Open data and environment simulation : environmental and social simulation on distributed process systems based on irregular cell space

Thanh Ngoan Trieu

COLLEGE DOCTORAL BRETAGNE — MATHSTIC BRETAGNE OCEANE

Université de Bretagne Occidentale

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ
DE BRETAGNE OCCIDENTALE

ÉCOLE DOCTORALE N° 644
Mathématiques et Sciences et Technologies
de l'Information et de la Communication en Bretagne Océane
Spécialité : *Informatique et Architectures numériques*

Par
TRIEU Thanh Ngoan

Open Data and Environment Simulation

Environmental and Social Simulation on Distributed Process Systems Based on Irregular Cell Space

Thèse présentée et soutenue à Brest, le 17 September 2024
Unité de recherche : Lab-STICC UMR CNRS 6285

Rapporteurs avant soutenance :

Christophe CLARAMUNT — Professeur des Universités, Ecole Navale, Lanvéoc
Congduc PHAM — Professeur des Universités, Université de Pau et des Pays de l'Adour, Pau

Composition du Jury :

Président :	Philippe SALIOU	Professeur des Universités, CHU-Université de Bretagne Occidentale
Examinateurs :	Anne MOLCARD	Professeur des Universités, Université de Toulon
	Philippe SALIOU	Professeur des Universités, CHU-Université de Bretagne Occidentale
	Christophe CLARAMUNT	Professeur des Universités, Ecole Navale, Lanvéoc
	Congduc PHAM	Professeur des Universités, Université de Pau et des Pays de l'Adour, Pau
Dir. de thèse :	Vincent RODIN	Professeur des Universités, Université de Bretagne Occidentale
Co-dir. de thèse :	Bernard POTTIER	Professeur Emérite, Université de Bretagne Occidentale

Invité :

Hiep Xuan HUYNH — Professeur - HDR, Université de Can Tho, Vietnam (Co-dir. de thèse)

Acknowledgement

I acknowledge the Brest Métropole and Can Tho University for providing me the funding to work on my thesis. This support gives me a good chance to continue to pursue my education path that has had a great impact on my career.

I would like to express my sincere gratitude to my supervisors, Prof. Vincent Rodin and Prof. Bernard Pottier from Université de Bretagne Occidentale and Prof. Hiep Xuan Huynh from Can Tho University. Their invaluable advice and great supports inspired me to make the right decisions at most of the important moments. I am glad to have a chance to work with them. I would like to thank Prof. Christophe Claramunt and Prof. Congduc Pham, who accepted to review my thesis document. Thank you for their constructive feedback and essential suggestions that enhanced the quality of my work. I am thankful to the members of my thesis jury, Prof. Philippe Saliou and Prof. Anne Molcard, for their precious time to be part of my thesis jury.

I greatly appreciate Madame Magali Gouez and staff officers at École Doctorale SICMA MathSTIC and Lab-STICC CNRS, UMR 6285, for supporting my activities during the study.

I am thankful to my colleagues at College of Information and Communication Technology, Can Tho University, who encouraged me to achieve complete investigations. Many thanks to all my friends and Ph.D. companions for their discussions and fruitful exchanges.

Last but certainly not least, I am extremely thankful to my family for their love and unconditional support throughout the years. Thank to my husband, Vo Chi Tam, who has and will accompany me throughout this life. And of course I cannot forget my little daughter. Your presence is also a precious gift that I receive.

Table of Contents

1 Introduction **9**

1.1 Environmental and Social Problems 9

1.1.1 Environmental Problems 10

1.1.2 Social Problems: Pandemics 10

1.2 Modeling and Simulation Complex Systems with Availability of Government Open Data 12

1.2.1 Complex Systems 12

1.2.2 Modeling and Simulation with CA 13

1.2.3 Emergence of Government Open Data 14

1.3 Motivation and Objectives 15

1.3.1 Motivation 15

1.3.2 Objectives 15

1.4 Thesis Outline 16

2 Open Data **19**

2.1 Data and Open Data 19

2.1.1 What is Open Data? 19

2.1.2 Why Open Data is Essential? 20

2.2 Geographical Data 21

2.2.1 Administrative Divisions 25

2.2.2 Neighborhood Computation 27

2.3 Meteorological Data 29

2.3.1 BUFR 30

2.3.2 Express Modeling Language 32

2.4 Open Data in Analysis and Simulation 34

3 Geographic Divisions Modeled as Distributed Process Systems **41**

3.1 Cellular Automata 41

3.1.1 CA Model Description 42

3.1.2 Related Works on Regular and Irregular Cell Spaces 44
3.2 Irregular Cell Space with Geographic Divisions 47
3.2.1 Process System Generation 47
3.2.2 Data Binding 49
3.3 Parallel Computation with Irregular Cell Space 52
3.3.1 Occam 53
3.3.2 CUDA 55
3.3.3 MPI and Multithreads 57

4 Epidemic Spreading Control on Geographic Divisions 61
4.1 Pandemics 61
4.1.1 Metrics For Epidemiology 61
4.1.2 Spreading Paradigm 63
4.2 From Parameters to Transition Rules: The Case of Covid-19 65
4.2.1 Selecting Parameters 66
4.2.2 Transition Rules 71
4.3 Experiments 75
4.3.1 Simulation Results 75
4.3.2 Variation of Behaviors will Assist Control 76
4.3.3 Epidemic Spreading Control 77
4.3.4 Discussion 77
4.4 Related Works 82
4.4.1 Mathematical Models 82
4.4.2 CA Models 85

5 Monitoring Shores: Hybrid System of Regular and Irregular Cell Spaces 87
5.1 Green Algae Issues 87
5.2 Modeling Pollution on Shores 89
5.2.1 Modeling Ground Activities 89
5.2.2 Ocean Activities Measurements and Modeling 91
5.3 Hybrid System: Regular in Combination with Irregular Cell Space 99
5.3.1 System Generation 99
5.3.2 Water Circulation with Currents 100
5.3.3 Tracking Virtual Markers 101
5.3.4 Monitoring Pollution Coverage Area 103

6 Conclusion and Perspectives **109**
6.1 Conclusion . . . 109
6.2 Perspectives . . . 111

Abbreviations **113**

References **115**

Appendices **133**
1 Storing and processing geographical data . . . 135
1.1 POSTGIS . . . 135
1.2 QGIS . . . 137
2 Decomposition of BUFR messages . . . 140
3 Process system generation using irregular cell space . . . 145
4 Simulation results . . . 150
4.1 Covid-19 spreading in Brittany, France . . . 150
4.2 Marker movement with currents . . . 153
5 Publications . . . 156
5.1 Open Data for Environment Sensing: Crowdsourcing Geolocation Data . . . 157
5.2 Interpretable Machine Learning for Meteorological Data . . . 166
5.3 Epidemic Spreading Simulation on Distributed Process Systems . . . 174
5.4 Shore Pollution Simulation Based on Tidal Currents and Ground Effects . . . 183

Chapter 1

Introduction

Chapter introduction: This chapter provides a brief look at the environmental and social problems occurring recently. We observed the changes in the environment, especially the green tides issues and the occurrence of epidemics. The general context of these problems and the availability of Open data related to health statistics and agricultural chemicals motivate us to build simulation models based on Distributed algorithms and Cellular Automata (CA) approach. In this work, we introduced the use of geographic divisions as irregular cell spaces to generate distributed processing systems. This approach provides a chance to connect cell processes with open data, which is increasingly developing and being focused in many countries all over the world. A case study of epidemic spreading simulation on irregular cell space is provided. The cooperation of regular tiles and irregular cells is also considered in monitoring pollution on the shores. This is a new approach to modeling the impact of ground activities in coastal areas with the ocean behaviors represented by tidal currents.

1.1 Environmental and Social Problems

Environmental changes drive consequences for humans with more frequent and intense droughts, storms, heat waves, and rising sea levels. These disasters immediately impact human lives causing thousands of deaths [1] all over the world. The severe impacts will be even worse with social changes as the human population, industrial activities, and energy consumption continue to grow. This clearly indicates the necessity of monitoring to manage and minimize the harmful effects on the natural environment and protect human beings. Especially, intelligent environmental monitoring requires the integration of wireless sensor networks, machine learning techniques, and IoT devices. In the scope of this work, we will focus on the two phenomena, green tides on the shores and pandemics,

1. https://ourworldindata.org/natural-disasters#annual-deaths-from-natural-disasters accessed on 18 October 2022

described as follows.

1.1.1 Environmental Problems

Fast-growing macro-algae bloom is one of the serious environmental pollution problems that can change ecological diversity with the dead zones being created. A wide range of problems occurs due to green tides, such as the threat to marine ecosystems, unattractive bays to tourism, the cost to collect the algae blooms, and human health threats from toxins during the decomposition. As an example, in July 2019, all of Mississippi's mainland beaches closed to the public because of toxic algae blooms [2] and at the same time, 6 beaches around Saint-Brieuc were closed due to the unmanageable quantities of green algae [3] (Figure 1.1). Green tides with drifting sea lettuces have affected Brittany coasts for several decades. The general consensus about the problem among scientists is greater amounts of nutrients (nitrogen and phosphorus) flowing towards the coast in places with low water exchange and high temperature condition [24, 38, 59]. Nutrient enrichment of coastal ecosystems that leads to eutrophication is recognized as a major pollution threat [33]. The problem is associated with human activities as chemicals intensively used in agriculture or aquaculture. The key to addressing the problem is to model the impact of ground activities on coastal marine through the interaction and exchange matter of the two neighboring regions.

1.1.2 Social Problems: Pandemics

An epidemic is a social issue attracting public attention to the bad experience of the Covid-19 pandemic. Throughout history, terrible pandemics have killed millions of people and heavily impacted on the economy and society such as plague, cholera, flu, severe acute respiratory syndrome coronavirus [127]. Until the 18^{th} century, three huge plague pandemics, and infectious bacterial diseases, occurred with millions of deaths [185]. During the 19^{th} and 20^{th} centuries, cholera pandemics [88] occurred with high mortality caused by ingestion of food or water contaminated with bacteria. In the 20^{th} century, the three flu pandemics, which are Spanish flu, Asian flu, and Hong Kong flu, occurred from the emergence of a novel influenza strain [136]. Covid-19 was found in December 2019

2. https://www.npr.org/2019/07/09/739874122/toxic-algae-bloom-closes-25-beaches-on-mississippis-coast-fed-by-fresh-floodwate accessed on 20 October 2022

3. https://edition.cnn.com/2019/07/16/europe/french-beaches-toxic-sea-lettuce-scli-intl/index.html accessed on 20 October 2022

that is a respiratory syndrome caused by coronaviruses originating from animals [61]. The outbreak quickly spread worldwide and was declared a global pandemic in March 2020 (global situation shown in Figure 1.2). The clinical symptoms of the disease include fever, cough, and dyspnea [70]. The viruses spread mainly between people who are in close contact with each other, from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing, or breathe to another person through the air inhaled at short range or if infectious particles come into direct contact with the eyes, nose, or mouth [4]. People may also become infected when touching their eyes, nose, or mouth after touching surfaces or objects that have been contaminated by the viruses. Close and direct contact helps to transfer viruses between humans.

(a) (b)

Figure 1.1 – Green tides in July 2019. a) Valais Beach in the bay of Saint-Brieuc, northwest France. b) Mississippi's beach, southeastern of the United States.

Efforts to control outbreaks have always been the research of effective vaccines, but they are often not available until after the pandemic had peaked in most countries. Some non-pharmaceutical interventions such as quarantines, school closures, and banning public gatherings have the potential to delay and flatten pandemic peaks. If people are less mobile and interact with each other, the virus has fewer opportunities to spread. However, the world is interconnected and increasingly globalized thus people and diseases are carried to any city in a matter of hours. Despite all the differences between pandemics, history has shown that pandemics happen in cycles, and the issue is not whether another pandemic will occur, but when. Thus, there is always a need for tools and methodologies to analyze pandemics, provide better understanding, and protect human life.

4. https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted

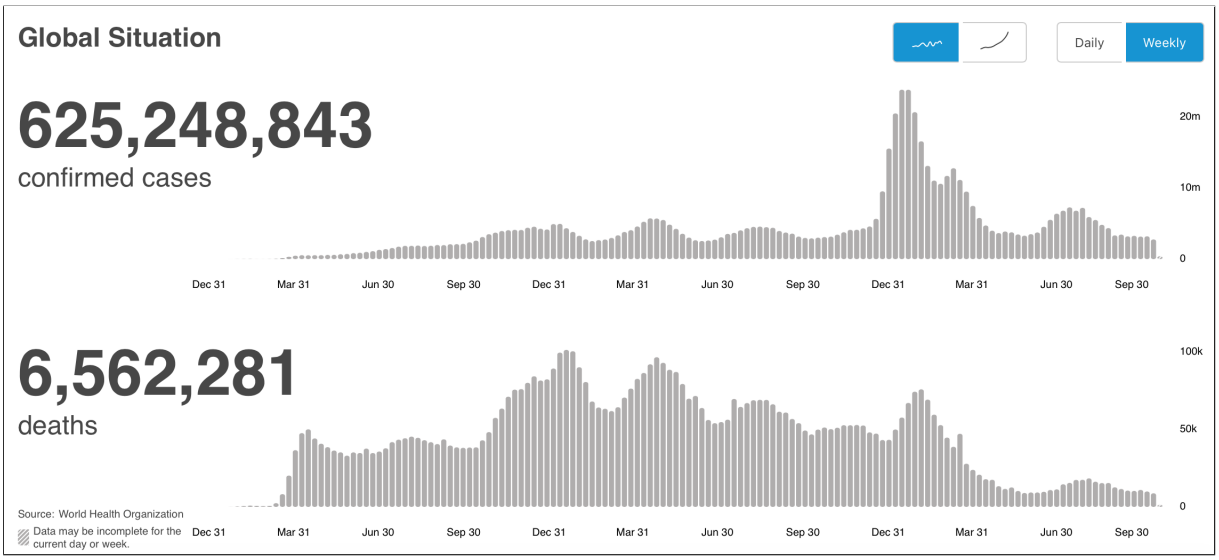


Figure 1.2 – Global situation of Covid-19 reported by World Health Organization. The confirmed cases are shown in the top of the figure and the deaths are in the bottom of the figure. The period is from December 2020 to October 2022.

1.2 Modeling and Simulation Complex Systems with Availability of Government Open Data

1.2.1 Complex Systems

The qualification of complex systems figures a difficulty to analyze systems highly variable in space, in local components defining their state [104]. Ordered systems are characterized by perfectly structured and therefore simple to predict. Disordered systems consist of independent components without any constraint thus they cannot be predicted on an individual level, but a prediction of average behavior is not only possible but trivial. Indeed, truly complex behavior is neither completely ordered nor completely disordered. Self-organization is a popular focus in complex systems science. The global organization in these systems results from distributed and localized interactions between their elements. Considering a group of fish staying together and swimming in the same direction in a coordinated manner, each individual fish moderates its behavior with reference to its immediate neighbors [128].

In environmental or social situations, quantifying is difficult due to contextual complexity. A practical example is the context of a real experiment related to tracking the life

of mussels as modeled by IFREMER, an oceanographic institution in Brest, France. The modeling process needs to go through a campaign, collecting samples, and studying the interactions by statistics [94]. The detailed report describes such a medium-term mission involving the mussel life cycle, marine currents, and species mobility tracked by satellite communication, and a significant spreading in the English Channel.

1.2.2 Modeling and Simulation with CA

Representing environmental behaviors is critical to understanding the internal logic that produces dangers. Phenomena as simple as excessive rain or sea rise will generate critical situations due to flooding. Simulations allow us to warn about the risks. The key point is to represent physical facts as program data, and then to process these data to represent what nature will do. Two classes of simulation models that have made major breakthroughs in regional science are agent-based models and CA [27]. These two are both complex systems approaches with elements and actions to form global behaviors. Agent-based systems require generally more modeling time and computing resources since they need to deeply understand all actors in the system such as behaviors, reasoning mechanisms, and resources. CA models are preferred when modeling space represented in the form of geographic cells that are suitable for modeling behaviors in the natural environment. Data models can be found from wide-scale distributed models such as CA of high density, and distributed algorithms, sparse and less regular. CA are discrete dynamical systems known to support complex computation and have been used to model complex systems in nature, including fluid flow [172], earthquakes [121], and biological pattern formation [37]. Their computation is qualitatively understood in terms of emergent coherent structures which are widely accepted to embody information storage, transfer, and modification. Most of these applications use regular patterns such as squares or rectangles to implement 2D grids.

A list of works has been done in our research team for natural environment simulation based on CA approach. Eloi Keita [81] focused on sound propagation simulation in an urban environment based on CA approach. A system was made of interconnected processes, that can reproduce collective behaviors suitable for sound propagation in consideration of streets, gardens, ring roads, buildings, and rivers. Pierre-Yves Lucas [106] produced QuickMap, a navigation software, that allows interfacing GIS databases and tile servers similar to Open Street Map, taking care of sensor locations and outputs to produce cellular systems oriented to physical simulations. Bao Lam [89] monitored the

environment with efficient sensors and radio communications as an evolution of manual light traps in the case of rice. Radio communication weaves these observatories into a network with a connection to databases storing measures and possible counteractions. Tuyen Truong [156] described simulation techniques based on geographic analysis to compute long-range radio coverages and radio characteristics in complex terrains. Geographic analysis was achieved using segmentation tools to produce cellular systems in the case of heavy rainfall and flooding. Thao Truong [154] described the environment evolutions using regular cell systems having geographic coordinates, channel communications, and propagations by physical dependencies. High-performance computing with MPI was focused on the case of a wide area with significant improvement in running time.

1.2.3 Emergence of Government Open Data

Open data provides valuable information for monitoring and managing problems. It ensures that individuals and groups of the social community can access, use, and redistribute the data without any special restrictions. The term open data has become popular and becoming a trend in developed countries. The interest starts since it is an indicator of the United Nations level of e-government development. Many countries have set up dedicated portals for sharing open data with widely recognized benefits. The United States is the first country to publish Government Open Data through the government data portal [5]. The portal was opened on May 21, 2009, at the initiative of President Barack Obama. Along with providing data, description data is also added to provide more information about each dataset such as data content, origin, and update time. After the United States published Government Open Data, the provision of Government Open Data quickly became the goal of information and data transparency commitments of many countries.

French government open data portal [6] was established in January 2011 to which public information is published, centralized, and shared freely by the government, public institutions, and local government. This portal provides a unique platform to promote data reuse and facilitate the creation of applications and services. Until January 2024, 47,317 datasets have been published and 3,720 datasets have been reused/included in applications or projects. A wide range of data is available through this platform including geography, economics, energy, and health. Different formats are available, in which the most common formats are XLS and CSV. There are also formats for voting records such as XML/JSON

5. https://data.gov accessed on 02 January 2023
6. https://data.gouv.fr accessed on 02 January 2023

and the land register such as Shapefile, MIF-MID, and GeoConcept export. Especially, Publish Health in France has published datasets related to Covid-19 since April 2020. This openness to data has fueled the work of researchers and paved the way for a myriad of visualizations in the press.

1.3 Motivation and Objectives

1.3.1 Motivation

This work is motivated by scientific models based on the CA approach in complex system simulations and the availability of open data through government portals related to geographical data, health statistics, and agriculture chemicals. CA models have shown the ability to imitate complex structures by simple systems while providing an intuitive understanding of the development process of natural problems. These models have been used in various fields because of their parallel nature, in which all cells change their states synchronously in discrete time steps. The initial configuration encodes the input data and time evolution yields final outputs. We observed the emergence of government open data related to health statistics, demographics, and environment linking geographical objects. Data and data models are important to provide a better understanding of the problems and effectively determine the cause of the problems. Geographical objects have the advantage of correct data bound into irregular cell processes and shapes reflect the complexity of objects. Connectivities and dependencies between geographical objects follow shape organizations.

1.3.2 Objectives

The objective of this work is to provide a general way of environmental and social simulation using geographical objects as irregular cell spaces. The CA approach is to counteract complexity, that will remain in large and irregular data systems. Open data are explored and exploited to provide transition rules for problem evolutions.

The three main contributions of this work are as follows:

— Introduce an approach of using geographic divisions as irregular cell space in modeling and simulation. We provide a new module (PickShape - see section 3.2) within this research framework to support modeling and simulation using government open data linked to irregular geographical objects. This is an alternative

to regular tiles that have been used in previous works in our research team with QuickMap/PickCell. The input data are geographical objects in shapefiles, and the neighborhood is the adjacent neighbors, objects sharing a point or line in common.

— Analyze open data available in the French government portal to learn the dependencies between the parameters and the Covid-19 incidence rate to produce transition rules for the epidemic propagation simulation. The data types are demographic data, the mobility data, and weather data. A case study of epidemic evolution in Brittany, France has been shown using geographic divisions as irregular cell space.

— Introduce an approach of cooperating irregular cells in land and regular cells in the oceans to monitor shore pollution related to agricultural chemicals released from land. The simulation strongly depends on the tidal currents to imitate the water circulation in coastal areas. An illustration of water circulation is done by dropping virtual markers in the sea area. We also show some possible places of green algae problems.

1.4 Thesis Outline

The thesis will be organized as follows:

— **Chapter 2** presents the Open Data concept and describes several types of open data. The computation of geographical objects is discussed for later use in the simulation system generation. The Express modeling language is used to model BUFR, a kind of meteorological data, to analyze and provide an accurate BUFR parser. An overview of the Open Data in environmental simulation is provided.

— **Chapter 3** describes the CA approach in modeling real life phenomena with distributed systems. The works using regular and irregular lattices are presented to have an overview of the scientific context of the approach. PickShape is introduced to generate distributed process systems using geographical objects. Each object is bound with a set of open data according to its identification.

— **Chapter 4** details the data analysis process using machine learning models. The important parameters are analyzed to provide transition rules for epidemic propagation in irregular cell spaces. A brief look at the research related to epidemic modeling is also presented, including the use of CA models to examine the spatial and temporal epidemic propagation.

— **Chapter 5** presents the approach of regular and irregular cell space cooperation

in modeling coastal pollution. Tidal currents are modeled to simulate water circulation on coastal areas and monitor green algae issues given agricultural chemicals released from land use.

— **Chapter 6** concludes the work with personal perspectives. Discussions of future plans are based on this research direction.

The general context, motivation, and objectives of this work have been introduced and the next chapter will give more details on Open Data.

Chapter 2

Open Data

Chapter introduction: This chapter presents the concepts of Data and Open Data. Open Data has received more attention, especially in developed countries since they recognize the value of openness to develop new applications, solve problems, and generate new values for society, the economy, and the environment. Geographical objects of administrative divisions publicly available in the government Open Data portal are retrieved and object neighborhood is computed allowing space segmentation irregularly. The description of meteorological data and data modeling language is presented in this chapter in the case of BUFR data, a type of partially Open Data in the sense that it is accessible and can be redistributed but it is not really usable in a non-common format.

2.1 Data and Open Data

2.1.1 What is Open Data?

Data is the lowest level of abstraction from which information and then knowledge are derived. It is at the bottom of a three-tier triangle with data, information, and knowledge. Data is captured and stored in a repository then it is processed and organized to become useful information, and this information is analyzed and interpreted to become knowledge [84, 101]. Data are the locations, photos, descriptions, and prices that one may need in order to plan a holiday for instance. All of the data are taken and organized into information about the place to go for a holiday and from that one contextualizes it and turns it into knowledge that is relevant to him. At the same time, others could take all of that data and compare the prices in different locations across the country, which is a completely different usage. It is critical to have the data to build different levels of information and allow people to contextualize them to build their own knowledge.

Open Data is data that anyone can access, use, and redistribute [1]. Data becomes

1. https://data.gov.ie/edpelearning/en/module1/#/id/co-01 accessed on 30 November 2022

accessible by being made available online and becomes usable by being made available in a common, machine-readable format. The most important thing is data can be used without a license or data can be licensed and this license allows for free reuse of the data in any way, including transforming, combining, and sharing it with others, even commercially. If a commercial entity uses an Open Dataset, the original dataset version will remain available and open without restriction. Restrictions can only be applied to the derivatives of the original dataset, never to the original dataset itself.

2.1.2 Why Open Data is Essential?

The term Open Data is introduced when people have problems accessing and using data that is commercially valuable. In fact, data is considered a new kind of resource, which has its intrinsic value. It is necessary to transform or refine the data to take full advantage of its internal value. However, Open Data does not necessarily mean without costs. In order to maintain and sustain the availability of the Open Data, there may be some marginal costs. The most common occurrence is that this cost should be no more than the reasonable reproduction cost of the data unit that someone is asking for.

The evolution of scientific research is characterized by an accelerating growth in scale, scope, and complexity of digital data. The extreme quantities of data produced by government agencies, research institutions, and industry are a fundamental component of scientific research. Thus, the roles and values of Open Data in global science have been discussed with the opportunities and challenges to the global science system associated with establishing an Open Data policy [118, 158]. Restricting access imposes structural inefficiencies, higher research costs, and slowing scientific progress. In healthcare, opening up medical data enables semantically to relate and enrich data on symptoms, diseases, diagnoses, treatments, and prescriptions offering the potential for improvements in care for individuals [85]. As an example, the Kaggle NIH Chest X-ray dataset [2] includes 112,120 disease-labeled X-ray images from 30,805 patients and the Kaggle head CT dataset [3] includes 2,500 brain window images and 2,500 bone window images from 82 patients. These datasets are publicly available on the Kaggle community supporting the clinical diagnosis using a variety of machine learning methods. This provides a better understanding of health outcomes as well as enabling more efficient ways of working for healthcare prac-

2. https://www.kaggle.com/datasets/nih-chest-xrays/data accessed on 03 February 2023
3. https://www.kaggle.com/datasets/vbookshelf/computed-tomography-ct-images accessed on 03 February 2023

titioners with effective technical support. From citizens' perspective, Open Data benefits come with improvements in care for long-term conditions using remote technologies enabled by easier access to information. Another example is the health indicators available on the Open Data Government portal. The public health agency in France provided different datasets related to diseases and health indicators. As part of an information mission related to the Covid-19 pandemic, the Public Health in France disseminates data related to this crisis since mid-March 2020 via their web geo-statistical observatory named Géodes [4]. The web portal was created in order to improve the visibility and accessibility of these indicators allowing users to explore the geographical dimension through interactive maps along with graphs and tables. The health data is either a number of disease cases, a proportion, or an incidence rate stratifying by categories such as gender or age group. These data are available online without any access restrictions.

