contacts.
- Daily Contacts of health professional with patients: according to the Spanish National Health System, on average a health professional is in contact with 30 patients per day on average. Due to the lack of data, we have used this value for all the simulated countries.
- Average contacts of health professional with COVID-19 patients. We estimate this value considering, on one hand, 180,000 health professionals in Spain. On the other, given a 10% of SARS-CoV-2 prevalence among the Spanish population at the end of the second wave, and the number of suspects per day, based on the number of diagnostic tests performed (1M per

[6] Note that each collective and sub-collective has different group sized based on the activity that they perform.

day, scaled by 1.3 in order to consider the tests not performed on suspects) will result in an average of 4.2 contacts with suspects per day, and a estimation of 0.4 contacts with COVID-19 patients per day, per health professional. According to [21] the hazard ratio for front-line health professionals is 3.3, so we assume that front-line health professional have a larger number of about 1.4 contacts with COVID-19 patients per day. Data are stored in configuration files formatted in xml.
- Catering contacts per hour. In our experiments we consider three levels of catering contacts: pre-pandemic, pandemic with a more reduced number of contacts per hour, and lockdown with catering services closed. Related data are stored in configuration files formatted in xml.

2.2 Scenario simulation

SARS-CoV-2 infection data. These data were extracted from research papers. They include the basic reproduction numbers (R0s) related to each disease stage, the state transition probabilities (for instance, the probability of an infected individual of being asymptomatic), the hospitalized and death probabilities, and the duration of each infection stage. Please refer to [24] for a detailed description of these parameters.

Non-pharmaceutical intervention data. This category includes different sources of heterogeneous data that record the NPIs imposed by each country during the pandemic.

- Social distancing policies consider three distancing measures collected from the Data on country response measures to COVID-19 database [8]: the closure of schools, the closure of public spaces of any kind, and the workplace closure. In this work we use the existing social distance measures for each European country in a simulation starting on December 27th of 2020.
- Face mask use. EpiGraph models the use of both surgical and ffp2-grade face masks, with different efficacies [24]. The results provided in this work are related to the simulation period at the beginnings of 2021 in which the entire European population was using mask outside the family circle.
- Sampling strategies [16] are modelled by the number of daily tests performed, the minimum time between two consecutive tests carried out to the same individual, the quarantine time, and the percentage of quarantine breakers, i.e. the fraction of people who do not comply with social distancing during quarantine time. These data was provided by the Spanish Ministry of Health.

COVID-19 incidence. We use the ECDC's weekly sub-national 14-day notification rate of new COVID-19 cases [8] for setting the initial percentage of infected population in each urban area (this value is only used at the beginning of the simulation). ECDC database provides sub-regional data for European countries, so the cities are set with regional values -instead of average country

values-. These data are automatically loaded using bash scripts that leverage the city's NUTS code to identify the incidence region values in the ECDC database.

We obtain the seroprevalence information related to each country from [20]. This information, which is uploaded using bash scripts, is only needed for setting the initial conditions at the beginning of the simulation.

Vaccination data. [23] presents the EpiGraph's vaccination model in detail. It includes both the vaccine effectiveness model, that depends on the individual age and the SARS-CoV-2 variant, and the vaccination strategy that is simulated, that defines aspects such as prioritization among target groups, and the time between the administration of the doses. The vaccination model was obtained from research papers and the vaccination strategy was provided by the Spanish Ministry of Health[7]. Please see [23] for further details about this model.

Meteorological data. In the current development of the simulator the meteorological data is provided by the Spanish Meteorological Agency (AEMET) [1] and is only used for Spanish-level simulations. The input data consists of cvs files with 10-minute samples taken during one year by all the meteorological stations in Spain. Data was processed by Matlab using interpolation algorithms (for estimating lost values) and collecting and processing the desired meteorological parameters of temperature, pressure and humidity for each city in Spain [22].