Open Data has the power to transform the interaction between governments, businesses, and society to unlock real value [18]. It helps to make governments more transparent and provides evidence that public money is being well spent and policies are being implemented. Currently, many countries have set up dedicated portals for sharing Open Data. Open Data is available in 97% of countries in a study of 115 countries [5]. Based on the available data, new applications can be developed and problems can be solved to generate new values for society, economy, and environment. The web is already an accepted part of our societal infrastructure and Open Data is the next critical part of this infrastructure.

In this chapter, we will take a look at the two main data types used in the content of this thesis that are geographical and meteorological data.

2.2 Geographical Data

Geographical data is defined as data having an implicit or explicit associated with a location relative to the Earth. There are two common formats, in which geographical data are stored: vector or raster [21]. Vector data defines objects, polygons, and other involved units so that they are displayed or analyzed based on their associated attributes. Raster datasets are stored as a set of uniform grid cells that represent a continuous surface. Geographical data is processed and analyzed using geographical information systems (GIS).

4. https://geodes.santepubliquefrance.fr accessed on 23 February 2023
5. https://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf accessed on 31 October 2022

These are computer programs that help users make sense of geographical data including management, manipulation and customization, analysis, and creating visual displays.

Maps are a common practice of presenting geographical data as this is an easy way for humans to perceive the Earth. Geographical data is not only used for graphical visualization but also used for statistics and analysis. Let us look at the way John Snow analyzed a cholera outbreak in Soho, London in 1854 [144]. The cholera pandemic [88] occurred in the 19^{th} and 20^{th} centuries has left an indelible mark on human and medical history. At that time, people believed that cholera was spread in the air but John Snow proposed a hypothesis of the epidemic transmission via drinking water. During the outbreak, he identified the homes of those who were impacted by the disease on a city map to better understand the epidemic. He discovered that there was a relationship between infected people and the place where they lived. From this spatial analysis, he identified the Broad Street pump that supplied contaminated water to be responsible for the spread of cholera. He drew a crude Voronoi region [6] around the Broad Street pump with an equidistance line delineating the area closer to the pump than to any other pumps (Figure 2.1). This is an early attempt to generate a Voronoi diagram for only one Voronoi generator. A Voronoi region $V(p_i)$ of spatial object p_i (generator) is defined as a set of locations $P1$ satisfying the condition that the distance $d(p, p_i)$ between $P1$ and spatial object p_i is equal to or less than the distance $d(p, p_j)$ between $P1$ and any other spatial objects p_j [163]. The mathematical expression of a Voronoi region is as follows.

$$V(p_i) = \{p | d(p, p_i) \leq d(p, p_j), j \neq i, j = 1, 2, ..., n\} \tag{2.1}$$

Voronoi diagrams were studied by Georges Voronoi, who extended the investigation of Voronoi diagrams to higher dimensions with a wide range of applications such as epidemiology, geophysics, and meteorology. The spatial analysis by Snow provided new insight and evidence to the medical community that cholera was not in fact transmitted by air but was a result of ingesting contaminated food and drink. Another example is the use of geographical data in the field of surveillance and monitoring of diseases. Nulda Beyers et al. determined the geographical distribution of tuberculosis in a high incidence community [15]. The results of the study showed that 1,835 out of 5,345 dwelling units (34.3%) had at least one case of tuberculosis during the past decade and in 483 houses, three or more cases occurred. Their findings were critical showing certain houses where

6. https://mathworld.wolfram.com/VoronoiDiagram.html accessed on 01 November 2022

tuberculosis occurs repeatedly. This information should be used to direct health services to concentrate on certain high-risk areas.

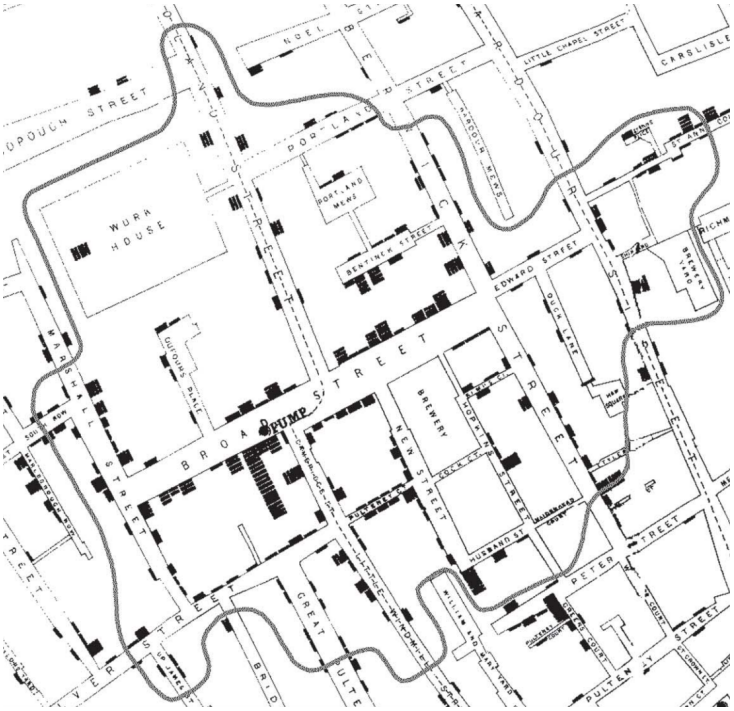


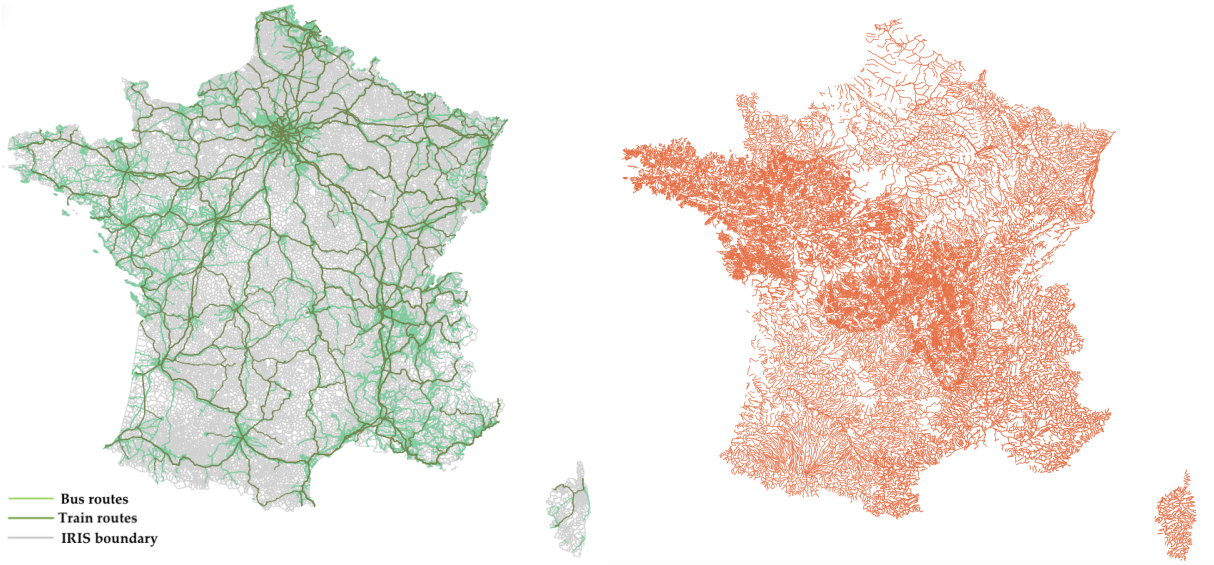
Figure 2.1 – Voronoi region of Broad Street pump by John Snow for identifying the contaminated water to be responsible to the spread of cholera.
(https://johnsnow.matrix.msu.edu/book_images12.php)

OpenStreetMap [7] is a project aiming to create a geographical database that is free to use and edit [13]. The database is built by volunteers who gather information by recording their moves using global positioning system receivers. The involvement of a large number of users in data creation, namely crowdsourcing [68], is a phenomenon of distributing tasks and gathering results using the Internet. The power of crowdsourcing had been shown in the way OpenStreetMap reacted to the heavy earthquake in Haiti in January 2010 [8]. Haiti was the poorest country in the western hemisphere with little information on the available maps at the moment. When the disaster occurred, aid was planned to support Haiti's citizens but there was not any way to identify the cities and buildings. Immediately, volunteers of the OpenStreetMap community initiated a mapping project to provide geographical data on roads and buildings to analyze damaged buildings, displacement camps, and triage centers. It quickly became the base map for many organizations

7. http://www.openstreetmap.org accessed on 01 November 2022
8. https://wiki.openstreetmap.org/wiki/WikiProject_Haiti accessed on 02 November 2022

involved in the response and reconstruction.

To extract data from OpenStreetMap, users browse the website and use search, pan, or zoom tools to find the area where they need to download the dataset. Besides, OpenStreetMap is a contributor to the Open Data Government Portal, particularly in France [9]. The geographical data of French territories can be downloaded as a shapefile, a common vector file format. Despite being called a "shapefile", the format is actually a compilation of many different files, in which the SHP, SHX, and DBF are mandatory to create a functioning shapefile. These file formats include Feature geometry, Index format for the feature geometry, and Feature attribute information, respectively. Currently, shapefile is supported by almost all commercial and open-source GIS software. An example of geographical data is the shapefiles of bus and train routes provided by OpenStreetMap (Figure 2.2a). The water surface is another example of environmental data provided by the Water Management Agency (Figure 2.2b).



(a) (b)

Figure 2.2 – Geographical objects such as public transportation routes and rivers. a) Train and bus routes in France represent the mobility data as the travel needs of people living in the country. b) Water surface in France provided by Water Management Agency including distinct parts of water surface such as rivers and canals.

Not only the transportation and territories, but other types of data such as health indicators, economy, and environment are provided as geographical objects. This will

9. https://www.data.gouv.fr accessed on 02 November 2022

improve the visibility of data over space.

2.2.1 Administrative Divisions

In this section, we present the geographical data of French administrative divisions used in this work. The geographic division is a way of managing a country by dividing it into smaller parts. It is at different hierarchical levels as defined by the Nomenclature of Units for Territorial Statistics (NUTS) and Local Administrative Unit (LAU). NUTS was established by Eurostat [10] to provide territorial units for the European Union. To be more specific, we consider France's geographic divisions with different hierarchical levels (Figure 2.3). Metropolitan France is divided into 13 regions (NUTS1) and 22 sub-regions (NUTS2). These divisions again are subdivided into 96 departments (NUTS3). Each department has an identification number, which is the first two digits (or letters) in a postal code. The departments are divided into districts and each district encompasses a number of cantons (LAU1). The cantons are subdivided into towns (LAU2).

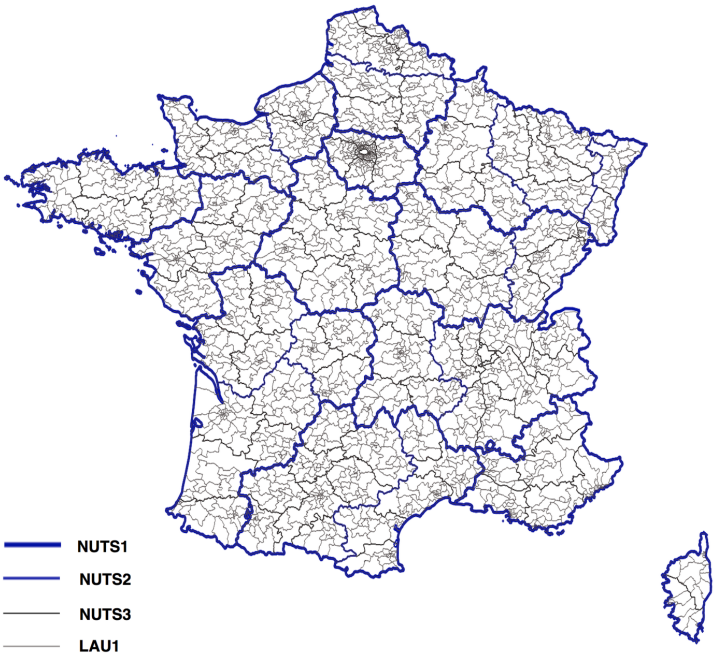


Figure 2.3 – Different administrative levels of geographic divisions in France. NUTS1 is the regions defined since 2016. NUTS2 is the sub-regions, which are former regions defined before 2016. NUTS3 is the departments. LAU1 is the cantons.

10. https://ec.europa.eu/eurostat accessed on 02 November 2022

In addition, the French National Institute of Statistics and Economic Studies (INSEE) has provided a definition for IRIS [11] to prepare for the dissemination of the population census. Towns with more than 10,000 inhabitants and a large proportion of towns with between 5,000 - 10,000 inhabitants are divided into several IRIS units. This separation represents a division of the territory, which carries its own data of the local area. There are four types of IRIS, the residential IRIS (code H), the business IRIS (code A), the miscellaneous IRIS (code D), and not divided communes (code Z). The general structure of the geographic divisions includes metric and topological properties such as area, perimeter, edges, length of edges, and population. Table 2.1 presents the average areas, edges, and population of the divisions at different levels.

Table 2.1 – Geographic divisions in France Metropolitan

Division	**Number**	**Avg Area** km_2	**Avg Num Edges**	**Avg Population**
NUTS1	13	42,245.69	3.54	4,945,714.46
NUTS2	22	24,963.36	3.90	2,922,467.64
NUTS3	96	5,720.77	4.97	669,732.16
LAU1	1995	275.285	5.63	32,227.71
LAU2	35,370	15.527	5.95	1,817.76
IRIS	48,590	11.30	5.99	1,323.20

The geographical data of French geographic divisions is provided by OpenStreetMap, except the IRIS which is provided by the National Geographic Institute (IGN) [12]. The data are in shapefile format downloaded from the French Open Data Government Portal, which can be stored in a local database for later use. The content is quite simple with three main attributes, an identity code provided by INSEE, the name of the division, and geographical data. In fact, local data related to population and government management linking to each IRIS are provided by the authorities. In this work, we store the geographical data in a PostGIS database and process it with spatial SQL queries. The data are used to generate irregular cell systems (will be discussed later in section 3.2). Detailed descriptions of how to store downloaded datasets in a local database and how to process these data are presented in section 1 of the Appendix.

11. https://www.insee.fr/en/metadonnees/definition/c1523 accessed on 02 November 2022
12. https://geoservices.ign.fr accessed on 02 November 2022

2.2.2 Neighborhood Computation

Neighborhood computation is an important part of using geographical data to generate simulation systems. The neighbors of a polygon (geometry in a shapefile) are the polygons that share a border with it one or more edges. Given N polygons, to detect the intersection of each polygon, the worst case is that we need to check the polygon with all other polygons: $O(n^2)$. Significant improvements in intersection detection time may be possible if preprocessing is allowed [117]. One of the best-known techniques to filter complex intersection tests is to compute an axis-aligned bounding box for each object [148, 184]. Two objects need to be tested for intersection only if their bounding boxes intersect. It is very easy to test whether two such boxes intersect by comparing their projections on each coordinate axis. The approach to calculating the neighborhood of each polygon by Filter-and-Refine is shown as follows.

1. For each polygon, compute the bounding box
2. For each bounding box
   - Filter: Find all bounding boxes overlapping with it to narrow down the polygons for checking neighborhood
   - Refine: For each polygon in the overlap bounding boxes, check if the two polygons are really intersect

Filter: A rectangle (bounding box) can be defined by just one of its diagonals. Let us say the first rectangle's diagonal is (x_1, y_1) to (x_2, y_2) and the second rectangle's diagonal is (x_3, y_3) to (x_4, y_4). The condition to check whether the rectangles are overlapped is $(x1 < x4)\&\&(x3 < x2)\&\&(y1 < y4)\&\&(y3 < y2)$ (shown in Figure 2.4).

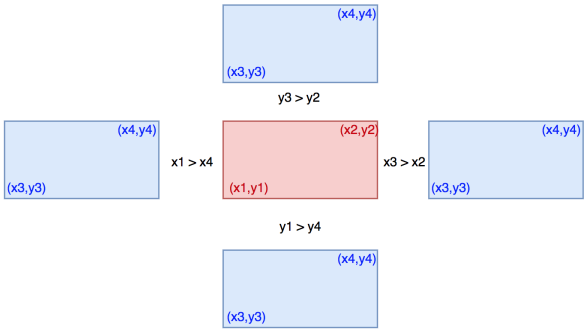


Figure 2.4 – Rectangles overlap conditions. To check whether the rectangles are overlapped, the condition is $(x1 < x4)\&\&(x3 < x2)\&\&(y1 < y4)\&\&(y3 < y2)$.

Refine: Given two polygons $P1$ and $P2$ with m and n vertices, do they intersect? If

$P1$ and $P2$ intersect, then either $P1$ contains $P2$, $P2$ contains $P1$ or some edge of $P1$ intersects an edge of $P2$. Since both $P1$ and $P2$ are simple (no edge intersections inside a polygon), any edge intersections that occur must be between the edges of different polygons. If no intersection is found, we still must test whether $P1 \subset P2$ or $P2 \subset P1$. The Intersection of simple polygons is linear-time transformable to line-segment intersection testing. Given N line segments in the plane $S = \{s_1, s_2, ..., s_n\}$, determine whether any two intersect. A naive approach is to check all pairs of segments (s_i, s_j) whether they are intersected. This algorithm has a running time $O(n^2)$. In [129, 139], Shamos and his colleagues presented the Plane-sweep algorithm that determines whether two simple polygons intersect in $O(NlogN)$ time. We can formalize the algorithm as follows. At first, we sort the line segments by their endpoints and handle them from left to right. We start to sweep with a line l from left to right over all segments. While we move line l, we always store which segments are currently intersecting the line. This allows us to find pairs of segments that possibly intersect that we can check for. Either we add a segment s_i, namely, when the sweep line moves over $s_i's$ left endpoint, or we remove s_i from this list, namely when l moves over $s_i's$ right endpoint. The list of segments is maintained in sorted order and if there is any change in the list structure, we check the segments in the set for possible intersections.

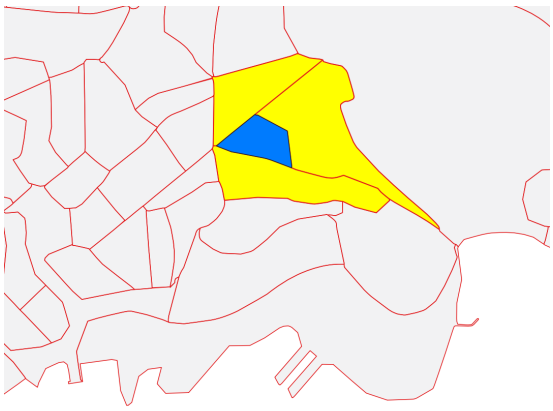


Figure 2.5 – A polygon (in blue) with its neighbors (in yellow). This is the adjacency neighborhood contained all polygons sharing a point or a line with it.

2.3 Meteorological Data

Meteorological data include any facts or numbers about the state of the atmosphere such as temperature, humidity, precipitation, wind speed, and pressure. Different instruments such as thermometers and barometers are used to measure different characteristics of the atmosphere in various locations. According to the World Meteorological Organization (WMO), billions of observations are obtained and exchanged in real time between their members and other partners every single day [13]. The information on the atmosphere is collected from 11,500 land-based stations, 1,000 weather radars, 1,300 upper-air stations, over 3,000 onboard aircrafts, 4,000 routinely reporting ships, 1,250 drifting buoys, more than 500 moored buoys, and a number of space-based components. The description of the global observing system is shown in Figure 2.6.

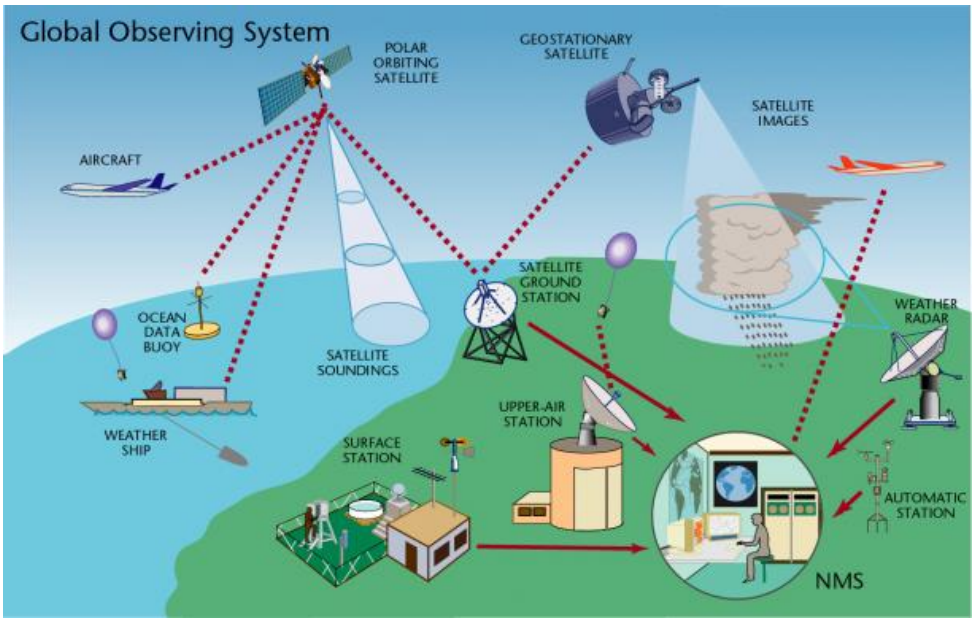


Figure 2.6 – Global observing system with weather stations, weather radars, aircrafts, buoys, ships, and satellites.
(https://public.wmo.int/en/programmes/global-observing-system)

The meteorological observations address weather phenomena that respect no national boundaries. The observing of the weather and even the forecast products need to be exchanged between the weather stations in international cooperation. Thus, it is essential to have a standard data representation for the interchange of data products. The WMO has a

13. https://public.wmo.int/en/resources/bulletin/global-observing-system accessed on 19 November 2022

long history in the development and operational use of meteorological data representation systems. BUFR (Binary Universal Form for the Representation of meteorological data) [145] is presented with the advent of computerized telecommunications and internationally accepted protocols capable of handling binary data. It is designed to represent any meteorological data in a logical and efficient way. The core concept of the BUFR standard is its self-descriptive nature, which helps this standard in accommodating changes. It only needs to have additional data description tables when there is new observation data.

2.3.1 BUFR

The term "message" refers to BUFR being used as a data transmission format. Each BUFR message consists of a continuous binary stream comprising 6 sections. Each section is a series of octets, coined to qualify one byte as an 8-bit sequence. It always consists of an even number of octets with extra bits added on and set to zero when necessary. In theory, there is no upper limit to the size of a BUFR message. However, by convention, BUFR messages are restricted to around 15,000 octets. An example of a complete BUFR message is presented in Figure 2.7. The message decomposition is shown as follows.

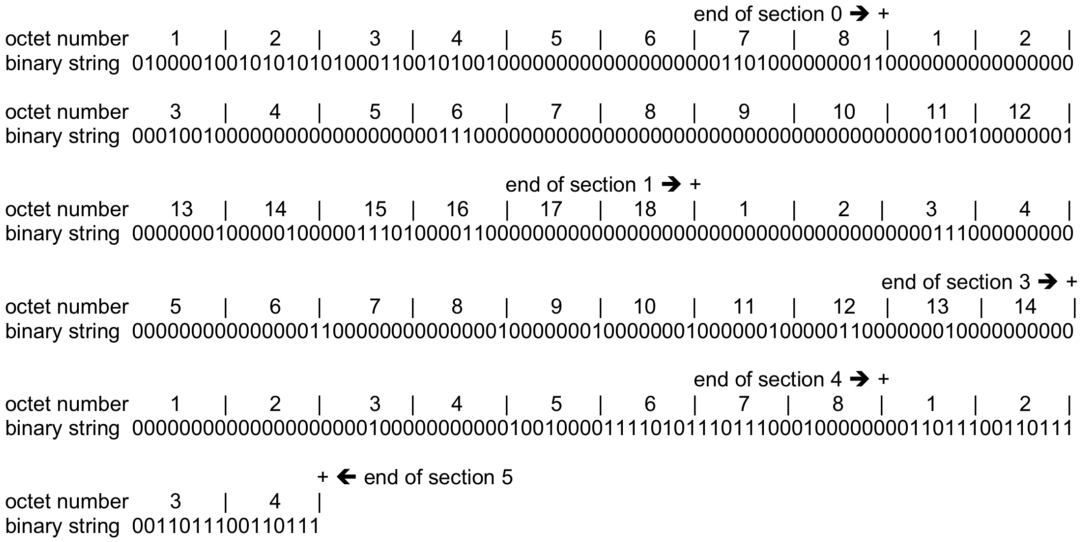


Figure 2.7 – Example of a complete BUFR message containing 52 octets [152] with 5 sections.

— Section 0 (Indicator section) includes 8 octets starting with "BUFR" characters in octets 1 to 4. The next three octets determine the entire length of the message. In this example, the length of the message is 52 represented by a binary number

"00110100" in octet 7. The last octet in the section contains the number 3, which is the BUFR edition number.

— Section 1 (Identification section) has different lengths between BUFR messages but not less than 18 octets. The length of the section is determined by octets 1 to 3, which is 18 in this example. The first bit of octet 8 is set to 0 indicating that Section 2 is not included in this message. The octets 13 to 17 of the section specify the date/time (year, month, day, hour, and minute) of the observation. Octet 13 represents the year in the century, meaning that "00000001" is the year 2001.

— Section 2 (Optional section) is not usually sent in international messages, but it is put to use in some computer centers that use BUFR frequently in a database context.

— Section 3 (Data description section) contains a collection of descriptors that define the form and content of individual data elements in Section 4 (Data section) starting from octet 8 of the section. The length of the section is determined by octets 1 to 3, which is 14 in this example represented by "00001110". The data descriptors are composed of three parts - F (2 bits), X (6 bits), and Y (8 bits). Thus, octets from 8 to 13 of the example are decoded in three descriptors (001001, 001002, and 012004), which can be found in the BUFR tables [14]. A short description of these descriptors is shown in Table 2.2.

Table 2.2 – Examples of data descriptors in Section 3

Descriptors (F X Y)	**Name**	**Unit**	**Scale**	**Data width (bits)**
0 01 000	WMO block number	Numeric	0	7
0 01 002	WMO station number	Numeric	0	10
0 12 004	Dry-bulb temperature at 3 m	K	1	12

— Section 4 (Data section) contains the binary data, as defined by the descriptors of Section 3, starting from octet 5 of the section. The length of the section is determined by octets 1 to 3, which is 8 in this example. The WMO block number occupies the first 7 bits, the WMO station number occupies the next 10 bits, and the temperature occupies the next 12 bits of octets 5 to 8 of this section. Thus, the WMO block number is 72, the WMO station number is 491, and the temperature is 295.2 degrees Kelvin (divided by 10 with a scale of 1).

— Section 5 (End section) is the four octets representing the four characters "7777". This is to notify the end of a BUFR message.

14. https://community.wmo.int/activity-areas/wis/latest-version accessed on 07 December 2022

Meteorological data in BUFR format can be downloaded from an FTP server [15] of the National Oceanic and Atmospheric Administration. This database consists of 13,000 weather stations all over the world, specifying by geographical locations from the year 2000 until the present. It is noted that the meteorological stations provided are all registered with WMO and each will receive a 5-digit WMO index for identification. The meteorological data in this case is partially open in the sense that it is accessible and can be redistributed but it is not really usable by being made available in a common format. In the next section, we introduce the Express modeling language, which allows us to analyze and provide an accurate parser for BUFR messages.

2.3.2 Express Modeling Language

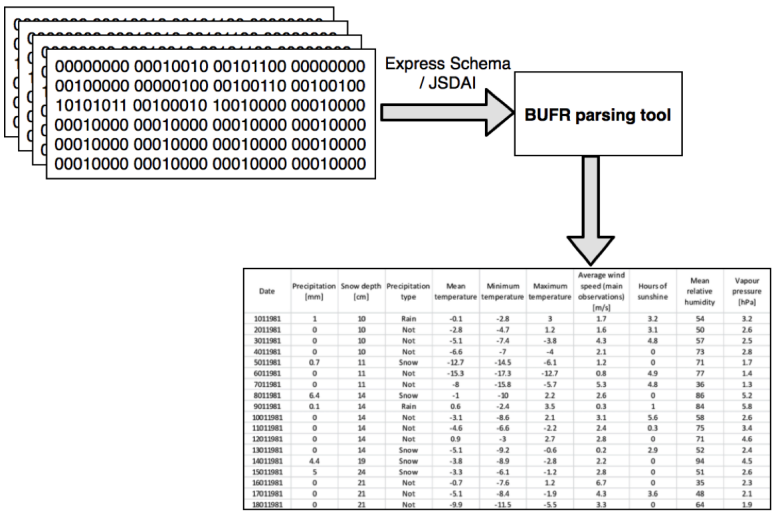
Express is a standard data modeling language, which is formalized in the ISO standard 10303-11 [149] (within the STEP - Standard for The Exchange of Product model data). It is an object-flavored lexical language that is firstly designed to represent the models of industrial products. The data modeling language helps define data objects and the relationships between the objects and enables the exchange of data between the objects. Express provides a series of data types for building blocks in a schema. The most important data type in Express is the entity data type. Entity attributes can relate an entity with other entities.

The BUFR is a strong and complex binary format with a self-descriptive nature. Thus, there is a need for a well-designed program for parsing the descriptors, matching the descriptors with the bit stream, and extracting the values out of the stream. In our work, we use Express modeling language to analyze the meteorological data in BUFR messages and provide an accurate BUFR parser. A data model for BUFR messages is defined in two ways, textual or graphical. In a textual form, a SCHEMA is clarified in which various data types and the structural constraints and algorithmic rules can be defined. The Express-G is the graphical representation for all details formulated in the textual form. The advantage of using Express-G is that the information is presented more understandably. The details of BUFR schema are presented in section 2 of the Appendix.

The data in BUFR format and its descriptive tables will be used in combination with the Express modeling language for transforming the raw data into any other machine-readable format such as CSV as shown in Figure 2.8. These data in machine-readable

15. http://www.meteomanz.com/ accessed on 08 December 2022

formats will later be used as the input for weather prediction with machine learning algorithms. We used some interpretability techniques to interpret the machine learning models on meteorological data as shown in Figure 2.9. This work has been published in the 5^{th} International Conference on Machine Learning and Soft Computing (section 5.2 of the Appendix).



Date	Precipitation [mm]	Snow depth [cm]	Precipitation type	Mean temperature	Minimum temperature	Maximum temperature	Average wind speed (main observations) [m/s]	Hours of sunshine	Mean relative humidity	Vapour pressure [hPa]
1011981	1	10	Rain	-0.1	-2.8	3	1.7	3.2	54	3.2
2011981	0	10	Not	-2.8	-4.7	1.2	1.6	3.1	50	2.6
3011981	0	10	Not	-5.1	-7.4	-3.8	4.3	4.8	57	2.5
4011981	0	10	Not	-6.6	-7	-4	2.1	0	73	2.8
5011981	0.7	11	Snow	-12.7	-14.5	-6.1	1.2	0	71	1.7
6011981	0	11	Not	-15.3	-17.3	-12.7	0.8	4.9	77	1.4
7011981	0	11	Not	-8	-15.8	-5.7	5.3	4.8	36	1.3
8011981	6.4	14	Snow	-1	-10	2.2	2.6	0	86	5.2
9011981	0.1	14	Rain	0.6	-2.4	3.5	0.3	1	84	5.8
10011981	0	14	Not	-3.1	-8.6	2.1	3.1	5.6	58	2.6
11011981	0	14	Not	-4.6	-6.6	-2.2	2.4	0.3	75	3.4
12011981	0	14	Not	0.9	-3	2.7	2.8	0	71	4.6
13011981	0	14	Snow	-5.1	-9.2	-0.6	0.2	2.9	52	2.4
14011981	4.4	19	Snow	-3.8	-8.9	-2.8	2.2	0	94	4.5
15011981	5	24	Snow	-3.3	-6.1	-1.2	2.8	0	51	2.6
16011981	0	21	Not	-0.7	-7.6	1.2	6.7	0	35	2.3
17011981	0	21	Not	-5.1	-8.4	-1.9	4.3	3.6	48	2.1
18011981	0	21	Not	-9.9	-11.5	-5.5	3.3	0	64	1.9

Figure 2.8 – BUFR decoding into machine-readable formats. BUFR is modeled with Express; An application programming interface reads and compiles the model to generate parsing tool; Decoding a binary data stream to receive machine-readable format.

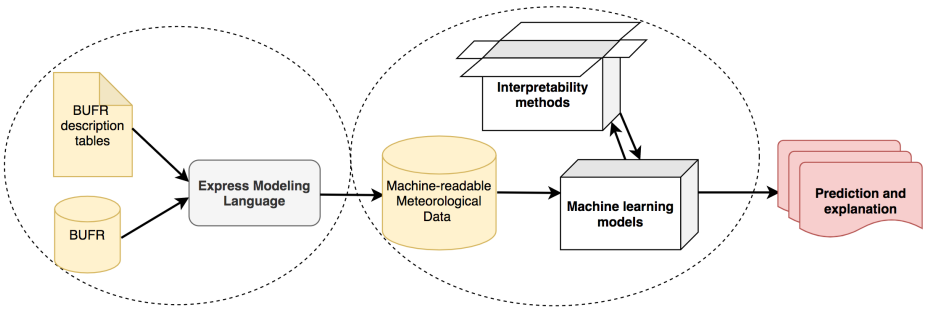


Figure 2.9 – Interpretability process of meteorological data. BUFR description tables and Express modeling language are used to decode BUFR into CSV. It is then used as the input for machine learning models with interpretability techniques to generate predictions and explanations.

2.4 Open Data in Analysis and Simulation

Open Data plays a crucial role in understanding and managing the environment, as it allows researchers, scientists, policymakers, and the public to access and analyze relevant data for a wide range of applications. These efforts promote transparency, collaboration, and evidence-based decision-making in environmental management and sustainability. Open Data in environmental simulation refers to the availability of publicly accessible data that is used to model and simulate various aspects of the environment. This data can include information about climate, weather patterns, land use, vegetation, water resources, and other factors.

In the field of climate modeling, historical climate patterns, such as temperature, precipitation, and atmospheric composition, can be used to simulate future climate scenarios. The availability of climate model data in Google Public Datasets [58] or Kaggle [83] is significant for researchers, scientists, and the general public. By making climate model data openly accessible through a widely used platform, researchers from diverse disciplines can access the data, conduct their analysis, and contribute to our understanding of climate dynamics and its implications. Jiantao Wu et al. presented a study focused on the development of an interoperable Open Data portal specifically designed for climate analysis [173]. The aim of the portal was to provide access to a Web-wide climate domain knowledge graph made for daily climate data for Ireland and England. The study proposed an architecture that leverages Open Data standards and technologies to enable seamless integration and interoperability of diverse climate datasets.

Open Data on past natural disasters, such as hurricanes, floods, and earthquakes, can be used to simulate and predict the behavior of such events. These data can include information about hazard maps, weather patterns, emergency services, infrastructure, population demographics, and other relevant factors. This helps in developing early warning systems, evacuation plans, and emergency response strategies. Crowdsourcing is a way to gather information from a large number of users in disaster management. Jens Ortmann et al. [122] proposed crowdsourced Linked Open Data (LOD) to ease the challenges of information triage in disaster response efforts. The study explored how LOD principles can be applied to integrate and link diverse data sources, enabling better data interoperability, analysis, and decision-making. It highlighted the benefits and challenges of crowdsourcing and LOD in disaster management and emphasized the importance of collaboration between stakeholders to leverage the power of citizen-generated data for effective disaster

response. Thushari Silva et al. [142] presented the same approach focusing on the application of LOD in disaster mitigation and preparedness. The authors discussed the use of LOD in areas such as hazard mapping, resource allocation, and early warning systems. Tao Lin et al. [103] focused on assessing the risk of urban water-logging by utilizing Internet Open Data. The researchers collected and integrated various Open Data sources, including meteorological data, topographic information, and urban infrastructure data. The study analyzed and evaluated the factors contributing to water-logging risks, such as rainfall intensity, land use, and drainage capacity. By utilizing Internet Open Data, the research identified high-risk areas prone to water-logging and provided insights into the spatial distribution of vulnerabilities. Ananya Gupta et al. [63] combined deep learning techniques and Open Data from OpenStreetMap for aerial image segmentation in disaster impact assessment. The authors utilized Open Data sources to train deep learning models for automatically segmenting aerial images. This helps to identify disaster-related features and assess the extent of damage. A comparison between the two proposed network models with several segmentation models was provided. The research highlights the potential of utilizing publicly available datasets and advanced machine learning techniques to automate and expedite the process of evaluating disaster-related damages. Si Wang et al. [164] proposed a methodology for conducting a quantitative risk assessment of storm surge by the use of the GIS software and Open Data. The building footprint data used for identifying elements at risk were extracted from the Open Data and the potential monetary damages were calculated by using depth-damage functions for different types of buildings. The study quantitatively assessed the potential impacts of storm surge events providing valuable insights on risk levels in coastal areas and identifying high-risk zones.

The impact of Open Data on urban studies and planning is another research topic. Ying Long and Lun Liu [105] examined the transformations and implications of data-driven approaches in urban research and planning. The work explored the utilization of various data sources in urban modeling and land-use analysis. The four major transformations of urban studies are explored in spatial scale, temporal scale, granularity, and methodology. The significance of freely available remote sensing data in supporting sustainable land management practices was analyzed in [130]. Various Open Data satellite missions, such as Landsat, Sentinel, and MODIS, showed their capabilities in capturing different aspects of land dynamics, including land cover change, vegetation health, and surface temperature. Open remote sensing data was used in diverse applications of land monitoring, including

urban growth analysis and agricultural monitoring. The WordPop project [16] focused on providing Open Data for spatial demography. The study [151] highlighted the importance of accurate and detailed population distribution data for various applications, including health planning, disaster response, and urban development. The methodology employed by WorldPop to produce high-resolution population distribution maps using diverse data sources, including census data, satellite imagery, and geospatial modeling techniques.

In epidemic simulation, an agent-based model is an approach with the ability to model at the individual level. To produce results that can be readily applied to a given population, agent-based models need to be data-rich in representing the population being modeled. Elizabeth Hunter et al. [72] demonstrated the use of Open Data in the development of epidemiological simulation models. The researchers utilized various sources of Open Data, including census data, transportation data, and social media data, to develop a representation of the population and its interactions within the towns. The agent-based model simulates individual agents with specific characteristics, such as age, occupation, and mobility patterns, to capture the dynamics of disease transmission. The research emphasizes the importance of Open Data availability and integration in epidemiological modeling, enabling more precise and context-specific simulations. In [71], the authors discussed the process of acquiring and processing Open Data for modeling, emphasizing the need for data quality assurance and integration techniques. By incorporating Open Data into the agent-based model, they generated realistic scenarios and assessed the impact of various intervention strategies. Teodoro Alamo et al. [3] provided an overview of the available open-data resources for monitoring, modeling, and forecasting the Covid-19 pandemic. The authors discussed various types of Open Data sources, including case, testing, hospitalization, mobility, and demographic data. They explored how these datasets can be utilized to track the spread of the virus, analyze its impact on healthcare systems, and predict future trends. The work also discussed the challenges associated with Open Data, such as data quality, standardization, and privacy concerns. Antonio Desiderio et al. [35] presented a study that combines a multiplex mobility network with meta-population epidemic simulations to understand the impact of human mobility patterns on epidemic dynamics. The study utilized various sources of Open Data to construct a multiplex mobility network that represents the interconnectedness between different regions in Italy. The authors then simulated epidemic outbreaks to assess the spread and impact of infectious diseases. The advantages of incorporating Open Data into the modeling approach

16. https://www.worldpop.org accessed on 16 June 2023

are highlighted as it is accessible by anyone, can easily be updated in the future, and can be tested against possible systematic errors. However, in their opinions, proprietary data typically have higher granularity, especially in the temporal dimension.

Rezník Tomáš et al. [131] presented a model that utilizes Open Data for precision agriculture applications and monitoring agricultural pollution. The authors integrated various data sources, including satellite imagery, weather data, soil data, and crop yield data, to provide comprehensive information for precision agriculture. This enabled farmers to make data-driven decisions regarding irrigation, fertilization, and pest control, leading to optimized resource usage and increased productivity. In addition, the authors discussed the application of the Open Data model in monitoring agricultural pollution. It highlighted the importance of tracking and analyzing data related to nutrient runoff, pesticide usage, and soil erosion to mitigate environmental impacts and ensure sustainable agricultural practices. Pavlík Jan et al. [125] explored the application of IoT devices and Open Data repositories in analyzing water pollution. The study focused on assessing the feasibility and effectiveness of these technologies in the Czech Republic. The authors discussed the deployment of IoT sensors to collect real-time data on water quality parameters such as pH, dissolved oxygen, and turbidity. They also utilized Open Data repositories that provide access to additional relevant data, including weather conditions and historical pollution records. The study highlighted the advantages of real-time data collection and the integration of various data sources for more comprehensive analysis.

Overall, the above studies provide insights into the utilization of global Open Data in different fields of study. It highlighted the importance of data availability and collaborative initiatives in advancing research and practical applications. With the same attention to the issue of environmental sensing, we presented a model for building environmental services with Open Data combining environmental data collection, Open database creation, and environmental service formation. This work has been published in EAI Endorsed Transactions on Context-Aware Systems and Applications 2020 (section 5.1 of the Appendix). Next, we provide our approach to using Open Data for environmental and social simulation. Figure 2.10 shows a general workflow of our approach with distributed processing systems. Administrative divisions are used as geographical objects in a set of tools to generate distributed process systems in different syntaxes (see Chapter 3). Open Data in a local place are analyzed with machine learning techniques to deduce transition rules for simulation using distributed processing systems and the cellular automata approach (see Chapter 4). The simulation results bring knowledge to humans and also can

be provided back to the Open Data sources.

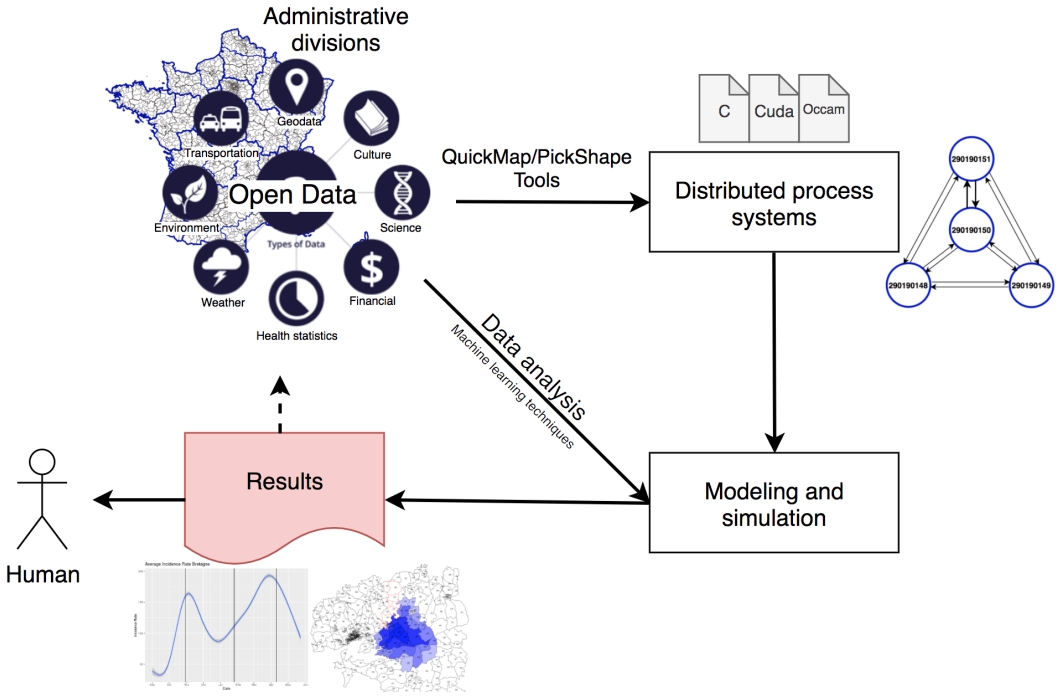


Figure 2.10 – General workflow of using geographic divisions as irregular cell spaces in modeling and simulation with Open Data. Open data are used to generate distributed process systems for simulation with models obtaining from historical data analysis.

Different Open Data types and study methods have been used in monitoring and modeling the environment. Table 2.3 presents a summary of the Open Data types and methods used in some related studies.

Chapter summary: The growing attention to Open Data in many countries has shown the benefits of openness in global science and government management. We present an overview of Open Data, its benefits, and two data types used in this thesis. The main data sources were available in the government Open Data portal in France. Geographical data have been shown to be useful not only for visualization but also for statistics and spatial analysis. Other data types can be provided in terms of geographical objects to improve the visibility of data over space. In particular, we focused on geographical data of French territories and meteorological data analysis in this chapter. The next chapter will describe the process system generation based on CA using these data.

Table 2.3 – Summary of Open Data types and methods in some related studies

Study	**Data**	**Method**
Tao Lin et al. [103] assess urban waterlogging	- Meteorological data (rainfall intensity) - Urban infrastructure (land use) - Topographic information (drainage capacity)	Analysis with ArcGIS
Ananya Gupta et al. [63] assess disaster impact	- Aerial image before and after a disaster	Deep learning models for segmenting aerial images
Si Wang et al. [164] assess potential building damage	- Meteorological data (tropical cyclones) - Socioeconomic data (land cover, population, GDP) - Geographic data (administrative boundary, building footprint)	- Hazard assessment with the Advanced Circulation (ADCIRC) model - Spatial analysis with ArcGIS - Depth–damage functions for risk assessment
Elizabeth Hunter et al. [72] develop epidemiological simulation models	- Census data, transportation data, and social media data	Agent-based model with age, occupation, and mobility patterns to model dynamics of disease transmission
Antonio Desiderio et al. [35] model the impact of human mobility on epidemic spreading	- Municipalities population, surface - Intra-province and inter-province adjacent for short range connections - Train and flights for long-range connections	Susceptible-Infected-Recovered (SIR) meta-population approach
Pavlík Jan et al. [125] analyze water pollution	- IoT sensors to collect real-time data (pH, dissolved oxygen, and turbidity) - Weather conditions and historical pollution records	ArGIS and BNHelp are used for spatial analysis
Our approach on Epidemic modeling	- Geographic data (administrative boundaries, transportation networks) - Meteorological data (temperature, humidity, wind speed, etc.) - Demographic (population density, age group, gender, immigrant) - Health statistics (infected cases, vaccination, lockdown measurement)	- Data analysis with machine learning models (Permutation feature importance, Accumulated local effects, Shapley additive explanation) - Distributed systems and Cellular automata approach – synthesize transition rules with Poisson distribution

Chapter 3

Geographic Divisions Modeled as Distributed Process Systems

Chapter introduction: Geographic divisions represent the spatial dividing of a country based on population and administrative levels. Governments monitor the country's status by these management levels thus data collection by geographic divisions will be more precise than a regular grid estimation over spaces. The Open Data emergence motivates us to use geographic divisions as irregular cell spaces in simulation using distributed processing systems. Geographic behaviors based on the CA model are presented in this chapter. Section 3.1 explains the CA model theory and related works with regular and irregular lattices. Section 3.2 provides details on simulation system generation using geographic divisions as irregular cell space and local data binding into cells. This is supported by a set of tools developed in our lab framework for environment modeling and simulation. Section 3.3 demonstrates the parallel computation with Occam/CUDA/MPI on the generated process systems. Simulations imply millions of cells and long run to cover spatial behaviors thus concurrent implementations allow to management of these needs.

3.1 Cellular Automata

Cellular automata (CA) is a common and simple model for parallel computation. John von Neumann and Stan Ulam initiated the concept of CA in the early 1950s [162]. The first idea was to study biological processes such as self-reproduction to gain a deeper understanding of complex systems having the ability to produce copies of themselves [161]. In the 1980s, Stephen Wolfram developed the computation theory of CA with comprehensive studies [170]. He has many fundamental studies presenting a gigantic collection of results related to CA [168, 169, 171]. CA is used in various fields of study by its parallel nature. It is considered a parallel processing computer, in which the initial configuration encodes the input data and time evolution yields the final output. Physics, mechanics,

biology, and sociology are domains where CA models are elaborated and supported by fundamental mathematical theories such as fluid mechanics [153], gas dynamics [54], or epidemiology [134].

3.1.1 CA Model Description

A cell system is a discrete spatial and temporal dynamic with four main elements: a grid of cells, a neighborhood, a set of initial conditions, and a transition rule. A discrete interpretation of reality is often the simplest way to separate physical concerns. In general, a system can be defined as regular tiles such as square, rectangle, or triangle shapes. Each cell has a finite number of possible states and acts as a finite automaton that interacts with other cells within a local neighborhood. A set of initial conditions is the initial states for each and every cell in the system. The state of each cell changes according to a function, known as the transition rule, and it is affected by the states of its neighbors (Figure 3.1). Based on its current state, the center cell receives/sends its information from/to the neighbors and changes its state under the global impact. All cells change their states synchronously thus the state of the entire lattice advances in discrete time steps. Synchronous behavior refers to the idea that all cells are updated simultaneously at each time step. This approach ensures that the interactions between cells are consistent and that the state of the system is updated in a consistent manner. The synchronous behavior of CA is advantageous in certain applications, such as in modeling physical phenomena or simulating biological processes. In these cases, it is important to ensure that the behavior of the system is consistent and that changes in one part of the system are reflected in all other parts. Although the behavior of each individual cell is simple, the interactions between all cells lead to intricate global behavior. The transition rules of the cell systems are strongly dependent on domain knowledge. It is usually according to the intuitive understanding of the development process of the problems.

The neighborhood is one of the major components in the process systems. The neighborhood structure is defined depending on the geometry lattices. The common neighborhood structures for two-dimensional CA of square lattices are von Neumann [161], Moore [115], and Margolus [153]. Von Neumann (Figure 3.2a) considered neighbors of a given cell to be the four adjacent cells that make up a diamond shape. Moore (Figure 3.2b) defined neighbors as the cells that have their coordinates differing by at most 1 from the coordinates of a given cell. The Margolus neighborhood (Figure 3.2c) is well-known of a block 2 x 2 cells neighborhood that specifies different neighbor cells in odd and even steps.