2.3 Analysis and visualization

EpiGraph generates simulation traces for each urban area, that contain, for each time step, the number of individuals each state of the infection and additional information as the number of vaccinated individuals (for each vaccine type and time when the doses have been administrated), use of masks, number of quarantined and hospitalized individuals, use of other NPI interventions. This information is processed in parallel and is combined with the population demographic and social data in order to generate both global and collective-specific statistics. For example, it is possible to obtain for a certain urban area, how many catering workers are infected or how many of the infected ones are vaccinated. In addition, we employ Matlab for providing overall statistics and graphical display of the results by means of images and videos.

3 Evaluation

In this section we provide simulation results for a national scenario related to Spain. The simulations were executed on the Tirant supercomputer, which is made up of 336 nodes each with two Intel Xeon processors Sandy Bridge E5-2670 and 32 GB RAM, interconnected with an Infiniband 40 Gbps network. In this section we provide simulation results for Spain. We simulate the third wave starting on December 27th of 2020.

[7] Note that the vaccination prioritization strategy is similar for all European countries.

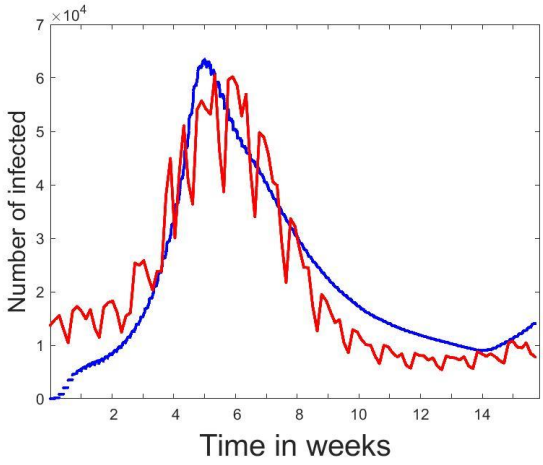

Fig. 3: Spain third wave: daily real (in red colour) and simulated (in blue colour) data related to the number of infections of the COVID-19. Simulation starts on December 27th of 2020.

This scenario simulates a population of 19,574,086 individuals related to the 63 most populated cities of Spain, using 109 processes. The simulation starts on December 27th of 2020, which was the starting date for the COVID-19 vaccination campaign. It reproduces the Spanish COVID-19 vaccination campaign and includes a given number of daily tests of 0.25% over the simulated population, and a percentage of positive tests of around 9% (which corresponds to the real testing rate and detection efficacy). Figure 3 shows in red colour the daily real reported cases for Spain and in blue colour the aggregated simulated cases for the considered cities. Real cases have been scaled by a factor of 1.42 in order to include the non-reported cases. Simulated cases represent the median of 10 different simulations.

A more detailed results of the simulation output is shown in Figure 4(left) in which the results are broken down by Spanish provinces. We can observe that both the real cases (in red) and the simulated ones (in blue) are similar although there are some differences for some of them. Note the high complexity of the problem that we are tackling. Figure 4(left) shows the geographic location of each one of the urban areas.

Note that this baseline simulation scenario is being used to evaluate different alternative scenarios. Examples of them are the scenarios where we introduce changes in the vaccination strategy (for instance, introducing changes in the prioritization process). Other interesting question is to assess alternative NPIs. For instance, evaluate school closing (instead of being opened, as has happened