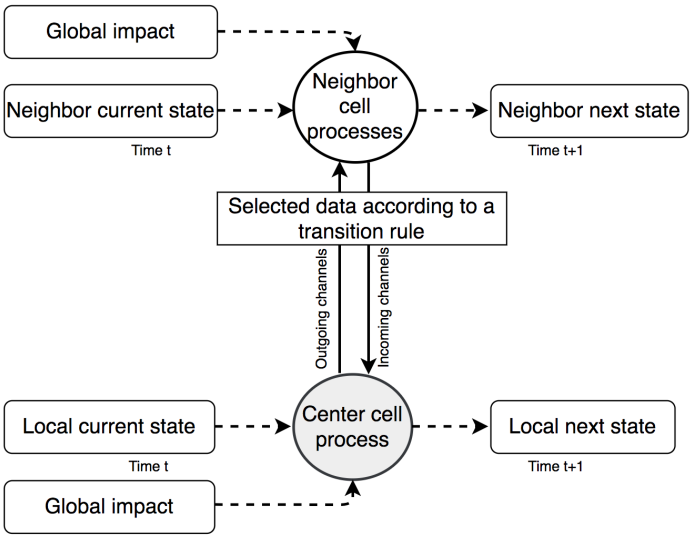
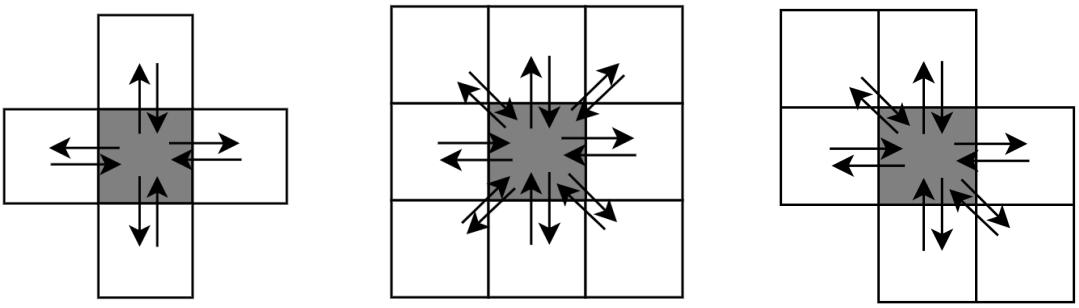


Figure 3.1 – Process state changes. Each cell has its current state at time step t. At time step $t + 1$, the state of each cell changes according to a transition rule. The center cell process sends and receives data from its neighbors via channels. These data will affect its state defined in the transition rule.



(a) (b) (c)

Figure 3.2 – Neighborhood structures with radius of 1 are cells having coordinates differing by 1 at most. a) von Neumann. b) Moore. c) Margolus

The process simulation of dynamic evolution represents a distributed computation across a defined space [91, 108]. In a directed network graph $G = (V, E)$, a distributed system is described as the computing elements allocated in the nodes of the graph. Each node $i \in V$ is a process in the distributed system. Each process has a set of states and a state transition function. Processes work synchronously, using blocking barriers where nodes emit messages and wait for neighbor messages. Communication is naturally represented by message passing systems, where messages are sent and received using queues. Associated with each edge $(i, j) \in E$, there is a channel holding at most a single message between the two processes i and j.

3.1.2 Related Works on Regular and Irregular Cell Spaces

Regular Lattice

In biological systems, CA models have been used to simulate a variety of phenomena, including the growth and spread of cancer cells [41] and the dynamics of bacterial colonies [46]. CA models are used to explore the behavior of these systems under different conditions and to investigate how changes in the underlying rules can affect the behavior of the system. Miller Julian proposed a method to create multicellular organisms of arbitrary size and characteristics [111]. The evolution of cell genotype will organize an organism into well defined patterns of differentiated cell types. The chemical can be taken into account as energy for the growth and death of cells. This is useful in constructing a more open ended evolution, where multicellular organisms compete for survival, linking morphology and behavior.

In the field of urban growth simulation, CA models get more attention from researchers since Wolfram showed that these simple systems can generate complex structures for exploring a wide range of fundamental theoretical issues in dynamics and evolution. White and Engelen [165] developed a CA model for urban land-use dynamics, addressing the issue of complexity in urban structure. The automaton is developed on a 50 x 50 grid of cells with a von Neumann neighborhood within a radius of 7 cells. Each cell has a state of vacant, housing, industrial, or commercial, and a state change is determined exogenously by applying growth rates for each urban function. Al-Ahmadi et al. [2] presented a calibration procedure within the fuzzy cellular urban growth model of 20m cells. The development suitability was a fuzzy function of a set of factors to determine the optimal combination that would result in the best performing model. They provided two scenarios

with a genetic algorithm and parallel simulated annealing to test the calibration performance. Zhang Yihan et al. proposed an approach using an ensemble Kalman filter [1] to estimate land-use changes by combining remote-sensing observations with urban simulation [98, 180]. They used maximum likelihood classification and artificial neural networks to obtain land-use classes and then used CA to simulate land-use changes for obtaining the process context information of an urban system.

One of the main research topics of CA is modeling real life phenomena. It is the ability to predict dynamic phenomena with an initial configuration and the transition rules to describe a long-term behavior. Real-life problems could be simulated using cellular automata such as fires, oil spills, and floods. The spreading of forest fires is predicted in a physical landscape under various weather conditions, in which wind speed and direction are the most important factors. Karafyllidis and Thanailakis [79] presented a model for fire spreading prediction that can determine the fire fronts in both homogeneous and inhomogeneous forests and can easily incorporate weather conditions and land topography. They provided transition rules for basic models of homogeneous/inhomogeneous forests and incorporated them with wind/elevation conditions. The experiments were done in a grid of 25 x 25 cells, 50 x 50 cells, and 100 x 100 cells. Yassemi et al. [176] developed a GIS-based CA modeling tool to simulate forest fire with a Moore neighborhood. The state of each cell ranges from 0 (unburned) to 1 (burned) a long continuous scale representing the ratio of the burning area to the total cell area. The transition rule was based on the assumption that fire spreads from one cell to another when it is fully burned and the fire travels according to speed and direction. Zhang Yihan et al. presented oil spill simulations based on the logistic-regression CA model [179, 182] and artificial neural network CA model [181]. The parameters in oil spills included proximity variables, wind, current, salinity, and temperature. The weights of these parameters were obtained by using logistic regression showing that distance is the most important factor, followed by currents, wind, salinity, and temperature. Tuyen Truong et al. [157] described simulation techniques based on geo-localized cellular systems with a case study of heavy rainfall simulation on complex terrain. They used a von Neumann neighborhood with water passing from cell to cell according to elevation differences. The water propagation is stopped or modified by ground obstacles such as hills, valleys, etc. The experimental zone is comprised of 58,275 cells corresponding to an area of 3 x 3 km in Morlaix, France. Thao Truong et al. [155] proposed a flood simulation model using the CA approach with a message

1. https://www.kalmanfilter.net/default.aspx accessed on 19 December 2022

passing interface framework. An experimental zone was decomposed into an organization of 26,214,400 cells representing 76m x 76m for each cell. The authors partitioned data to run parallel on multiple computing nodes with multithreads computing techniques.

Irregular Lattice

The cell space used in the above studies is regular cell space (mostly square shape) with all cells having the same shape and size. In an irregular lattice, the regular cell space is modified to be suitable for certain problems. David O'Sullivan [119] presented an approach to irregular CA models with a relaxation of the neighborhood definitions. A neighborhood is not necessary to be immediately adjacent to the cell, instead, it can be defined as a distance criterion. He introduced the concept of graph CA, in which the neighborhood is a graph of adjacency relations. The CA formalism consisted of a graph where each cell is a vertex, and its neighborhood is an edge with a constraint that each edge consists of an unordered pair of vertices. Any lattice can be regarded as a graph giving a convenient way of thinking about and investigating the implications of CA models that run on irregular lattices. Stevens and Dragićević [146] used cadastral parcels as irregular cell systems to model land-use changes at the land parcel scale. The use of a regular raster grid of CA models creates areas of assumed homogeneous land use that may contain variability in reality. Instead, land parcels avoid the problems associated with raster cells misrepresenting boundaries at a low resolution or requiring multiple cells to represent one discrete unit at a high resolution. The use of cadastral parcels can complicate the definition of neighborhood. The authors proposed three irregular neighborhoods suited to land parcel proximity functions: the adjacent neighborhood, the distance neighborhood, and the clipped distance neighborhood. The adjacent neighborhood contained all parcels sharing a point or a line with it (likened to the traditional Moore neighborhood). The distance neighborhood contained all parcels within a certain distance with its border. This can be a parcel that resided entirely within the distance limit or a parcel that resided partially within the distance limit. The clipped distance neighborhood returned only the portion of the outer parcels that lie within the specified distance.

Moreno et al. [116] introduced a dynamic neighborhood where the neighbors are defined during the modeling process and may vary for each object at each time step. They incorporated the concept of distance decay to minimize the impact of neighborhood size on the simulation outcomes. The neighborhood is the whole geographic space, in which object A is a neighbor of B if they are adjacent or separated by other objects whose

states are favorable to the change of state from B to A. Khila Dahal and Edwin Chow [30] presented a comprehensive discussion of neighborhood definitions that are possible in irregular CA for simulating urban growth. The two main concepts for neighborhood definitions are adjacency and distance. Based on adjacency relations, the neighborhood is defined as neighbors sharing points or lines in common or sharing line segments of at least a prescribed proportion of the focal polygon. Based on a specific distance, the neighborhood is defined in two ways using boundaries or centroids. A boundary neighborhood includes all polygons that intersect a buffer of a specific distance from the focal object and a centroid neighborhood includes all polygons whose centroids are completely contained in the buffer. There is also a possibility of considering the neighbors divided by topographic barriers such as major highways, rivers, faults, gorges, and hills. Another case of neighborhood based on distance is in the weights of influence. The influence of surrounding polygons of the focal object is optimum within close proximity and fades linearly in the outward direction.

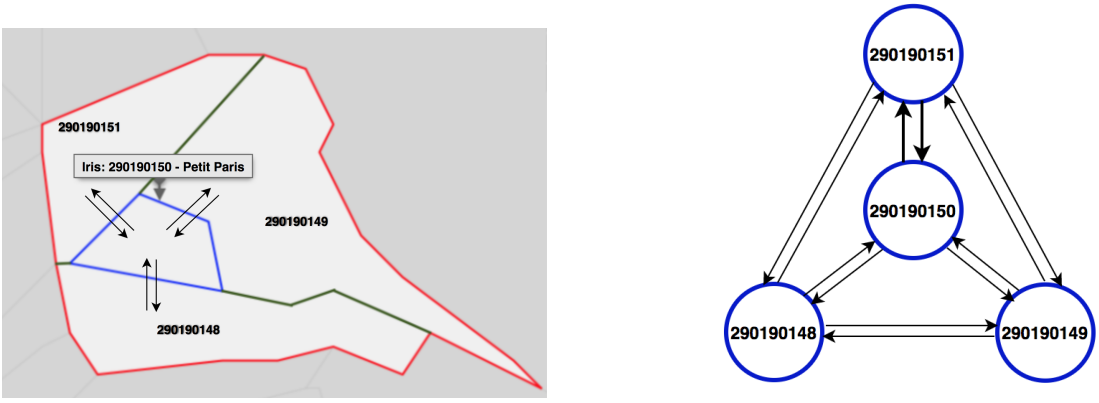
3.2 Irregular Cell Space with Geographic Divisions

The principle of using geographic divisions as a type of cell space in distributed processing systems is to collect precise data belonging to the territories for spatial and temporal simulation. IRIS is the smallest level of administrative division in France used to prepare for the dissemination of the population census. This type of spatial representation is more natural for human knowledge when trying to understand a phenomenon over space. Based on CA, a process system contains four main elements as described in section 3.1.

3.2.1 Process System Generation

The IRIS divisions with a variety of shapes and sizes are used as a set of cells in this work. Each division is a process in a simulation system that is bound by a set of input data, the initial conditions or initial states of a process. Each process interacts with other processes within a local adjacent neighborhood. The adjacent neighbors contain all divisions sharing a point or a line with the central division. Given a set of IRIS (Figure 3.3a), the neighborhood represents a set of the closest surrounding IRIS. Based on its current state, the center process receives/sends its information from/to the neighbor processes and changes its state under the global impact with a transition rule. A processing

system from these IRIS is generated as shown in Figure 3.3b. As an example, the central process in Figure 3.1 is a specific place in Brest, France, namely Petit Paris (Figure 3.3a), and the neighbor processes are the surrounding IRIS.



(a) (b)

Figure 3.3 – Neighborhood structures. a) A polygon with its neighbors; each polygon has an identification code. b) Process communication via channels between the polygons. Each polygon is bound to a process and each influence with its neighbors such as epidemic and pollution is carried over channels between the polygons.

QuickMap/PickCell/NetGen is a set of tools developed at the University of Brest, France with the initial purpose of designing and developing software for environment modeling and simulation using graphical interfaces. QuickMap [106] supports standard map tiles, including Open Street Map, and a variety of other items, either for maps or aerial images. The tool allows one to zoom and pan to select specific locations and PickCell generates regular cell systems with Moore/von Neumann neighborhoods in the locations. This has been done for environment simulation in previous studies in the case of sound propagation [82], brown plan-hopper surveillance [90], and flash flood [155, 157]. Figure 3.4 shows an example of a regular representation of a cell system. Each cell in the grid of raster cells represents a 287-meter square and communicates with its neighbors in pre-defined directions, for example, North, East, South, and West.

In this work, a new module, namely PickShape, is introduced based on the principles of PickCell to generate distributed processing systems using geographic divisions. It allows to read geographical data in shapefile format and visualize a graphical window for zone selection as shown in Figure 3.5a. Cell segmentation is obtained on each division and channels between cells are computed based on their neighbors (see section 2.2.2 for

neighborhood calculation with irregular cell systems). Precise information related to each division is bound to each cell using its identification. The data are stored in shapefiles or local databases and then queried in the system generation process. After loading the shapefile and zooming in to an area, the geo-coordinates of the bounding box of that area are used to query into the local database to take all divisions in the area for processes and channel generation. The generated system consists of all divisions partially or fully contained in the selected zone (Figure 3.5b). Occam/C code is automatically generated for simulation on CPU and high-performance simulation on GPU.

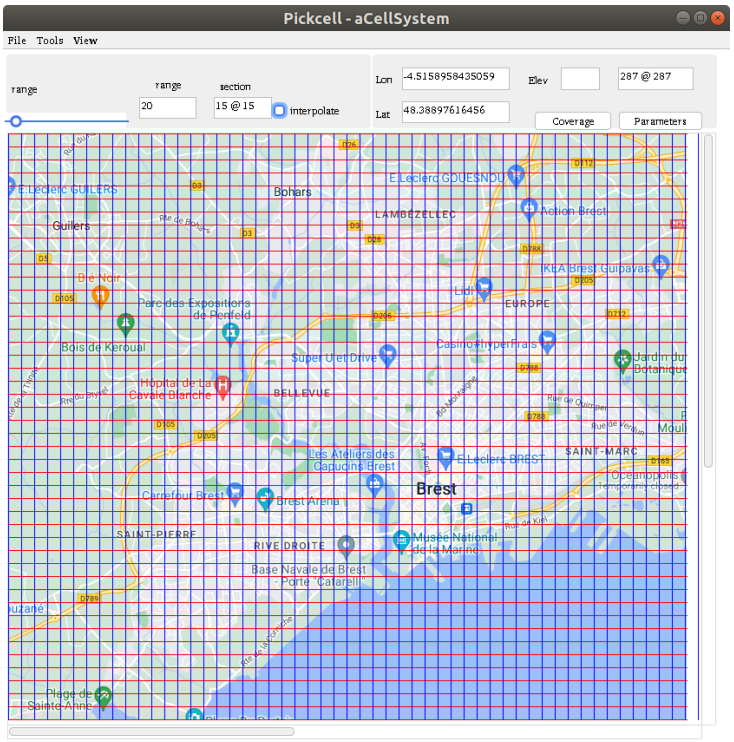
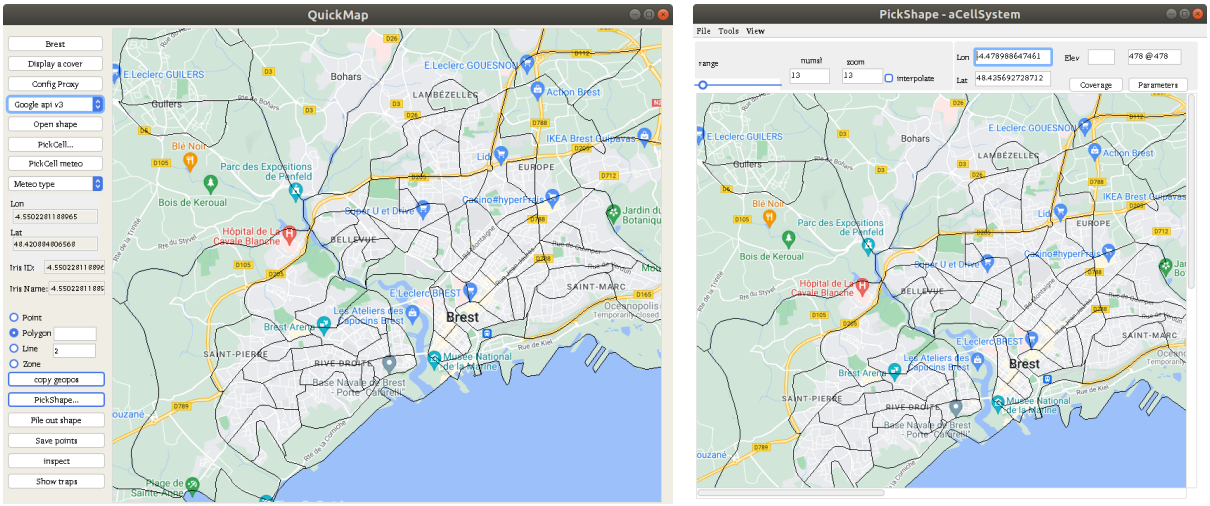


Figure 3.4 – PickCell tool with a map overlapped by a grid. Each cell in the grid ranges according to 15 x 15 pixels, and the corresponding actual cell size is presented in the figure top-right as a 287 meter square.

3.2.2 Data Binding

The data bound to geographic divisions are more precise than data estimation for regular grid cells. Let's consider binding population density into raster cells and geographic divisions in South France. The gridded population density per km^2 of South France is presented in Figure 3.6a. Each square cell is identified by top-left latitude-longitude co-

ordinates and a cell size of $4,800m^2$. The estimated population [2] is calculated from the growth rates based on a global comparison between the countries. It provides a global distribution of population on a continuous surface. The service is available through several protocols supporting direct data requests with proper syntax. The France population density per km^2 by IRIS is presented in Fig. 3.6b. By 2019, Metropolitan France was divided into 48,590 IRIS, of which each IRIS has an average area of 11.3 km^2 and an average population of 1,323 people. Each IRIS has a clear geographical boundary that is bound by a set of local data. The population density [73] is bound to each IRIS process through an identification code representing the correct local data compared with the gridded population density. We compare the gridded population density and the IRIS population density by 274 top-left latitude-longitude coordinates of the gridded cells that have data (Fig. 3.6a). Only 96/274 coordinates (35%) have the correct data thus the estimation error in gridded population density is quite large.



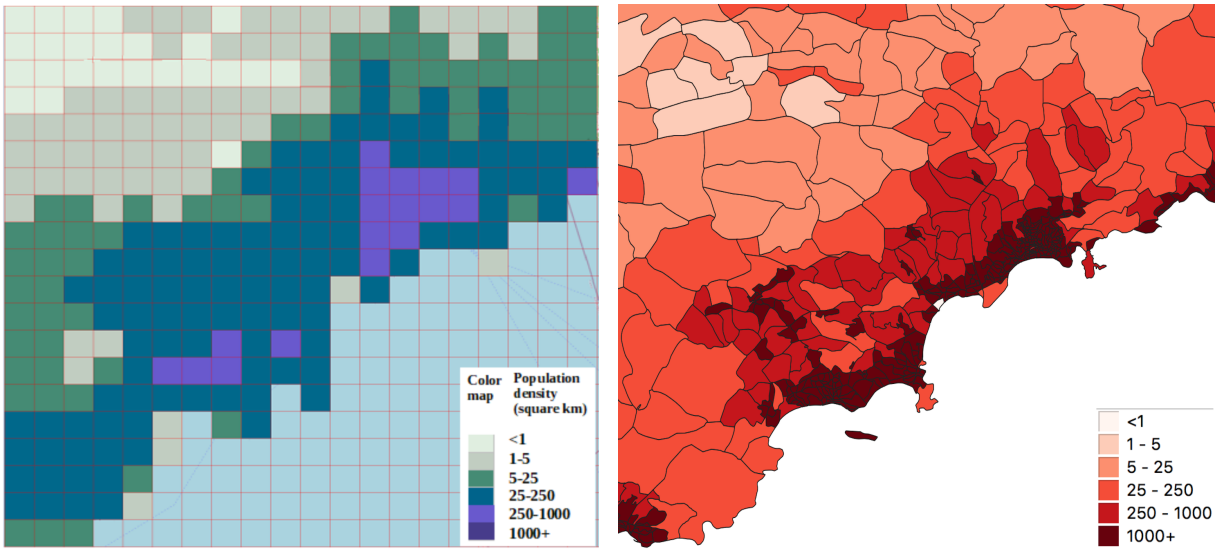
(a) (b)

Figure 3.5 – A cell system organization over geographical divisions: Divide the surface by population. a) Read geographical data and visualize on QuickMap allowing to zoom and pan for zone selection. b) PickShape allows to generate process systems with all divisions partially or fully contained in the selected zone. Each division is bound with a set of input data according to its identification.

A wide range of information can be passed into IRIS cells. As an example, besides population density, the economically active and inactive population is another demographic

2. https://sedac.ciesin.columbia.edu/data/collection/gpw-v3/maps/gallery/search

data that can be gathered into cells for research on environmental and human development. An active person is defined as a person who has a professional activity (employed) or a person who is looking for a job (unemployed). An inactive person is defined as a person who is not in the labor force, neither salaried nor unemployed. The number of active and inactive people aged 15 to 64 by IRIS is presented in Figure 3.7a and 3.7b. Meteorological data is another piece of information that can be passed into cells. Figure 3.11 shows the representation of temperature and humidity in Brittany, France. The information is important, especially in learning the connection of weather factors with disease propagation.

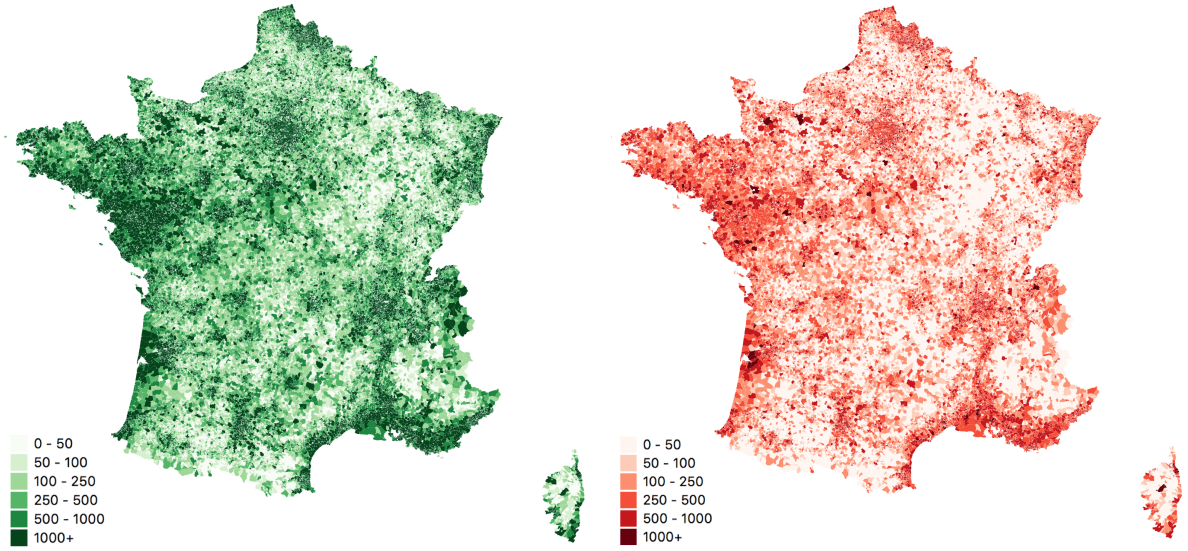


(a) (b)

Figure 3.6 – South France population density. a) population density in a grid of 400 raster cells provided by the Socioeconomic Data and Applications Center [154]. b) population density by 928 geographic divisions provided by the INSEE.

Mobility data is a piece of valuable information when modeling based on geographic divisions. Mobility data mainly means the information related to movements from one place to another. It can be the localization of cell phones provided by phone operators [8]. It can also be the number of incoming and outgoing travelers of a region according to the national census [14, 56]. Another way that may be possible is to collect the passengers' statistics from transportation companies. More simply, mobility data can be the transportation networks that represent the travel needs of people living in a region and a country. The intersection of geographic divisions with the transportation networks

provides information on the transportation routes of each division (Figure 2.2a). Besides, the mobility data can also be used as the channel for communication between cell processes. ROUTE500 [3] is the road database describing 500,000 km of classified network roads (motorways, national, and departmental roads) in Metropolitan France. It allows us to analyze statistical data and manage road trips.



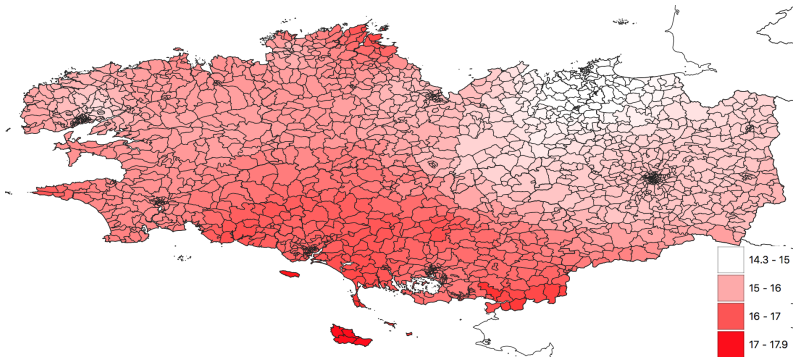
(a) Active people aged 15 to 64 (b) Inactive people aged 15 to 64

Figure 3.7 – France Metropolitan active and inactive population by IRIS divisions. a) Active people (employed or unemployed people) aged 15 to 64. b) Inactive people (people not in the labor force) aged 15 to 64.

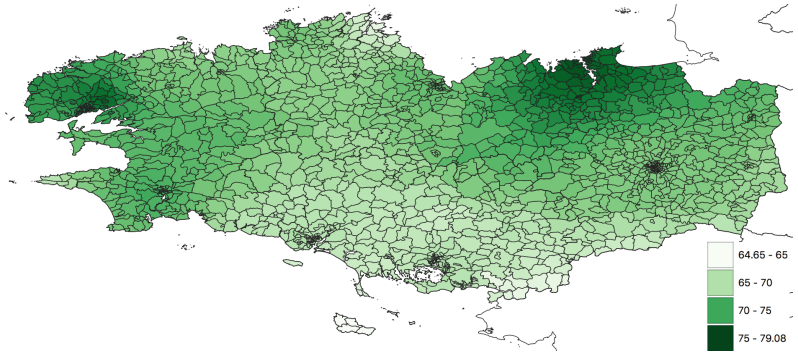
3.3 Parallel Computation with Irregular Cell Space

The process systems generated with QuickMap/PickShape are in C/Occam formats. C programs are sequential in nature. Generally, a sequence of statements is written in order to accomplish a specific activity. The parallel computing of the generated process systems can be achieved with additional techniques, including CUDA and MPI. In this section, we will present various ways to achieve parallel computing on process systems generated based on irregular cell spaces. The examples given here are to understand the main concepts associated with parallel computing. Real case studies will be given later in the thesis.

3. https://www.data.gouv.fr/en/datasets/route-500/



(a)



(b)

Figure 3.8 – Brittany France binding meteorological data into IRIS cells. a) Temperature in Celsius. b) Humidity percentage.

3.3.1 Occam

Occam [138] is known as one of the early parallel programming languages based on message-passing synchronization primitives. It enables a system to be described as a collection of processes, where the processes execute concurrently, and communicate with each other through channels. This gives the program a clearly defined and simple structure as well as allows it to exploit the performance of the system which consists of multiple parts. Occam-π is the common name for an Occam variant implemented by later versions of KRoC, the Kent Retargetable Occam Compiler [4]. It contains a significant number of extensions to the Occam 2.1 compiler. A set of primitive constructs facilitates handling not only processes in serial, SEQ(uency), but also in parallel, PAR(allel), as well as multiple events, ALT(ernate).

The PickShape system generates two files namely 'aCellSystem.occ' and 'aCellSystemData.occ' by default. The first file contains the general definitions of processes and channels as listed below. A process is defined by incoming channels, outgoing channels, an identification, and a channel to the center process (line 1). The incoming channels

4. http://projects.cs.kent.ac.uk/projects/kroc/trac/ accessed on 10 February 2023

`P01.in` represent an array of channels belonging to the first process `P01` (line2). Through these channels, the process `P01` will receive data from the processes `P39, P71, P69, P38, P73,` which are the neighbors of `P01`. The outgoing channels `P01.out` represent an array of channels belonging to `P01` for sending data to its neighbors (line 3). A simple channel such as `P01.P39` is defined to carry real values between processes (line 4). The `toMux[0]` is a channel connection between the current process to the central process for visualization purposes.

```
1        Node (P01.in, P01.out, 0, toMux[0])
2        P01.in IS [ P39.P01, P71.P01, P69.P01, P38.P01, P73.P01 ] :
3        P01.out IS [ P01.P39, P01.P71, P01.P69, P01.P38, P01.P73 ] :
4        CHAN OF REAL32 P01.P39, P01.P71, P01.P69, P01.P38, P01.P73 :
```

In addition, the first file also provides a set of variables as the initial state of each process that will evolve during the simulation. The variables are local data linked to a process such as number of population, area, and number of infected people by a disease. The second file is generated for the purpose of visualization. This file consists of data points representing the shape of each process in the evolution space. The data points provided in latitude-longitude coordinates are converted into Cartesian coordinates for visualizing on map figures.

The system generation by PickShape will not include a detailed definition of process operation. Thus, developers need to define the state changes over time steps, and communication between processes. A fragment of Occam code in Listing 3.1 describes the sending/receiving data and the updating local status in each process. Sending and receiving data will be done in parallel using different buffers ($bufIn[i]$ and $bufOut[i]$). Process status is updated after receiving data from its neighbors. An example of simulation results visualized by Occam code is presented in Figure 3.9. Each process represents a department in France and communicates with its adjacent neighbors to send/receive information. For visualizing purposes, the French map is divided into 8@8 pixels corresponding with 8,148 square cells generated by PickCell. These square cells with random colored points in each department show the density of that department.

Listing 3.1 – Sending and receiving data between processes in Occam.

```
1 PROC Node ( [] CHAN OF diam.proto in,out, VAL INT id,
2             [] CHAN OF diam.proto toMux)
3       //preparing data
```

```
 4       SEQ i=0 FOR SIZE out
 5         bufOut[i] :=  (diam.proto mydata)
 6       //send and receive data
 7       PAR
 8         PAR i=0 FOR SIZE in
 9           in[i] ? bufIn[i]
10         PAR i=0 FOR SIZE out
11           out[i] ! bufOut[i]
12       // updating data
13       SEQ j=0 FOR NLoop
14         SEQ
15           SEQ i=0 FOR SIZE in
16             SEQ
17               diff := REAL32 bufIn[i]
18               sum := sum + diff
19           current := current + diff
20 :
```

3.3.2 CUDA

Occam is a good start to show the parallel computation of process systems. The communication between processes is clearly defined in a simple structure. However, with the Kroc compiler, the number of possible simultaneous blocking calls is limited, currently 8,192 [10]. CUDA (Compute Unified Device Architecture) is considered to take advantage of graphics processing units (GPU). It is a parallel computing platform and an application programming interface that allows software to use certain types of GPU for general purpose processing. A GPU consists of thousands of cores to process parallel works effectively and it has a memory to provide fast access to essential data.

The processing system generated by PickShape is in C syntax, which is suitable to be included in CUDA programs [5]. Similar to Occam's case, the PickShape system generates two files namely 'aCellSystem.cu' and 'aCellSystemData.cu' by default. The former contains the general definitions of processes and channels and a set of initial variables as the

5. https://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf accessed 13 February 2023

initial state of each process. The latter consists of data points for process visualization in the evolving space. The C syntax gives more choices of supported libraries to display the simulation results compared with Occam.

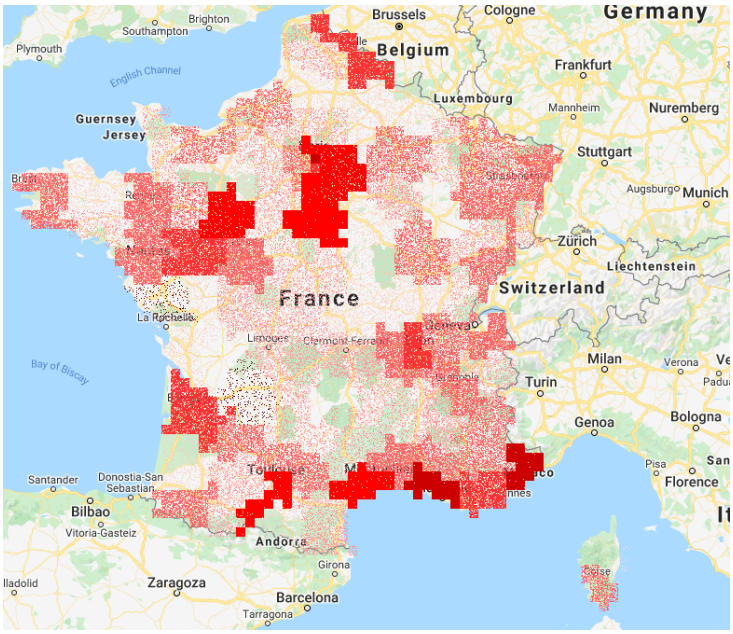


Figure 3.9 – An example of simulation results visualized by Occam code. Each process represents a department in France and the random colored points in square cells shows the density of a process.

The detailed description of process behaviors will be the responsibility of developers. They need to control the data flow between the host and device memory. In CUDA, the host refers to the CPU and its memory, while the device refers to the GPU and its memory. Thus, we must declare the buffers that will be located in the device memory (*nowState_device* and *nextState_device*). The data are copied from host to device for computation using the *cudaMemcpyHostToDevice* indication and copied from device to host for visualization using the *cudaMemcpyDeviceToHost* indication.

```
cudaMemcpy(nowState_device,nowState_host,stateSize,cudaMemcpyHostToDevice);
cudaMemcpy(nextState_host,nextState_device,stateSize,cudaMemcpyDeviceToHost);
```

An example of simulation results visualized by CUDA code is presented in Figure 3.10. Each process represents an IRIS in Brittany, France and communicates with its adjacent neighbors to send/receive information. The colored polygons show the density of a process. One obstacle we encountered when using CUDA for simulation was the

memory capacity of the graphics card. The card used in our lab is an NVIDIA GeForce GTX1070 card consisting of 1,920 cores with 256MB memory. This is not enough for handling simulations that need more processes and more data representing the status of each process.

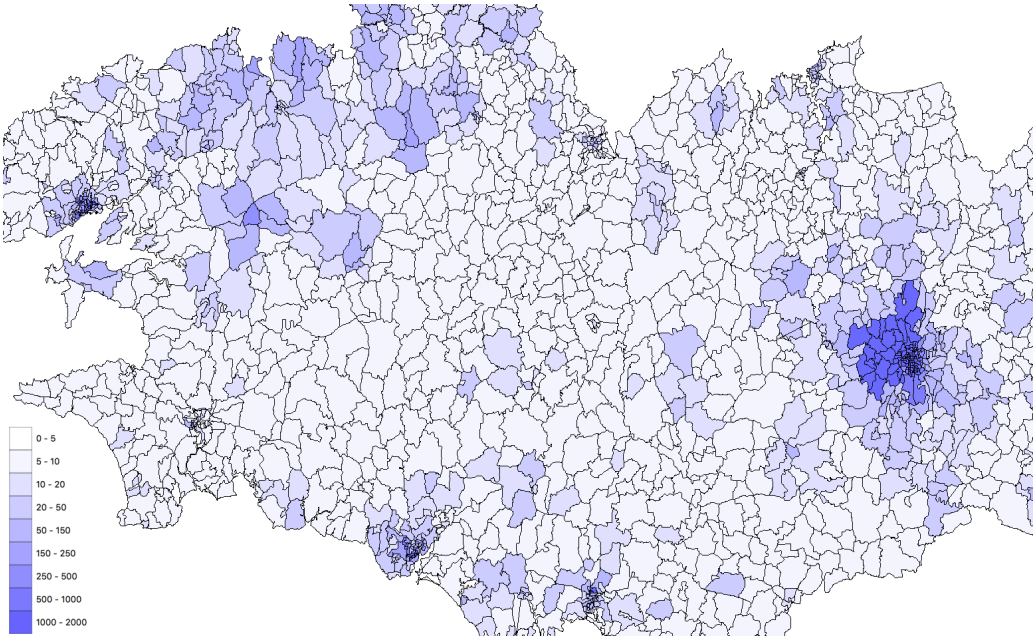


Figure 3.10 – An example of simulation results visualized by CUDA code. Each process represents an IRIS in Brittany, France and the colored polygons shows the density of a process.

3.3.3 MPI and Multithreads

MPI is a standard specification of the message passing interface for parallel computation in distributed systems. Multiple processes can concurrently run on separate nodes of clusters and communicate with each other by passing messages to exchange data. Parallelism occurs when partitioning a program into smaller chunks and distributing those chunks among the computing processes. Two common implementations of MPI are MPICH [60] and OpenMPI [55]. In this part, we choose to use the word "computing process" to indicate processes operating under the MPI framework. A computing process can be responsible for the computation of many cells in a cell system and it can be distributed across different computing nodes. The cell system generated in C syntax can easily call MPI library functions. In our case, the IRIS system needs to be divided into

blocks of cells, which will be distributed in computing nodes of a cluster. Figure 3.11a shows an example of partitioning the IRIS in Finistère, Brittany into 7 different computing processes represented by different colors. It appears that the communication between computing processes is not the same as regular grids because of the IRIS dividing by boundaries. The communication is not only with the computing process before and after it, but can also be before and after two or even three computing processes.

In some cases, it is useful to combine multithreads and MPI, such that MPI is used to parallelize a program across different nodes but the intra-node parallelization is handled with multithreads. This can be particularly useful for applications whose performance requires efficient memory use. By reducing the number of MPI ranks in favor of more threads, shared memory can be used more efficiently and communication computational penalties may be reduced. A hybrid parallel implementation combines OpenMP [23] with OpenMPI is also a good consideration [48]. We tested our cell systems of hybrid models from OpenMP and OpenMPI with the communication between computing processes as shown in Figure 3.11b. This is the case MPI_THREAD_MULTIPLE thread safe level, where any thread may issue MPI calls and different threads may issue MPI calls at the same time. The communication between computing processes is performed by the two main functions, i.e., MPI_Send and MPI_Recv presented below. Each computing process sends the current state of a number of cells, which it is responsible for, to the center computing process. The center computing process receives data and updates the current state of the entire system. This center process is responsible for managing the state of the system and displaying the simulation results.

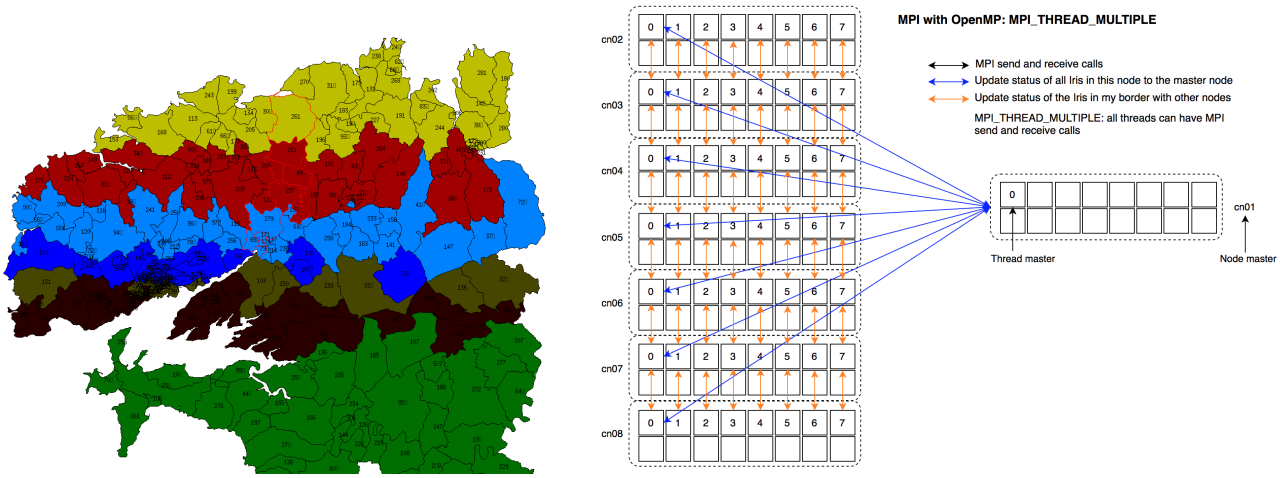
```
MPI_Send(subnowStates, nbcell, datatype, 0 , 0, MPI_COMM_WORLD);
MPI_Recv(nowState,nbcell,datatype,MPI_ANY_SOURCE,0,MPI_COMM_WORLD,&status);
```

One problem we encountered is that our IRIS system is not big enough as compared to the communication penalties. The simulation with 47,848 IRIS cells takes around 1 second per iteration (including writing images with the SDL2 library [6]). The processing time does not change much while changing from four to eight computing nodes and from 2 threads per node to 48 threads per node.

We also considered the execution time of multithreads with Pthreads [96] and OpenMP versus pure C programming in a simple simulation with the IRIS process system. These two models take a very different approach to allow parallel computing. Pthreads takes

6. https://wiki.libsdl.org/SDL2/Introduction accessed on 14 February 2023

a more low level approach to threading in the sense that it requires a more tailored program than OpenMP. The performance comparison between Pthreads and OpenMP is shown in [110, 183]. Given a new type of species appearing in the North of France, it can spread through the IRIS with a radius of 5 from the coastal line. A simulation is done with 323 iterations from North to South of France with 48 threads (Figure 3.12). The initialization time is around 60 milliseconds to calculate the IRIS with a radius of 5 from the coastal line. The execution time is 3,300 milliseconds without multithreads, about 240 milliseconds with Pthreads, and around 230 milliseconds with OpenMP.



(a) (b)

Figure 3.11 – IRIS partitioning into different processes. a) The partition is based on centroids latitude-longitude coordinates represented by different colors. b) Communication between processes in hybrid models combining OpenMPI and OpenMP. The thread safe level is MIP_THREAD_MULTIPLE, where any thread may issue MPI calls and that different threads may issue MPI calls at the same time.

Chapter summary: In this chapter, the CA theories have been presented with four main elements. We give an overview of the related works using CA models with regular and irregular lattices. The PickShape tool is introduced to generate process systems using geographic divisions as irregular cell space. This adds another option to the toolset QuickMap/PickCell with the purpose of developing software for environment modeling and simulation. We have shown the local data binding into cells and several works have been done to demonstrate the parallel computation with IRIS process systems. In the next chapter, we will present in more detail the use of process systems based on administrative divisions to monitor an epidemic spreading.

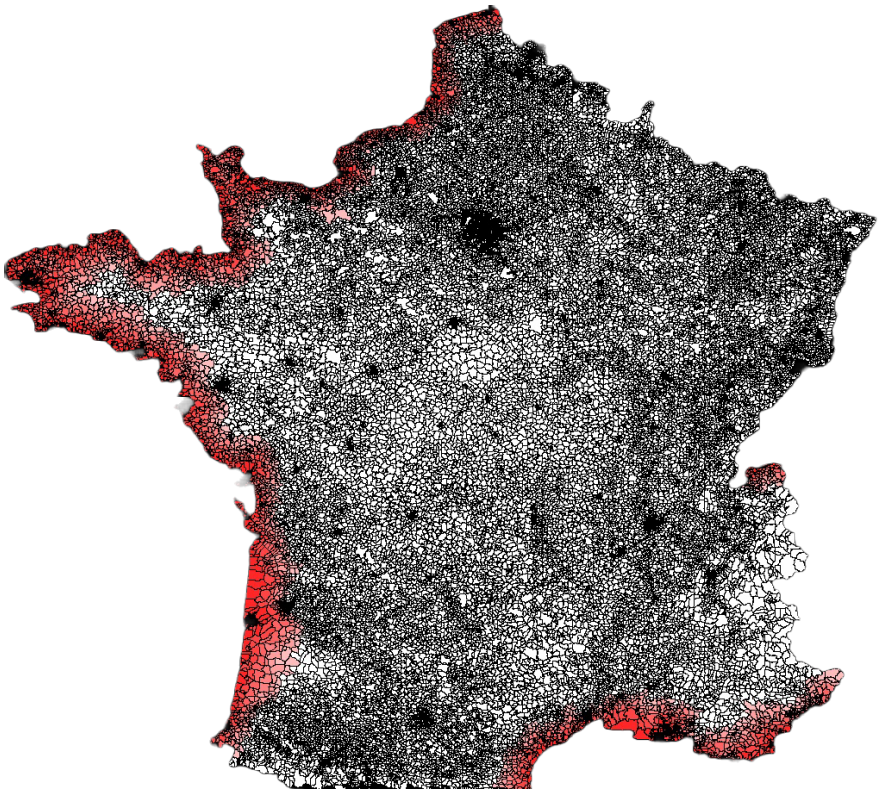


Figure 3.12 – IRIS with a radius of 5 from the coastal line in France. The more closer to the coastal, the more influence it can be affected by changes in shores. The initialization starts with IRIS location is 0 if the IRIS are in the coastal line. After 5 iterations, the IRIS with a radius of 5 from the coastal line can be calculated.

Chapter 4

Epidemic Spreading Control on Geographic Divisions

Chapter introduction: Epidemic spreading is still an attractive topic for public attention because of the regular occurrence of pandemics throughout history. Authorities collect health statistics based on geographic divisions and make it open for everyone to create potentially impacted tools. Understanding the spatial spread of the outbreak is critical to predicting and developing public health policies. In this chapter, we will explain how to model epidemic propagation based on geographic divisions and spreading control before applying measures. A cellular automata model is defined on this irregular cell space with the initial conditions acquired from Open Data repositories. The spreading process is performed by local exchanges in an adjacent neighborhood. The organization of the chapter is as follows. Section 4.1 discusses the metrics and spreading paradigm to model epidemic propagation. Section 4.2 presents the data analysis process and transition function synthesis for modeling the propagation of Covid-19. Section 4.3 presents the experiments with simulation results. Section 4.4 provides the related works with mathematical models and CA models in the epidemic examination.

4.1 Pandemics

In this section, we will have a brief look at the metrics and spreading paradigm to model epidemic propagation.

4.1.1 Metrics For Epidemiology

Control of contagion is a problem that regularly occurs throughout history. The Great Plague of Marseille arrived in France in 1720 [159]. From the original source of the infection, the epidemic spread to the surrounding localities. Attempts to stop the spread of

plague tried to separate Marseille and the rest of Provence with a plague wall and the remains of the wall can still be seen today. In the last few years, Covid-19 is one of the most dangerous diseases that threaten humans. The epidemic spread from one source to other communities and places far from the source due to the flights of many inhabitants. Before the availability of vaccines, some non-pharmaceutical interventions such as quarantines, school closures, and banning public gatherings were used to delay and flatten pandemic peaks.

Health authorities have established a counting policy based on administrative data collection with an antigen test or a real-time polymerase chain reaction test. Counting is done in medical entities and follows several criteria such as age, gender, vaccines, and medical history of patients. The common point is that every country is divided into smaller geographic divisions such as departments, districts, and communes for management and data collection related to the population. These divisions are bound by a set of data, e.g., demographic data, social measurement, meteorological data, and spatial dimension. Counting is done as government policies to measure the widespread of an epidemic. However, it is difficult to maintain a counting policy since it is impossible to take everything and everyone into account. Thus, we need to rely on statistics and models in order to estimate the transmission and propagation in the community.

The basic reproduction number (R_0) is defined as the average number of secondary cases generated by a primary case in a completely susceptible population [39]. This is a measure of the potential for disease spread in a population. Any factor having the potential to influence the contact rate, including population density and seasonality, will ultimately affect R_0 [34]. Because R_0 is a function of the effective contact rate, the value of R_0 is a function of human social behaviors and organizations, as well as the innate biological characteristics of particular pathogens. R_0 is dependent on population density and this dependence varies between countries. In [5], the number is estimated as $R_0 = N/N_t$ where N is the size or density of the host population and N_t is γ/β (immune capita rate divided by transmission coefficient). $1/\gamma$ represents the infectious period, which can be a few days, a few weeks, or a few months. Kate E. Jones et al. [76] studied the global trends in emerging infectious diseases and found that population density was a common significant predictor of emerging infectious disease events in all four categories including zoonotic pathogens from wildlife, zoonotic pathogens from non-wildlife, drug-resistant pathogens, and vector-borne pathogens. This supported their hypothesis that infectious disease emergence is a product of anthropogenic and demographic changes.

Karla Therese L. Sy et al. [150] showed the association between population density and the basic reproductive number even when the percentage of individuals that use private transportation and median income was accounted for. They concluded that a population density threshold of 22 people/km^2 was needed to sustain an outbreak in the United States. An increase in one unit of log population density increased R_0 by 0.16 likely due to increased contact rates in areas with greater density.

4.1.2 Spreading Paradigm

Spreading is a phenomenon where solid or abstract elements are moving, reproducing, or contracting in a defined space. Physical elements include gas, liquids, contagions, animals, or vegetal species. Abstract examples are knowledge, ideas, music, and languages. Fixing the topology of spreading effects is difficult because the mechanism cause is local while the effect is wide and general. It necessitates taking spatial context characteristics into account. The context variability requires a precise description expressed as a geography, as a solid, or as a living object geometry. A geographic spread phenomenon occurs in a variety of fields including biology, ecology, medical science, and species abundance.