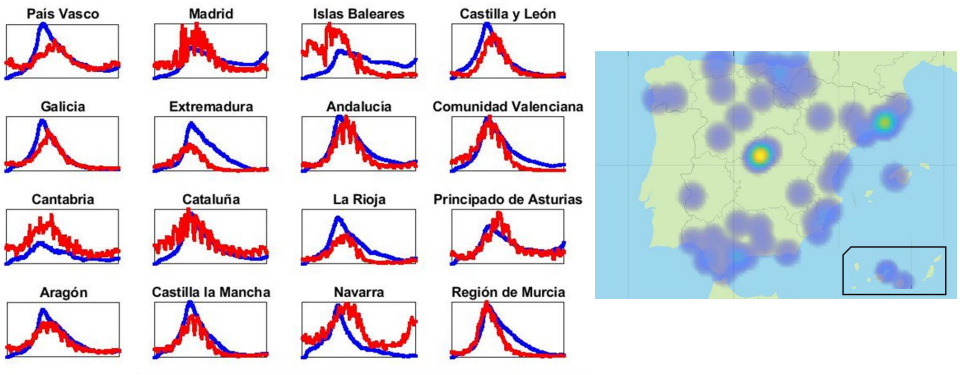

Fig. 4: (Left) Simulations results of the Spanish third wave, broken down by communities. Red line represents the real infected and the blue ones the simulated. Simulation starts on December 27th of 2020. (Right) Infected dispersion map of the 64 simulated cities.

in Spain), reducing the activity in the catering sector, lifting up the imposition of using face masks in open spaces, etc.

4 Related work

There are many approaches to model the COVID-19 propagation. A starting approach is the SEIR model based on solving the differential equations like in [4]. More complex versions of the SEIR model include, for instance, a quarantine class and a class of isolated (hospitalized) members [3, 18]. The main limitation of this approach is the lack of details in the simulation. An alternative way of modelling the infection spread are the models based on machine learning [15]. The work in [11], developed in the Imperial College of London introduces an extension of a semi-mechanistic Bayesian hierarchical model that infers the impact of interventions and estimates the number of infections over time. In [12], the authors use the discrete renewal equation as a latent process for the modelling of infections and propose a generative mechanism to connect infections to death data. They use this joint Bayesian hierarchical model to produce short-term predictions, and they apply their model to 11 different countries.

The European Centre for Disease Prevention and Control (ECDC) [13] has built a Monte-Carlo based model of COVID-19 that they use for forecasting. To model the behaviour of the people and how well they are responding to the measures, they compare the predictions with Google data about mobile phone use and they use the daily confirmed COVID-19 cases and daily deaths to calibrate it. It is interesting to note that some models perform forecast, like

COFFEE model from Los Alamos National Laboratory [5], and other are also capable of performing projections. A projection involves simulating alternative hypothetical scenarios. In the case of EpiGraph, this tool belongs to the models that perform projection.

5 Conclusion

This work describes the data management of EpiGraph, an agent-based simulator for influenza and COVID-19 propagation. The approach followed in EpiGraph is to combine several models that reproduce the different aspects of the environment where the disease spreads. This involves the simulation of complex phenomena that are modelled by employing different, complex and heterogeneous data sources. This work provides a description of the data management involved in EpiGraph's simulations including both the input data acquisition and pre-processing, and the output data post-processing and analysis. In our simulation framework, the use of Python and Bash scripts allows to quickly gather and perform simple processing (as data filtering or, specific data gathering) from many heterogeneous sources. Alternatively, Matlab was used for computing more complex tasks. This tool permits using advanced processing algorithms, because of the large number of toolboxes that includes, although its performance (considered as execution time) is low. We used Matlab for tasks that are usually performed once (like meteorological data interpolation) or for data visualization where data that was already processed. C programming language was employed for complex, computational-intensive tasks, like the social contact generation (that involves processing Enron/Facebook graphs) and the simulation processes. Given the particular characteristics of EpiGraph, it was possible to divide this data-intensive processing into different tasks, related to different sections of urban areas. In this way, several processing scripts can be simultaneously executed in order to speed-up the pre-processing and post-processing stages. For both the scenario generation and the simulation, MPI was used to execute the C program in parallel on multiple compute nodes.

References

1. AEMET. Agencia estatal de meteorología (aemet). https://www.aemet.es, 2021.
2. Sheena Batra. Facebook data. https://www.kaggle.com/sheenabatra/facebook-data, 2018.
3. F Brauer and C Castillo-Chavez. *Mathematical Models in Population Biology and Epidemiology*. Springer, 2012.
4. José M. Carcione, Juan E. Santos, Claudio Bagaini, and Jing Ba. A simulation of a covid-19 epidemic based on a deterministic seir model. *Frontiers in Public Health*, 2020.
5. Lauren Castro1, Geoffrey Fairchild, Isaac Michaud, and Dave Osthus. *COFFEE: COVID-19 Forecasts using Fast Evaluations and Estimation*. Los Alamos National Laboratory, LA-UR-20-28630, 2020.