Modeling a spreading effect necessitates representing space, elements' positions, or densities. Evolution rules and distributed influences will then describe the internal behavior. In the case of an epidemic, the evolution is measured as the number of infected cases, and the influence is its propagation, according to the transmission rate and the infection rate. The key point is to represent physical facts as program data, and then to process these data to represent what nature will do. Data models follow the CA approach with distributed algorithms showing the interactions between all cells. This allows the reproduction of physical evolution and dependency between cells, according to a transition function.

Entities appearing in a model represent real activities and values. The model itself is bound to real data to follow changes. The model definition related to CA and distributed processing systems is as follows.

— **Segmentation:** The space is segmented by geographic divisions represented as polygons on a map. A segment is a discrete interpretation of reality, which allows separating physical concerns and easing the description of behaviors. Each segment is a cell process in a distributed system. The set of cell processes can be represented as a matrix $C_{m,n}$, where m is the number of segments and n is the number of local parameters bounding to a segment (population, temperature, social measures, etc.).

$$C_{m,n} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix}$$

— A **neighborhood** of a segment is defined as the adjacent polygons containing all segments sharing a point or a line with it (likened to the traditional Moore neighborhood) [146]. Given a set of polygons, the adjacent neighborhood represents a set of closest surrounding polygons. The cell processes in distributed systems interact with others within a local neighborhood by exchanging messages that represent physical influences.

— In a simulation model, **influences** are represented by one-to-one messages carrying influence semantics such as epidemics. Message passing involves processes sending and receiving messages and a barrier is a synchronization point that requires all processes to reach before proceeding.

— Each segment holds a **state** representing a set of variables strictly isolated from its neighbors. A set of variables S specifies the initial state of a cell process and all cells change their states synchronously in discrete time steps based on influences from its neighbors. The evolution time is segmented according to a given clock or predictable events. In the case of an epidemic, the state is the density of infected cases and the evolution time is one day for capturing the changes of impacted cases.

— The interaction and local evolution are modeled by a **transition function**. This function receives the states of n neighbors one time step before determining the current state of a cell. Although the transition function is very simple, the interactions of all processes lead to global complex behaviors. This illustrates the complex behaviors of a system can arise from simple rules as in the Game of Life [12]. The evolution of a system is determined by an initial state and each cell changes states simultaneously over time according to the rules.

In epidemic modeling, the transition function is defined as in equation 4.1, where D_t is the density representing the infected cases at time step t, inf is the infection rate in a segment, s_{send} is the density sending to its neighbors, $s_{receive}$ is the density receiving from its neighbors, and r is the reduction rate.

$$D_{t+1} = D_t + D_t * inf + s_{receive} - s_{send} - D_t * r \tag{4.1}$$

The infection rate shows the increase of the density in a segment. This variable varies

depending on the actual situation of each segment as temperature, humidity, and population density. A high infection rate will cause the epidemic to quickly spread out to a broader area. The reduction rate shows the decrease of the density in a segment. This variable is defined by the government measures against the epidemic including social distancing and vaccination. The density D_t of the segment also changes depending on the interactions with its neighbors. Assuming that the epidemic is transmitted to its closest neighbors, the infected cases of a cell moving around and transmitting the epidemic to its neighbors but they still live in their place. Thus, in this case, the s_{send} is zero. The influence ($s_{receive}$) of the neighborhood is defined in equation 4.2, where n is the number of neighbors, D_{t_i} is the density of neighbor i at time step t, and trs_i is the transmission rate of neighbor i. The transmission rate may vary depending on the actual situation of the epidemic in the neighbors.

$$s_{receive} = \sum_{i=1}^{n} D_{t_i} * trs_i \tag{4.2}$$

4.2 From Parameters to Transition Rules: The Case of Covid-19

We provide a case study of Covid-19 spreading simulation using geographic divisions as distributed processing systems. A processing system has been generated in which each process is an IRIS in Brittany, France. The next important step is to define transition rules for epidemic propagation between processes. The Covid-19 incidence rate daily is provided by the French public health agency [1]. The epidemic incidence rate measures the total number of cases reported per 100,000 people. The parameters used to model the spreading of the Covid-19 pandemic are collected from a wide range of sources of different types including demographic data, the mobility data, weather data, and government measurements. These data are collected within a one-year period from May 2020 to May 2021. This work has been published in the proceedings of the 4$^{\text{th}}$ International Conference on Computers and Artificial Intelligence Technology (CAIT) (section 5.3 of the Appendix).

Demographic data are used to better understand diseases [92]. Getting to know the type of people who can be at risk with the Covid-19 pandemic helps to determine the pattern of the epidemic and identify the necessary intervention measures. The first con-

1. https://geodes.santepubliquefrance.fr accessed on 23 February 2023

sideration is the age distribution of the patients, which provides information on the age group that is more likely to be affected by the disease. The protection of this age group will become more pronounced in reducing the evolution of the disease. Gender can also be taken into account since there are diseases that usually occur in one gender rather than the other. Population density is a factor that cannot be neglected when considering any epidemic outbreaks. The higher the population density a city has, the more interactions it has that affect the propagation of epidemics. It is understandable that with more people living in a unit area, the connecting links between them are dense. This will affect the infection rate of the epidemics. Ruiqi Li et al. [97] investigated the relationship between population density and the death rate of influenza and pneumonia in the US. The study showed that large densities lead to high narrow peaks in death rate while small densities observe low and broad humps. The large population density during a pandemic can cause the hospitals to be overwhelmed so that the patients cannot have better treatments. French demographic data is provided by INSEE [73]. The data provided is in CSV format thus it is convenient for data analysis.

While population density can have an effect on the infection rate of a place, the mobility between places affects the transmission in different territories. In this work, mobility data are the transportation networks (bus, tram, metro, train, and road) that represent the travel needs of people living in a region and a country. These data are provided by OpenStreetMap in several formats, including CSV, JSON, KML, and Shapefile. In addition, learning the impact of temperature on the Covid-19 outbreak is one direction on the way of understanding the epidemic. Since most respiratory viruses are known to show seasonal infection, a number of studies have been done in this direction [6, 80, 132, 133]. It shows that the temperature plays an important role in the epidemic spreading and a moderately cool environment is the most favorable state. Warm places and regions have a lower risk of respiratory viruses. Cold and dry conditions can be the factors affecting the spread of the virus [109]. The emergence and replication of the virus have been shown to have a connection with weather factors. Thus, it is valuable to take into account meteorological data while modeling the Covid-19 outbreaks.

4.2.1 Selecting Parameters

There are different parameters that can be used to model the epidemic spreading. However, the more parameters are available, the harder transition rule synthesis is. We need to take into account all parameter changing effects on the propagation of the epidemic

and this is not easy. Thus, a data analysis process is done to analyze the dependencies between these parameters and the epidemic incidence rate. The aim of this process is to select important parameters for modeling the propagation of the epidemic. The general data analysis process is described in Figure 4.1. Data are collected from different sources on different days in different formats. The pre-processing stage will combine these data in a CSV format to use as the input for machine learning models. Then, we predict the Covid-19 incidence rate based on the input dataset with a random forest model. The model is a black box to human cognition since we cannot understand why these predictions are made and what impact these data have on the Covid-19 incidence rate. Thus, the next step is to use interpretation techniques to explain the relationship between the parameters and the incidence rate. The results are the important parameters having a major impact on modeling the epidemic propagation.

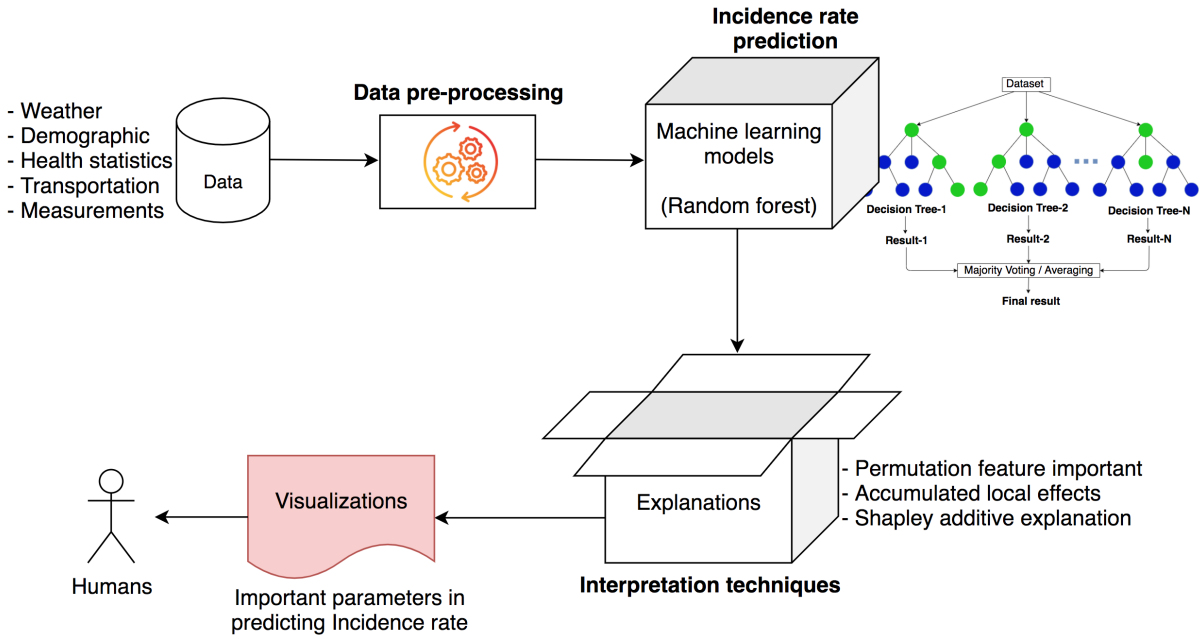


Figure 4.1 – Data analysis process. Data are analyzed with machine learning models and interpretation techniques to figure out the important parameters for epidemic propagation.

Machine learning model

A random forest model is fit [19] for interpreting the relations between parameters in the pandemic incidence rate predictions. It is a combination of a large number of classifiers $\{\mathbf{h(x,\Theta_k)}, \mathbf{k=1},\dots\}$ where the Θ_k are independent identically distributed random vectors that cast a unit vote at input x. The common prediction outcome of these decision trees

will be used as the final output. The decision trees are built on an approach to produce the purest tree nodes with impurity degree measuring methods, such as the Gini index used in classification and regression trees. Gini index [31] is calculated by the sum squared probabilities of each class from 1 (equation 4.3). The tools used for analyzing important parameters are mainly packages in R. The randomForest package [100] provides methods to create random forest models.

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2 \tag{4.3}$$

Interpretation techniques

Machine learning models are useful to provide predictions based on a set of data. However, these algorithms are mostly considered as a black box without explanations on predictions. The interpretable machine learning techniques [160] will help to gain insight into the available data, providing knowledge that is helpful to exploit the use of existing data. "Interpretability is the degree to which a human can understand the cause of a decision" [112]. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made [113]. There are several types of techniques for making machine learning models interpretable [42, 44], in which global techniques enable users to understand the entire model by its structures and parameters and local techniques examine the reasons for a specific prediction is made.

Permutation feature importance [19] is a global interpretation method that will shuffle the values of the features and measure the drop in performances. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature. An important feature is a feature that will increase the model error if permutes its values. The process of calculating permutation feature importance [50] is described as follows.

- Estimate the original model error (e_{orig}) (e.g., mean of absolute error or mean squared error)
- For each feature $j = 1, ..., p$ do:
  - Generate a feature matrix by permuting feature j
  - Estimate the model error (e_{perm}) after permute feature j
  - Calculate Feature Importance FI (e_{orig}/e_{perm} or $e_{perm} - e_{orig}$)
- Sort importance features by FI descending

Accumulated local effects [7, 113] is a global interpretation technique that describes

how features influence the predictions on average. The accumulated local effects plots average the changes in the predictions and accumulate them over the grid (the quantiles of the distribution of the feature). The method calculates the prediction differences conditional on features j and integrates the derivative over features j to estimate the feature's effect.

Shapley additive explanation (SHAP) [107] is a method to explain individual predictions based on the game theory Shapley values [140]. Shapley values assume that each feature is a player in a game and the prediction is the payout. For each feature j, it will evaluate the model of every combination of the features with and without j. SHAP comes with many global interpretation methods based on aggregations of Shapley values. SHAP feature importance [113] is an alternative to the permutation feature importance that is based on the magnitude of feature attributions. Features with large sum absolute Shape values are important. SHAP feature importance is calculated by formula 4.4, where $\Phi_j^{(i)}$ is the Shapley value of feature j at prediction i and n is the total number of features. The Shapley values are calculated as formula 4.5, where S is a subset of the features, $i(S)$ is the prediction i for feature values in set S.

$$FI_j = \sum_{i=1}^{n} |\Phi_j^{(i)}| \tag{4.4}$$

$$\Phi_j^{(i)} = \sum_{S \subseteq Nj} \frac{|S|!(n - |S| - 1)!}{n!}(i(S \cup j) - i(S)) \tag{4.5}$$

Implementation tools and results

IML [114] is an R package providing interpretability methods for machine learning models. The package allows to access a variety of methods such as Feature importance and Accumulated local effects. The SHAP methods are provided by package xgboost [25].

Permutation feature importance is used in this case to analyze data of the study region. The analysis shows that the important features affecting the predictions are temperature, humidity, and the number of days applied lockdown measure. The mean squared error is chosen to measure the loss in performance. Features associated with a model error increase by a factor of 1 meaning no change were considered not important for predicting the pandemic incidence rate. In Figure 4.2, the most important feature is temperature associated with an error increase of 1.72 after permutation.

SHAP summary plot (Figure 4.3) allows to analyze the contributions of the features.

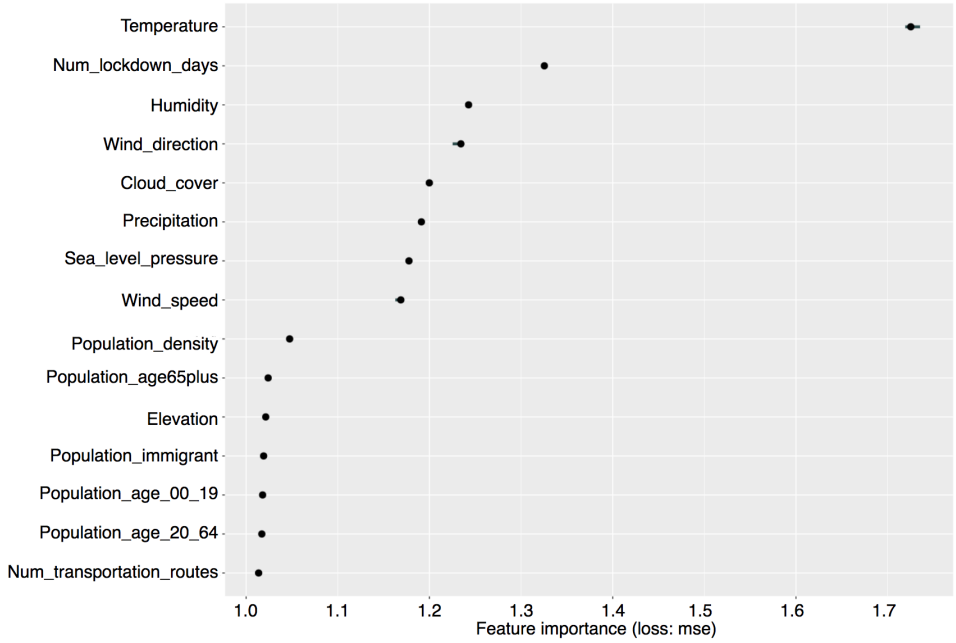


Figure 4.2 – Permutation feature importance. Temperature, humidity, and the number of days applied lockdown measure are the three most important features in predicting the incidence rate.

The color represents the values of the features from low to high. The features are ordered according to their importance with sum absolute Shape values.

Figure 4.4a shows the accumulated local effects for the feature 'temperature'. It shows how the incidence rate prediction changes with the temperature values. The average prediction falls with the increase in temperature and flattens above 19 degrees Celsius. For cool weather, the model predicts on average a high incidence rate. This ratio peaks when the temperature falls between 10 and 12 degrees Celsius. The incidence rate dramatically decreases when the weather gets warmer. The number of days that applied lockdown measurement also has a strong effect on the incidence rate prediction (Figure 4.4b). The average prediction falls with the increasing number of days applying lockdown measures. After 15 days, this measure has a negative effect: the higher the days of lockdown, the lower the prediction. When this number exceeds 30, the average prediction does not change much.

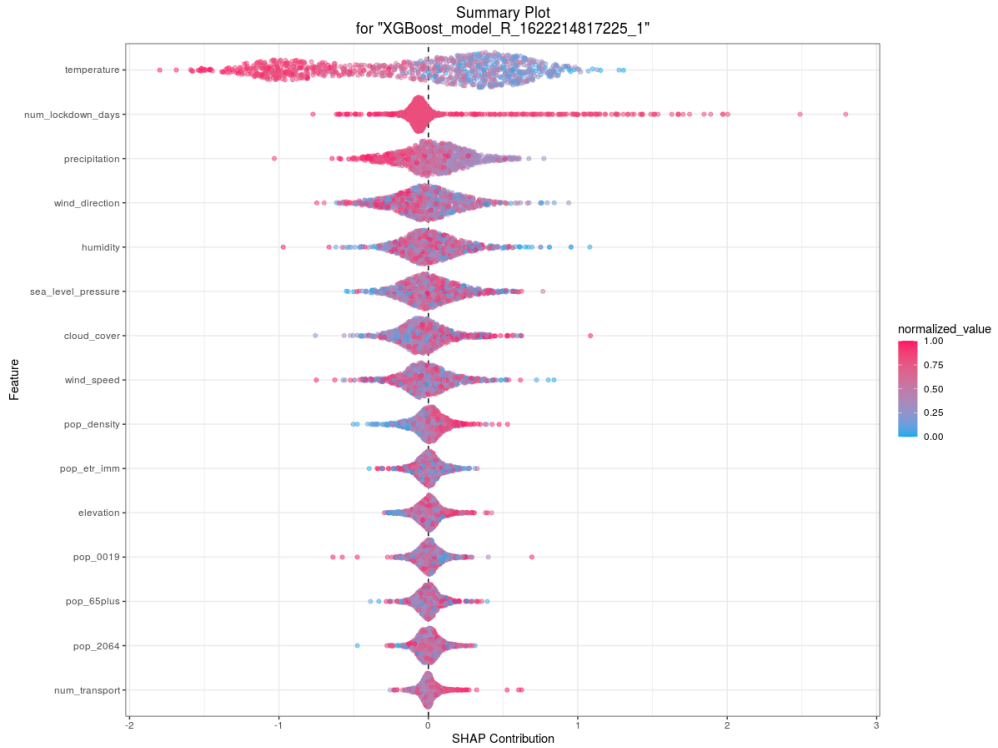


Figure 4.3 – SHAP values. Temperature, precipitation, and the number of days applied lockdown measure are the three most important features in predicting the incidence rate.

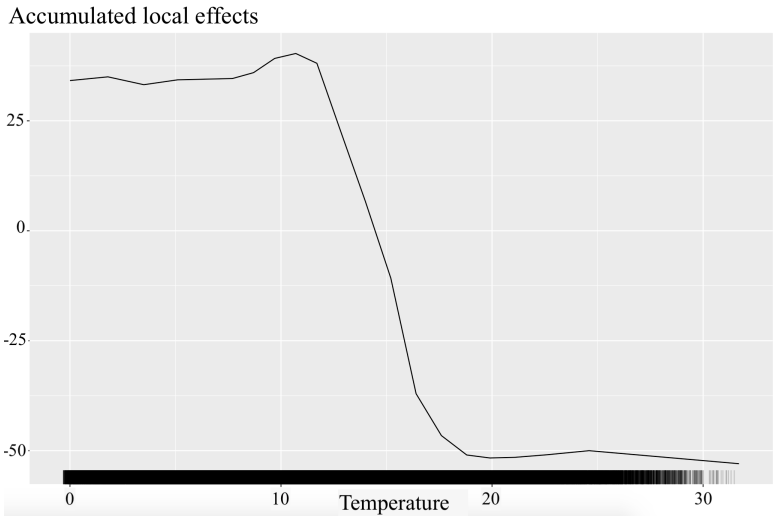
4.2.2 Transition Rules

The transition rules that provide state changes of cell processes are strongly dependent on domain knowledge. It is usually according to the intuitive understanding of the development process of the problems. Thus, transition rules analysis is a critical issue in a process system. Classification of parameters allows us to select the most important ones for modeling the pandemic spreading after the data analysis process. The temperature, humidity, previous day incidence rate, and the neighbors' incidence rate are the factors that will decide the infection and transmission rate of the pandemic. The lockdown measure and vaccination will affect the reduction rate.

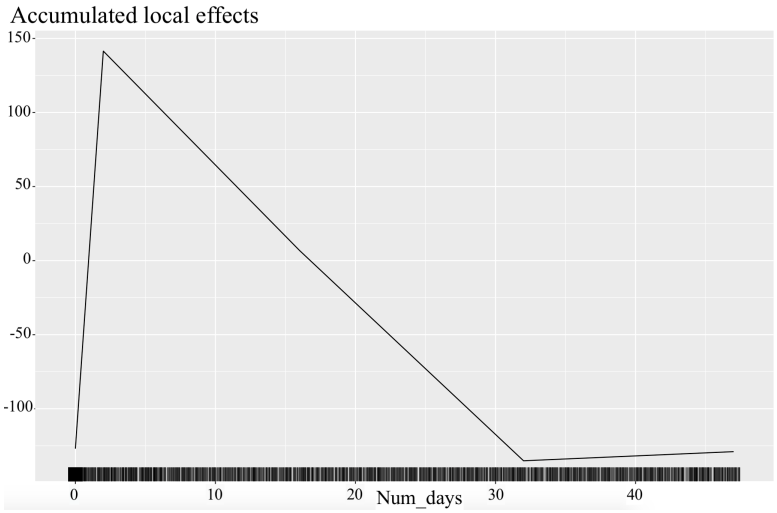
Poisson distribution [135] is a discrete probability distribution used to test the relevance of realistic stage-period distribution on the dynamics of epidemic outbreaks [64, 65, 87, 174]. Suppose some events occur λ times with an interval, the probability P of k times occurrences of the same event in the same interval is given by equation 4.6.

$$P(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{4.6}$$

Poisson regression models the dependency between the response and covariates by



(a)



(b)

Figure 4.4 – Accumulated local effects describe how temperature and lockdown measurement influence the incidence rate predictions on average. a) Accumulated local effects on temperature. b) Accumulated local effects on number of days applied lockdown measure.

Table 4.1 – Variables and explanations

Variable	**- Value**	**Variable**	**- Explanation**
β_1	0.005151	$temp_t$	temperature at time t
β_2	0.01403	$humid_t$	humidity at time t
β_3	0.00002795	$popu$	population density
β_4	-0.0512048	$lock_t$	number of days applied lockdown measure at time t
β_5	-0.0455449	$vaccin_t$	percent of vaccination/population at time t
β_6	0.0008964	nir_{j_t}	incidence rate of neighbor j at time t
		m	number of neighbors

assuming that the response y has a Poisson distribution. The dependency $\hat{y}$ is calculated as equation 4.7.

$$\hat{y} = e^{\lambda} \qquad with \qquad \lambda = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{4.7}$$

The remaining is to find the parameters $\beta_0, \beta_1, ..., \beta_n$ which maximize the possibility P by the maximum likelihood estimation.

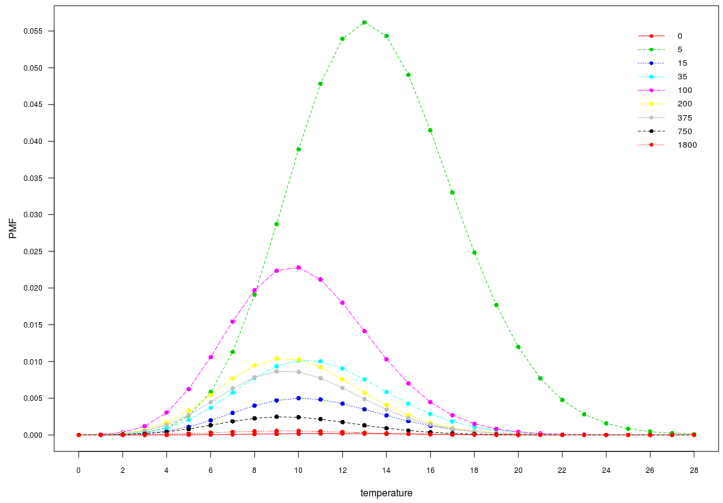
A Generalized Linear Model with Poisson distributions is fit to model the dependency between the pandemic incidence rate and the important parameters. This will provide the β values showing changes in the parameters affecting the incidence rate as in equation 4.8, where t is a time step, ir_t is the incidence rate at time step t, and δ_i is the dependency of parameter n on the incidence rate.

$$ir_{t+1} = ir_t + ir_t * e^{\lambda} \qquad and \qquad \lambda = \sum_{i=1}^{n} \delta_i \tag{4.8}$$

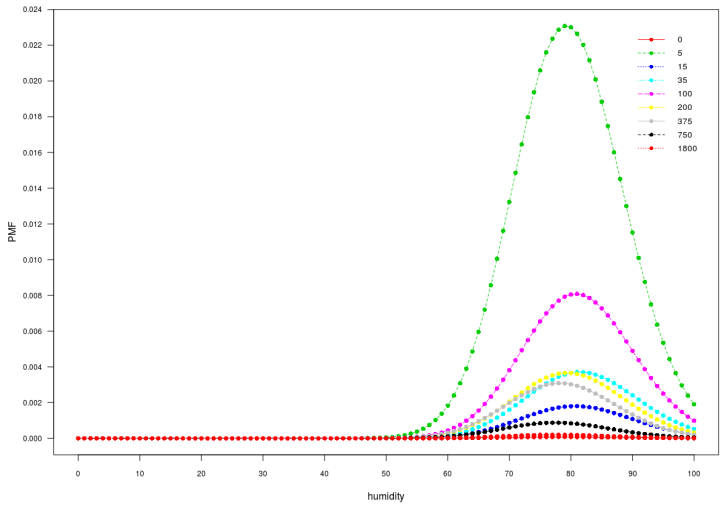
The general value of λ is calculated as equation 4.8. We analyze each parameter to deduce the value of δ_i. Figures 4.5a and 4.5b show the Poisson distribution of incidence rate based on temperature and humidity. It is observed that the probability mass function is expected to be highest when the temperature is around 10 degrees and humidity is around 80 percent. The rules are calculated with the variables as presented in Table 4.1.

Rule 1: δ_1

$$\delta_1 = \begin{cases} (temp_{t+1} - temp_t) * \beta_1; & \text{if } temp_t < 10 \text{ and } temp_{t+1} < 10 \\ (temp_t - temp_{t+1}) * \beta_1; & \text{if } temp_t >= 10 \text{ and } temp_{t+1} >= 10 \\ (10 - temp_t + 10 - temp_{t+1}) * \beta_1; & \text{otherwise} \end{cases}$$



(a)



(b)

Figure 4.5 – Poisson distribution of incidence rate based on (a) temperature and (b) humidity. The probability mass function is expected to be highest when the temperature is around 10 degrees and humidity is around 80 percent.

Rule 2: δ_2

$$\delta_2 = \begin{cases} (humid_{t+1} - humid_t) * \beta_2; & \text{if } humid_t < 80 \text{ and } humid_{t+1} < 80 \\ (humid_t - humid_{t+1}) * \beta_2; & \text{if } humid_t >= 80 \text{ and } humid_{t+1} >= 80 \\ (80 - humid_t + 80 - humid_{t+1}) * \beta_2; & \text{otherwise} \end{cases}$$

Assuming that when the lockdown measure is applied, the transmission between people in the population is trivial, and without any measurement, the transmission probability is around 10 percent. There is a moderate correlation between the incidence rate and the number of days applied lockdown measures. The incidence rate tends to decrease when there is an increase in the number of lockdown days. The Pearson correlation coefficient is a measure of the strength of a linear association between two variables. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables. The measured number is -0.36691, meaning that these two variables tend to lie on opposite sides of their respective means. The effect of lockdown measures, vaccination, and neighborhood on incidence rate is calculated as rule 4.

Rule 3: δ_3

$$\delta_3 = \begin{cases} 0; & \text{if } lock_t > 0 \\ 0.1 * popu * \beta_3; & \text{otherwise} \end{cases}$$

Rule 4: $\delta_4, \delta_5, \delta_6$

$$\delta_4 = lock_t * \beta_4;$$
$$\delta_5 = vaccin_t * \beta_5;$$
$$\delta_6 = \sum_{j=1}^{m} nir_{j_t} * \beta_6;$$

4.3 Experiments

In this section, we present the simulation results with the transition rules provided above. Other aspects are also considered: differences between geographic regions and tuning parameters in the transition rules to test the effectiveness of control measures.

4.3.1 Simulation Results

We simulate the epidemic propagation from September 2021 to May 2022. The Covid-19 incidence rate on 01 September 2021 is shown in Figure 4.6a. The places with higher incidence rates are mostly in urban areas with high population densities. Figure 4.6b

shows the simulation results of Covid-19 incidence rates after 60 days with the transition rules presented in section 4.2.2. The epidemic propagates from a cell to its neighbors shown by colored polygons in comparison between Figures 4.6a and 4.6b. We did not provide simulations over the whole of France because data collection and pre-processing are time-consuming. The simulation in this work is done in a specific region, but the general approach can be applied in any other places taking into account local available data.

Figure 4.7a shows historical data of the average incidence rate in time series. After the ease of restrictions in June and July 2020, during August 2020 incidence rate began to rise again. The average incidence rate continued to rise and the government decided to enter a second nationwide lockdown from 30 October 2020. This measure ends on 15 December 2020. The third national lockdown was proposed by the government starting on 5 April 2021 and lasted until 3 May 2021. Figure 4.7b shows simulation results of the average incidence rate. It shows that we can follow the trend of the average incidence rate. People send and receive viruses depending on population density, vaccination, lockdown measures, and weather conditions. Close contact helps virus transmission. Each process in the system has a state defining the number of infected people and updates its state by the transition rules considering the neighboring states. The simulation software computes epidemic spreading over irregular cell systems, allowing us to examine its evolution with transition rules synthesized from historical data. Although the global propagation trend is caught, it has to be noted that we cannot obtain high accuracy in the simulation because of the chaotic behaviors of the pandemic. It is observed in Figure 4.7b that the simulated incidence rates are generally less than the real data collected. Another limitation of the study is that we only consider the infected cases while ignoring the hospitalized, recovered, or dead cases. This can be one of the reasons for decreasing simulation accuracy. More figures showing the simulation results are available in the Appendix section 4.1.

4.3.2 Variation of Behaviors will Assist Control

Different regions will have different orders of important parameters. Figure 4.8 shows the first three important parameters in 3 different zones in Brittany, France including Brest, Carnac, and Landivisiau. We consider data in Brest (administrative & industry zone) with 27 IRIS having an average population density of 8,130 people/km^2. In Carnac (tourism zone), we take 9 IRIS having an average population density of 158 people/km^2. In Landivisiau (agriculture zone), we take 9 IRIS having an average population density of 593

people/km^2. The number of days applied lockdown measurement is the most important feature in places with high population density (Brest). In places with lower population density (Carnac, Landivisiau), temperature is the most important feature in predicting the epidemic incidence rate. The humidity and wind are also important features in these places. Thus, different behaviors in zones need to be taken into consideration to improve control measures. As an example, places, where lockdown measures are less important are candidates to reduce the number of lockdown days after checking the simulation.

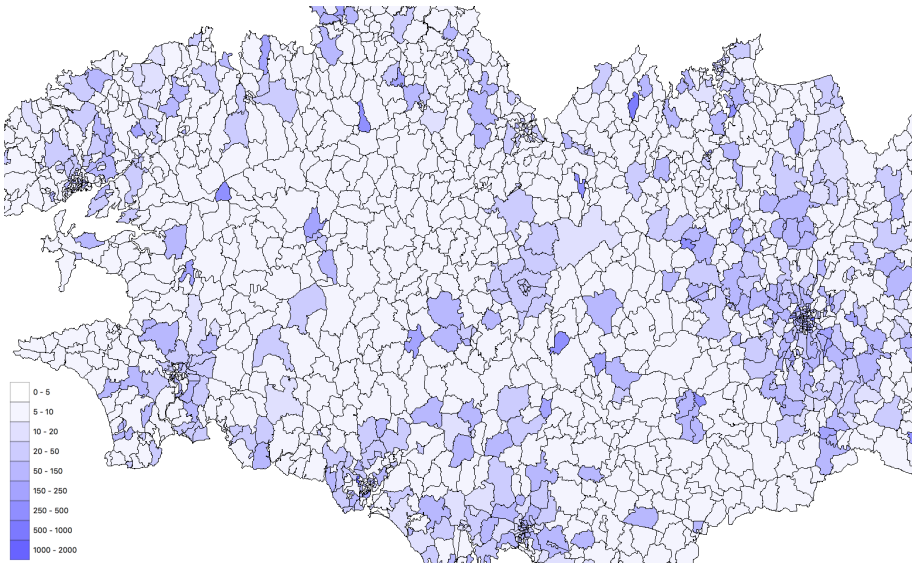
4.3.3 Epidemic Spreading Control

Figure 4.9a shows an illustration of an epidemic spreading from the first place highlighted with a red circle. We will discover the geometry of the epidemic spreading around the initial spot. Segments are presented on a map with color intensity representing the density changes and the points inside each segment represent the density (impacted cases). The increasing density is shown in red polygons, the decreasing density is shown in yellow polygons, and the white polygons are less affected places with density less than 5.

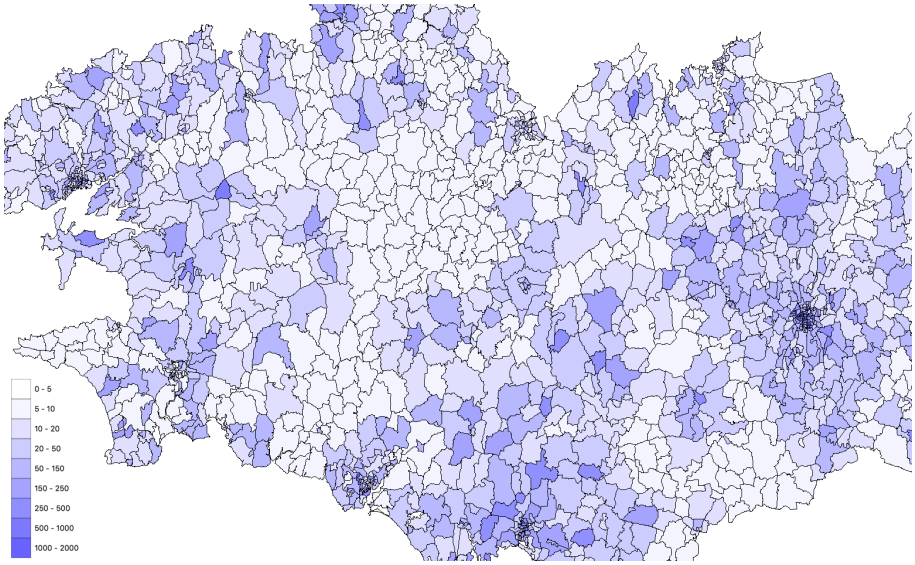
Figure 4.9b shows the epidemic spreading to the broader area after 90 rounds without any counteractions. It is noticed that, in some segments, the epidemic disappeared because herd immunity was achieved through previous infections. Not shown in the figures that the epidemic can oscillate between increasing and decreasing. Spreading could be managed given government open data and measures to counteract the epidemic. As an example, a barrier of vaccination or social distancing can be erected in an effort to slow down the epidemic spreading. In Figure 4.9c, a virtual wall in red line was built to control the spreading to the north of the region by tuning the infection rate and the transmission rate of the polygons in the red line. After 90 rounds, the north part of the region is less affected than the case without counteractions shown in Figure 4.9b.

4.3.4 Discussion

The important parameters pointed out and used in our model are similar to the results provided by several previous studies. Bastos and Cajueiro [11] analyzed epidemic data in Brazil with the SIR model to consider the effect of the social distance policy. It was shown that the policies of social distance can flatten the contamination. However, a short-term distancing policy can only shift the peak of infection into the future. This conclusion is close to our simulation results. When there is a lack of vaccination, social distancing can

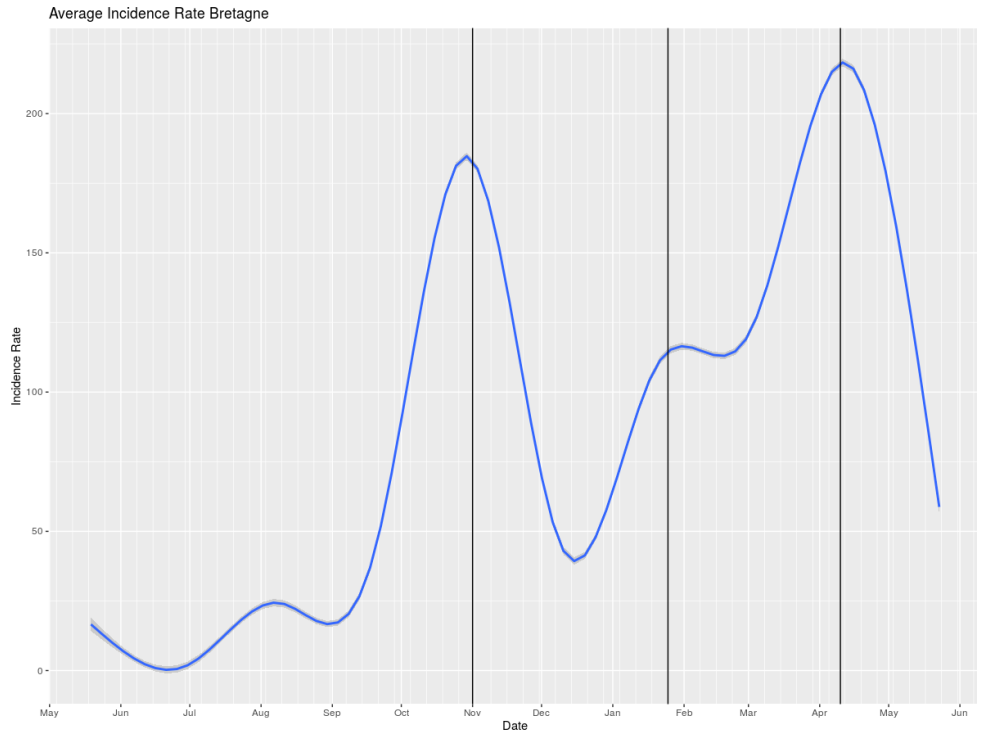


(a)

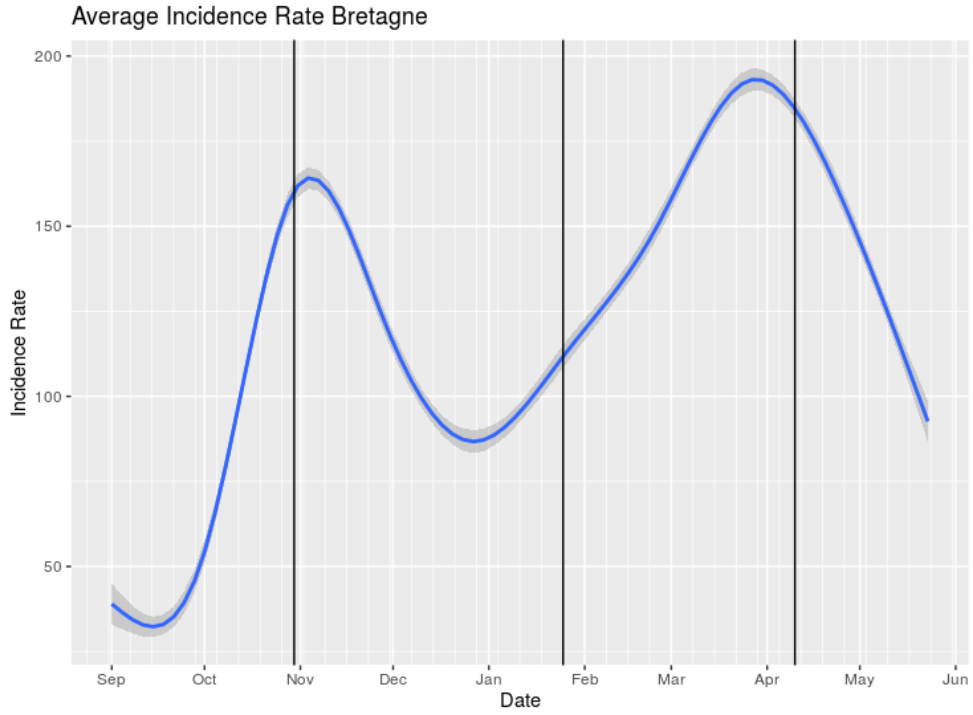


(b)

Figure 4.6 – Covid-19 incidence rate over time. a) Incidence rate in 01 September 2021. b) Simulation results after 60 days.

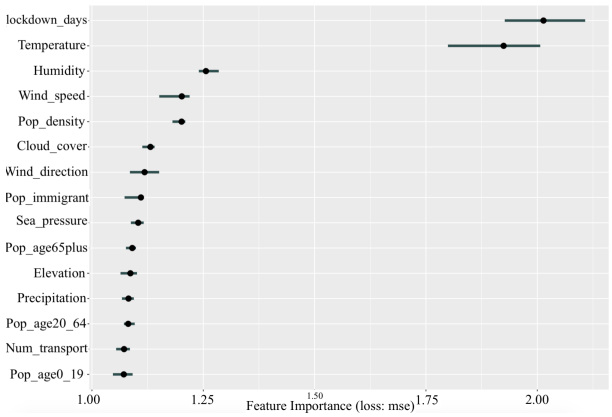


(a)

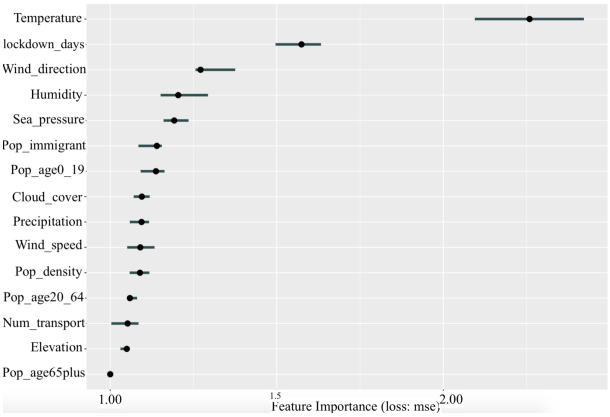


(b)

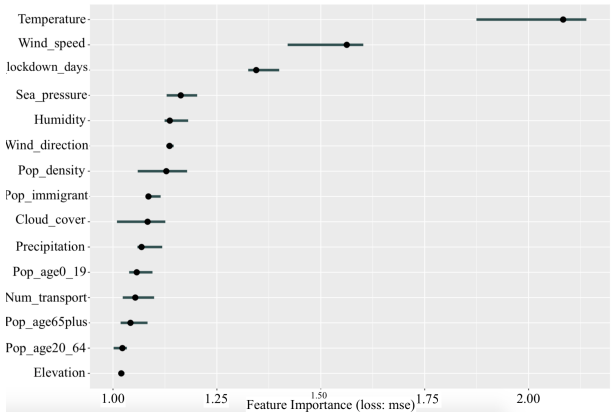
Figure 4.7 – The average incidence rate in time series. a) Historical data. b) Simulation results. The simulation results follow the trend of the average incidence rate in time series.



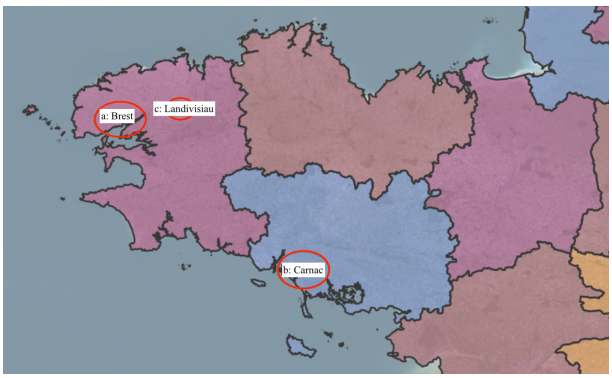
(a) Brest



(b) Carnac



(c) Landivisiau



(d) Zones in Brittany map

Figure 4.8 – Permutation feature importance in 3 different zones in Brittany.

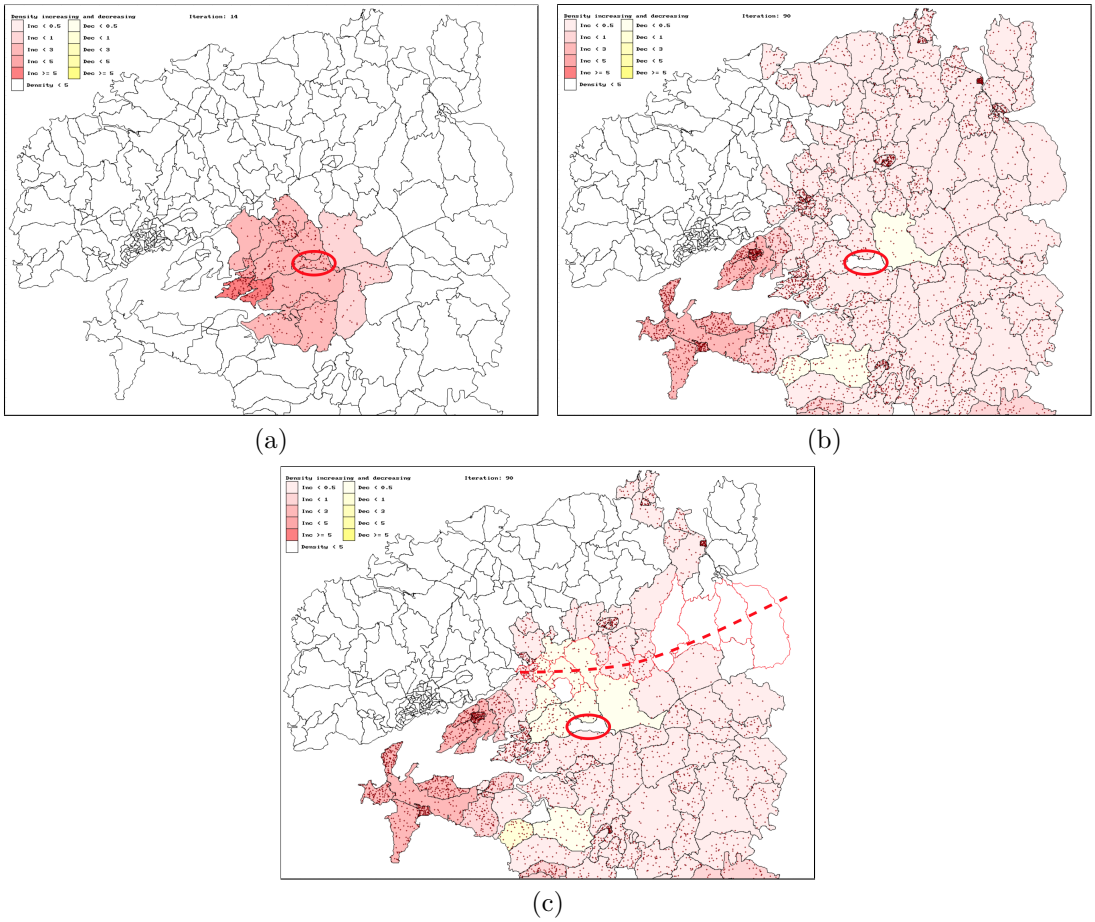


Figure 4.9 – Epidemic simulation with points represents the density and colored polygons represent the density changes. a) A source of contamination in red circle spreads to its surrounding neighbors after 14 days. b) The spreading after 90 days without control. c) The spreading after 90 days is controlled with vaccination and social distancing (polygons in the red line).

only move the epidemic peak to the future as shown in Fig. 4.7b (the epidemic peak ships from November 2020 to April 2021). The emergence and replication of the virus have been shown to have a connection with weather factors. Adly Anis [6] demonstrates that the temperature plays an important role in the epidemic spreading. He showed that cool weather is the most appropriate for virus activity and transmission. Aly Kassem [80] also showed the relationship between temperature and transmission speed of the Covid-19 virus. Most respiratory viruses are known to show seasonal infection [109]. Warm places and regions have a lower risk of respiratory viruses. Cold and dry conditions can be the factors affecting the spread of the virus. This is also one of the results shown in our study that two important factors affecting the spread of the virus are temperature and humidity.

In addition, we base our simulation on transmission between neighboring geographical areas assuming that people in neighboring areas will spread the disease to others. This is a result shown in Schimit Pedro's study [137] describing the transmission of Covid-19 with a square lattice of n x n cells. The local interactions were based on an extended Moore neighborhood with a radius affecting the probability of interacting between one cell and the others. Their results show that Covid-19 is likely to spread in densely populated regions and between geographically adjacent regions since people in these regions are more likely to interact with each other.

4.4 Related Works

In this section, we present the related works in the field of epidemic modeling. The mathematical models have been used for a long time considering the number of suspected, infected, and recovered people. Other extension models also consider the number of exposed and asymptomatic people. In addition, we provide a general context of the use of CA in epidemic modeling.

4.4.1 Mathematical Models

The most commonly used models for predicting the evolution of Covid-19 are SIR (Susceptible-Infected-Recovered) and A-SIR (Asymptomatic-Susceptible-Infected-Recovered). These models describe the evolution of diseases in a population split into classes: susceptible (S), exposed (E), infected (I), asymptomatic (A), and recovered (R) or death (D). Fanelli and Piazza [47] used the SIR model to analyze the Covid-19 data in China, Italy,

and France in a 23-day period from February 2020 to March 2020. It is indicated that the kinetic parameters describing the recovery rate seem to be the same over the three countries, while the infection and death rates appear to be different. This is likely to be connected with culture-related habits and underlying health conditions leading to a larger variability from one country to another. Bastos and Cajueiro [11] analyzed epidemic data provided by the Ministry of Health of Brazil from February 25, 2020 to March 30, 2020 to model the evolution of the Covid-19 pandemic. The authors modified the SIR model to consider the effect of the social distance policy imposed by the government. The common measure to respond to the pandemic at the moment is social distancing, including school closure, restricting commerce, and working on distance. It was shown that the policies of social distance can flatten the contamination. However, a short-term distancing policy can only shift the peak of infection into the future, keeping the value of the peak at almost the same value. Castilho et al. [22] provided an age-structured SEIR (Susceptible-Exposed-Infected-Recovered) model with quarantine policies. The population was split into three groups of different ages including youngsters, adults, and the elderly and applied 5 different quarantine strategies. They estimated the number of deaths in each age group for these quarantine strategies and showed that strategy 2 (stronger isolation of the elderly, twice as much as the other groups) was the best among these. They also argued that the isolation for only the elderly leads to two main problems: the quarantine effort would be too small to be significant due to the small percentage of the elderly class and this allows for a much higher number of infected individuals burdening the health system.