6. Matteo Chinazzi and Dina Mistry. Mixing patterns. https://github.com/mobs-lab/mixing-patterns, 2021.
7. Will Cukierski. The enron email dataset. https://www.kaggle.com/wcukierski/enron-email-dataset, 2015.
8. European Centre for Disease Prevention and Control(ECDC). Home-ecde. https://www.ecdc.europa.eu/en, 2021.
9. European Commission (EC - DG ESTAT). Happygisco python interface to gisco web-services. https://happygisco.readthedocs.io/en/latest/, 2021.
10. Eurostat. Home-eurostat. https://ec.europa.eu/eurostat, 2021.
11. Seth Flaxman and et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584:257–261, 2020.
12. Seth Flaxman and et al. Swapnil Mishra. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: technical description update. *arXiv:2004.11342*, 2020.
13. European Centre for Disease Prevention and Control. Updated projections of COVID-19 in the EU/EEA and the UK. Technical report, ECDC: Stockholm, 2020.
14. Grazzini J., Museux J.-M. and Hahn M. Eurostat. https://ec.europa.eu/eurostat/web/nuts/background, 2021.
15. Youyang Gu. COVID-19 projections using machine learning. https://covid19-projections.com, 2020.
16. Miguel Guzmán-Merino, Durán, Maria-Cristina Marinescu, Christian , Concepción Delgado-Sanz, Diana Gomez-Barroso, Jesus Carretero, and David E. Singh. Assessing population-sampling strategies for reducing the covid-19 incidence. *Preprint*, 2021.
17. LOVASZ L. Random walks on graphs: A survey. *BOLYAI SOCIETY MATHEMATICAL STUDIES*, 2:46, 1993.
18. Michael Lingzhi Li, Hamza Tazi Bouardi, Omar Skali Lami, Nikolaos Trichakis, Thomas Trikalinos, Mohammad Fazel Zarandi, and Dimitris Bertsimas. *Overview of DELPHI Model V3 - COVIDAnalytics*, 2020.
19. Ministry of Economic Affairs and Digital Transformation (MINECO), http://www.ine.es. *National Statistics Institute (INE)*, 2021.
20. Ali Rostami, Mahdi Sepidarkish, Mariska Leeflang, Seyed Mohammad Riahi, Malihe Nourollahpour Shiadeh, Sahar Esfandyari, Ali H Mokdad, Peter J. Hotez, and Robin B. Gasser. First "snap-shot" meta-analysis to estimate the prevalence of serum antibodies to sars-cov-2 in humans. *Clinical Microbiology and Infection*, 27(3):331–340, 2021.
21. Anoop S V Shah and et al. Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study. *BMJ*, 371, 2020.
22. David E. Singh, Maria-Cristina Marinescu, Jesus Carretero, Concepcion Delgado-Sanz, Diana Gomez-Barroso, and Amparo Larrauri. Evaluating the impact of the weather conditions on the influenza propagation. *BMC Infectious Diseases*, 20:265, 2020.
23. David E. Singh, Maria-Cristina Marinescu, Miguel Guzmán-Merino, Christian Durán, Concepción Delgado-Sanz, Diana Gomez-Barroso, and Jesus Carretero. Evaluation of vaccination strategies for spain. *Preprint*, 2021.
24. David E. Singh, Maria-Cristina Marinescu, Miguel Guzmán-Merino, Christian Durán, Concepción Delgado-Sanz, Diana Gomez-Barroso, and Jesus Carretero. Simulation of covid-19 propagation scenarios in the madrid metropolitan area. *Frontiers in Public Health*, 9:172, 2021.