Lavezzo et al. [93] carried out two surveys in Vo', a small town in Italy, collecting nasopharyngeal swabs from 2,812 and 2,343 study participants, respectively. They showed that 42.5% of all confirmed cases across the two surveys were asymptomatic and among the confirmed cases, there are no significant differences in the frequency of asymptomatic infection between age groups. None of the children under 10 years old tested positive despite at least 13 of them living together with infected family members suggesting that children may be less susceptible than adults. Gudbjartsson et al. [62] carried out two strategies for Covid-19 testing in Iceland - targeted testing of persons at high risk for infection and population screening. They noticed that 43% of the tested positive participants reported having no symptoms, although symptoms almost certainly developed later in some of them. These patients can contaminate other people and greatly affect the modeling of the epidemic since they present no clinical symptoms. The results also showed that young children and females were less likely to test positive than adolescents

or adults and males.

The effect of spatial dimension is important for modeling disease outbreaks. Guan et al. [61] extracted data regarding 1,099 patients with laboratory-confirmed Covid-19 from 552 hospitals in China until January 29, 2020. They analyzed the distribution of patients by province, the characteristics between residents of Wuhan and non-residents, and the history of direct contact with wildlife and non-residents of Wuhan who visited the city or who had contact with the citizens there. This early study is important allowing the clinical characteristics of the affected patients to be more precisely defined. Kang et al. [77] explored the spatial epidemic dynamics of Covid-19 in mainland China using Moran's Index I spatial statistic with various definitions of neighbors. In models 1 and 2, spatial adjacency was defined by geographical information including border sharing and centroid distancing. In models 3 and 4, the ranking of population and population density was considered thus a province was defined as adjacent to both the previous and following ranked provinces. Models 5 and 6 considered the ranking of a number of doctors and medical centre beds for defining neighbors. The dataset for statistical analysis included newly confirmed cases, population, population density, number of licensed doctors, and health centre beds per 1,000 inhabitants in 31 provinces in mainland China from 16 January to 06 February 2020. The spatial dependency detected in the study implied the possibility that Covid-19 had spread from Wuhan to other areas via transportation. This could be linked with the government policy to close off Wuhan City on 23 January. Desjardins et al. [36] used a prospective space-time statistic to detect emerging Covid-19 clusters in the United States at the county level. The authors showed the ability of the approach to add updated counts and re-execute the statistics to identify new emerging clusters while tracking the previously detected clusters to determine if they are growing or shrinking in magnitude. The study results introduced several epidemic centers in the U.S. when adding the updated daily cases to the prospective scan statistic. The authors suggested that the counties belonging to emerging clusters should be prioritized when allocating resources and implementing various quarantine and isolation measures to slow viral transmission. Gatto et al. [56] modeled in space and time the countrywide spread of the epidemic using an extended version of SEIR model including mobility data of 107 provinces in Italy. They estimated the generalized reproduction number is 3.6 in the absence of containment interventions and suggested that the sequence of restrictions posed to mobility and human-to-human interactions have reduced transmission by 45%. The spatial nature of the model was helpful in terms of making different mobility restriction

policies for the authorities.

4.4.2 CA Models

Cellular automata are used to examine the spatial and temporal propagation of epidemics. Doran and Laffan [40] proposed SIR CA model to simulate the spatial dynamics of foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia. The cell system is a square lattice system implemented using the map algebra system in ArcInfo GRID version 8.2. The probability of interactions between the grid cells depends on the density of susceptible herds. The research showed that depending on the season the outbreak is initiated as wet or dry, the evolution results were completely different in the two regions assessed. Sirakoulis et al. [143] used the CA approach in order to study the effect of population movement on epidemic propagation. Each cell represented a part of the population, which may be found in one of three states: susceptible, infected, and recovered. The state of each cell started with susceptible and became infected if its neighborhood was infected and it remained infected for a defined period before becoming recovered. As the population moves randomly in the lattice, the disease spreads. Hoya White et al. [166] introduced a mathematical deterministic model based on CA, and three classes of the population were considered: susceptible, infected, and recovered. The cell space was a two-dimensional array of 50 x 50 cells standing for a square portion of the land. The cell state was obtained from the fraction of the number of individuals, which are susceptible, infected, or recovered from the disease. The authors considered two cases for simulations based on the connection of each cell with its neighborhoods as a constant or not. Alejandro Salcido [134] used a lattice gas approach to model the Covid-19 epidemic spreading. The model consisted of a regular lattice and the state at each lattice site indicated the presence or absence of the particles traveling in nine directions. The particles represented susceptible, exposed, infectious, hospitalized, recovered, and death cases. The particles that arrive at one site may collide with each other. The collisions between particles moving on the lattice will change from one species to another type of species according to a given transition probability. Schimit Pedro [137] described a SEIR model in terms of probabilistic CA for the transmission of Covid-19. The model consisted of a square lattice of n x n cells, which n is 14,500 in the case of the Brazilian population. Each cell represented one out of eight states including susceptible, exposed, infected, and recovered. The local interactions were based on an extended Moore neighborhood with a radius r affecting the probability of interacting between one cell and the others. The

simulation ran for 3,000 time steps, in which one time step was considered to be one day. The simulations would take two months to be completed on a regular PC with a 4GHz processor with 16 GB RAM. The authors made use of the Graphics Processing Unit to reduce the time to one week and a half. The results showed that Covid-19 is likely to spread in densely populated regions and between geographically adjacent regions since people in these regions are more likely to interact with each other.

Chapter summary: In this chapter, we demonstrate epidemic spreading control on irregular cell space with geographic divisions. Open data repositories offer multiple parameters that can be used to approximate local behaviors inside automata specifications. This is interesting in the case of dynamic systems such as epidemic monitoring such as Covid-19 in France. As many as 15 parameters are referenced by related databases, it is necessary to be correlated inside territories such as regions or clusters of districts. Permutation feature importance is used to select parameters for defining local transition rules. The practical interest is to understand the epidemic spreading in time and space. It is shown that control measures can be defined, and their effects are predicted by simulation. Principles can be reproduced in a number of situations provided that accurate geographic segmentation and related data are available. The last section provides the state of the art related to modeling and predicting the evolution of epidemics. The next chapter will provide another approach to combining regular and irregular cell space in distributed processing systems.

Chapter 5

Monitoring Shores: Hybrid System of Regular and Irregular Cell Spaces

Chapter introduction: Marine pollution comes from different sources including agriculture, industry, and domestic wastewater discharge from human activities in coastal areas. Environmental simulation can represent ground and sea characteristics, modeling spreading occurring in both spaces. These characteristics are variable, due to soil capability and reaction, and sea behavior, in particular currents and tides. This work presents a heterogeneous tiling approach modeling sea behaviors in coastal areas based on tidal currents and ground effects. The ground is segmented into irregular cells following geographic divisions for collecting observations while the sea area is segmented into regular geographical tiles. The impact of the interactions is represented by messages carrying qualities and quantities of physical pollution. Channels link cells following cellular automata or distributed system paradigms. This system architecture allows to produce a synchronous message passing program suitable for massive parallel execution. The status of cells and messages are produced step by step and can be interpreted graphically. Green tides caused by eutrophication appear when nutrients circulate in high concentrations in coastal waters. These nutrients come from land use, accumulate, and propagate to the shores mainly through rivers end up joining the sea or the oceans. Our simulations show when and where tides are able to increase concentration levels, producing space and time characteristics.

5.1 Green Algae Issues

Green tides have occurred in an increasing scale and frequency over the last few decades [51]. The phenomenon refers to the sudden and excessive growth of macro-algae, such as

Ulva, in coastal areas. A large marine area is affected by green tides including America, Europe, and Asia-Pacific [177]. From May to July 2008, prior to the 29^{th} Olympic games in Beijing, China, the Yellow Sea coastline experienced a massive green tide with approximately 1 million tonnes of drifting biomass covering an area of 13,000 - 30,000 km^2 [147]. The problem of green tides has occurred yearly in late spring and summer on the coast of Brittany, France since the 1970s with intensive agriculture providing nutrient enrichment [38]. The most important attribute of these algae is the competitive response to the high nutrient content of the surrounding waters for rapid uptake and storage of nutrients. Optimal temperature, light, and nutrients contribute to the formation and high relative growth rates of green macro-algae. As an example, Ulva prolifera, a type of macro-algae, is growing at an approximately two-fold growth rate per day in optimum conditions [67]. Macro-algae need to be in contact with the water surface to photosynthesize and grow, thus calm water condition is another reason allowing them to remain at the surface for extended periods of time. On the other hand, wind and currents contribute to the transport of the blooms to larger coastal areas. These blooms have significant ecological, economic, and social impacts, including the loss of biodiversity, degradation of habitats, and reduced access to fisheries and tourism activities. An additional severe effect is the release of toxic products during the decomposition process. The most widely treatment of these blooms is physical removal and transportation to landfill sites at a considerable cost. The total volume of green algae collecting in Brittany, France in 2004 is 69,225 m^3 with an expense of around 610,000€[24].

Eutrophication is one of the major reasons causing this phenomenon. It is a process of natural or man-made enrichment with inorganic nutrient elements - mainly nitrogen and phosphorus [57]. When these nutrients are presented in high concentrations in coastal waters, they can stimulate the growth of macro-algae, leading to green tides. Sources of nutrient pollution can include agricultural runoff, untreated sewage, and industrial discharges. In agriculture, nitrates are essential plant nutrients, but in excess amounts, they have a negative impact on the water quality. Together with phosphorus, nitrates in excess amounts can accelerate eutrophication causing a dramatic rise in aquatic plant growth as well as changes in the types of plants and animals that inhabit the stream. This, in turn, affects dissolved oxygen, temperature, and other indicators. In [126], the authors showed that the nitrate concentration in rivers must be limited between 5 and 15 mg/l, depending on the bay to reduce macro-algae blooms. Nitrates from land sources end up in rivers and streams more quickly than other nutrients like phosphorus. This is due

to the fact that they more easily dissolve in water than phosphates. As a result, nitrates serve as an indicator of the possibility of a source of sewage or manure pollution. Nitrate has no color, smell, or taste, making it difficult to detect in water thus it is important to get it tested in a laboratory. Although nitrate is generally recognized to be the main nutrient source of macro-algae, some genera such as Ulva can also rapidly take up and utilize ammonia, another form of nitrogen.

5.2 Modeling Pollution on Shores

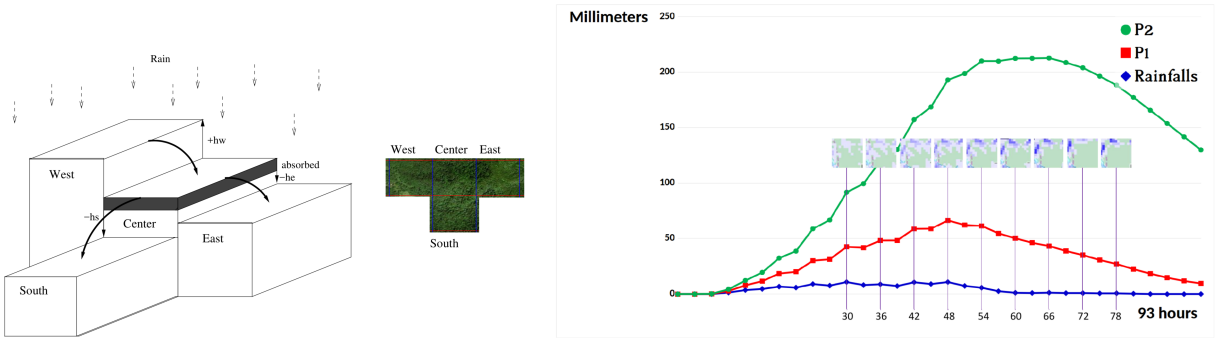
Coastal interactions appear to be mainly from the ground to the sea, especially in the case of intensive agriculture, urban development, and industry. The nature of these exchanges is chemical, biological, or sediments, with sometimes huge quantities of pollutants sent to the sea. These exchanges are in fact bi-directional, ocean activity producing, in turn, sediments and biological effluents that are spread back to the coast. A first classification inside this system can distinguish a set of behaviors: ground circulation depending on the nature of ground pollutants, ocean circulation, and the nature of spreading.

5.2.1 Modeling Ground Activities

In the case of pollution on shores impacted by ground activities, simulation tracks the pollution sources and their propagation to the shores. Tracking pollution spreading can be approximated by simulation or observed by sensors.

During rainfall events, water flow can pick up pollutants from the land and carry them into water bodies. This runoff can contain contaminants, which can impact the water quality and marine ecosystems downstream [167]. The pollutants accumulate and propagate to the shores mainly through rivers end up joining the sea or the oceans. The rivers are the channel connections for sending ground impacts to the oceans. The propagation of pollution with rainwater can be simulated on a 2D ground segmentation according to a geographic grid, augmented by elevation. As an example, the real problem is water distribution in regular tiles with rain dropping on the ground [157]. Water disappears locally because of absorption or evaporation and water moves from cell to cell according to elevation differences as shown in Figure 5.1a. This is an incomplete neighborhood showing a center with three neighbors. Water is measured as a volume, and influence is its circulation, according to ground topology and preferred top-down circulation. Each cell

will receive rainwater, dispatch part of this water to neighbor cells with lower elevation, and absorb another part. While real dependencies are water leaking from cell to cell, the abstract behavior is represented by messages sent and received to and from neighbors. Messages are volumes of water exchanged between neighbors. Figure 5.1b shows the water levels of two different locations during a four-day rainfall. Position P_2 (green line) has a lower elevation than position P_1 (red line). Due to the significant difference in the slope of the ground, the position with a lower elevation accumulated a lot of water causing flash floods.



(a) (b)

Figure 5.1 – Water distribution with rain dropping on the ground [157]. a) Physical exchange during a rain episode. Water flow downward in the effects of elevation differences. b) The water levels of two different positions during 4 days of rainfalls.

Another way to monitor ground activities is to observe pollution sources by dedicated sensors. The observation results are produced and can be queried within administrative boundaries. The core segmentation to model the problem on the ground is geographic divisions to collect or observe human activities generating pollution sources. The data collected belonging to these territories for spatial and temporal simulation as the government policies to monitor the environment. Nitrate has been monitored in Brittany, France since 1995 as a parameter of water condition [1]. In 2020, 718 stations were taken to evaluate nitrate concentrations based on five classes: bad (> 50 mg/l), poor (> 25 and ≤ 50 mg/l), medium (> 10 and ≤ 25 mg/l), good (> 2 and ≤ 10 mg/l), and very good (≤ 2 mg/l). The limitation value is 50 mg/l used as an indication of bad water condition. The Brittany region is classified as a "vulnerable zone" with high nitrate concentration

1. https://Brittany-environnement.fr/nitrates-cours-eau-bretons-datavisualisation accessed on 12 April 2023

reported as shown in Figure 5.2. Action programs have been initiated in these territories concerning balanced fertilization and good agricultural practices must be respected.

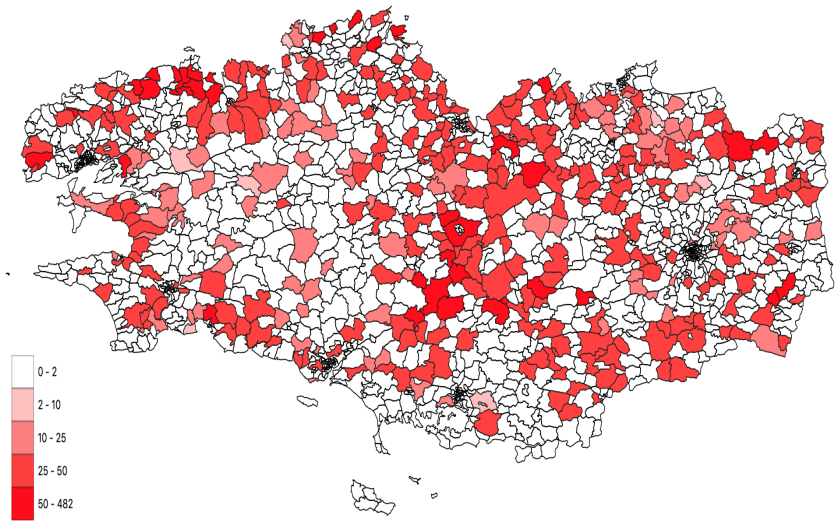


Figure 5.2 – Nitrate concentration in Brittany, France in 2020. The value is evaluated and classified in 5 classes as water quality evaluating system. Data are collected and managed by the Environment Observatory in Brittany.

5.2.2 Ocean Activities Measurements and Modeling

The oceanic behavior is based on currents carrying objects on movement such as algae blooms. Aerosolized toxins from harmful algae blooms are transported on land with wind effects and subsequent exposure and inhalation of the generated aerosols can induce adverse human health effects [102]. The nature of spreading in sea areas is different with different kinds of pollutants. As an example, green algae tend to develop and spread out on top of the water whereas plastic trash can be transported around the world by ocean currents. Waves produce sprays carrying back marine pollution to the ground.

Physical causes of marine currents

The term “current” when used in association with water describes the horizontal motion of water. Marine currents are complex motions of water produced locally or globally by physical forces: wind, water density differences, and tides resulting from the sun and moon attraction and the Coriolis effect. Winds drive currents that are at or near the ocean’s surface. Winds have a tendency to drive localized currents along coastlines, which

can lead to phenomena like coastal upwelling. On a larger scale, winds create currents in the open ocean that circulate water thousands of miles throughout the ocean basins. Another factor that drives ocean currents is thermohaline circulation. The process is driven by density differences due to variations in temperature (thermal) and salinity (haline) in various regions. Thermohaline circulation-driven currents are significantly slower moving than tidal or surface currents and can be found in both deep and shallow ocean depths. Currents driven by wind and thermohaline circulation are non-periodic currents to be predictable since they are associated with changing weather. The gravitational pull of the moon and the sun causes tides. Water moves up and down over a long period of time during tides. The currents produced by tides are varying near the shore, in bays, and in estuaries along the coast. They are referred to as "tidal currents". Strong tidal currents can travel at speeds of eight knots or more. The currents are generally measured in knots (1 knot = 1.85 kilometers per hour) and the direction of a current is the direction it is headed for or where the current is flowing towards counted from 0^o to 360^o (0^o being the geographic north) clockwise.

Currents estimation and measurement

The current measurement is of fundamental importance to the understanding of circulation patterns in the oceans.

Sensing observation: Current measurement is typically done with current sensors, which are divided into two groups: single-point current meters and current profilers. A current meter is an oceanographic device for flow measurement by mechanical, tilt, acoustical, or electrical means. It is an instrument for measuring the velocity of the flow of a fluid in a stream. The current meters are usually incorporated in a magnetic compass to determine the orientation of the instrument with respect to the magnetic north. Single-point current meters will only measure the current in the exact depth where they are installed [28]. Four classes of current meters are distinguished, based on the method used for measuring current magnitude. Tilt current meters are based on the principle that a tethered object will experience a tilt induced by the force balance of buoyancy, drag, and mooring tension. Mechanical current meters use a propeller-type device to measure the current speed, and a vane to determine current direction. Electromagnetic current meters combine a coil to produce a magnetic field and two sets of electrodes to determine the speed and direction of the ocean current. Typical acoustic current meters will have two orthogonal sound paths and the difference in arrival time for the sound traveling in opposite directions gives

the water velocity along the path. The class that is more commonly used for irrigation and watershed measurements is the mechanical type. However, the use of electromagnetic current meters is very popular among water districts. As sea currents often tend to vary significantly with depth, it is often preferred to measure the current profile for the entire water column by using a current profiler. The Acoustic Doppler Current Profiler is extensively used to measure ocean currents [20]. It operates on the same principle as acoustic current meters using reflections of the sound wave from drifting particles for the measurement. The profiler sends an acoustic signal into the water column and that sound bounces off particles in the water. The instrument calculates the speed and direction of the current by knowing the frequency of the return signal, the distance it traveled, and the time it took for the signal to travel. The current profiler can be attached to the bottom of a boat to estimate currents in different locations. Another method for estimating large-scale current circulation is the tracking of drifters built to follow the ocean surface. A drifter [2] consists of a surface float, tether, and drogue (i.e., sea anchor). The surface float is buoyant and remains at the ocean surface, while the sub-surface drogue extends to roughly 20m depth designed to move with the near-surface ocean currents. Without the presence of a drogue, the surface float will also be transported by wind and waves. A drifter can transmit data and information via satellite communications or the global positioning system. Drifters can also submerge for long periods of time to measure ocean currents at a particular depth [3]. The drifter would then periodically resurface to send a signal with its data and position to observers on the ground. This method of estimating the surface current field is time-consuming and, like the current meter approaches, quite expensive because the drifters are typically watched by aircraft, radar, ships, or satellites.

Radar observation: Radio antennas and high-frequency Radio Detecting and Ranging systems (radar) are used to map surface ocean current patterns over a large area in coastal areas [66]. Similar to the Acoustic Doppler Current Profiler, these shore-based instruments use the Doppler effect to determine when currents are moving toward or away to/from the shore to measure the velocity of a current. The utilization of high-frequency radar systems in coastal areas has rapidly increased alongside the use of moored current meters [74]. It has been demonstrated that coastal high-frequency radar can resolve quick changes. However, their coverage remains limited although the number of high-frequency radars has been augmented.

2. https://www.aoml.noaa.gov/phod/gdp/faq.php#velocity accessed on 21 March 2023
3. https://oceanservice.noaa.gov/facts/currentmon.html accessed on 21 March 2023

Satellite observation: A combination of surface current measurements by satellite and high-frequency coastal radar is a promising approach to cope with both the resolution and fast dynamics characteristic of coastal areas and the medium size and slower evolution of surface currents in the open ocean regions. However, direct measurements of surface currents by satellites remain quite limited [74]. Emery et al. [45] showed how to trace the displacements of sea surface temperature characteristics between successive infrared satellite photos in order to determine the surface currents. Choi et al. [26] and Yang et al. [175] determined the surface currents off the west coast of Korea based on the total suspended particulate matter concentration pictures taken from the Geostationary Ocean Color Imager. The satellite-derived surface currents were compared with instantaneous moored current observations to demonstrate the retrieval methodology. In [69], the semidiurnal tidal current ellipses derived from the Geostationary Ocean Color Imager are validated with a comprehensive set of the historical surface drifter and moored current meter observations.

It is obvious that many devices would be needed in order to map large areas of the ocean. The Global Ocean Observing System [4] is operated by more than 80 countries. The observing network includes ships, both academic and merchant, surface and subsurface mobile instruments, and fixed platforms, e.g., 4,000 profiling floats, 1,500 drifting buoys, 400 moored buoys, 150 HF radar operational stations, and 200 autonomous underwater vehicles.

Observation data are collected from different types of sensors (currents, salinity, temperature, oxygen, and pressure) at selected locations to track the movement of various water masses. These data are used to generate current predictions in the locations. To create accurate prediction models, it is necessary to periodically resurvey various coastal and estuarine locations [5]. Extensive numerical models have been used to study the characteristics of tidal currents based on physical and geographical parameters temperature, salinity, water depths, vertical turbulent viscosity, and diffusion [16]. These numerical models need to be assessed and calibrated using observational data, such as observation with current sensors, to produce reliable current models.

4. https://public.wmo.int/en/resources/bulletin/global-ocean-observing-system-oceans-of-data-earth-system-predictions accessed on 21 March 2023

5. https://oceanservice.noaa.gov/facts/current-survey.html

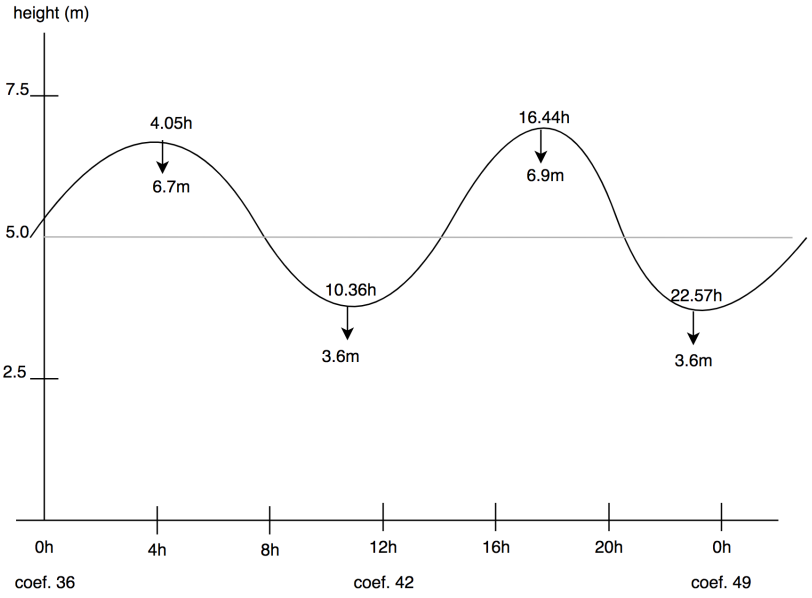
Tidal Currents

Tidal currents are important to analyze since it can affect coastal activities such as fishing, shipping, and tourism. Tidal currents can break up some pollutants or carry others back to shore and contribute to shoreline changes such as erosion. However, the regular reversal of the tidal currents can give a reliable supply of energy, a renewable energy powered by nature. Predictions of the tidal current are reliable at the precise depth and location where the current data were collected. If the bathymetry does not change significantly, such estimates might be applicable at nearby locations ideally at the same depth. Dramatic changes in bathymetry, such as due to dredging or rapid shoaling or due to the dramatic movement of sediments after a large storm, can make the tidal current predictions invalid [75]. In addition, the shoreline and bathymetry can gradually change over time, which implies that the tidal currents can similarly gradually change.

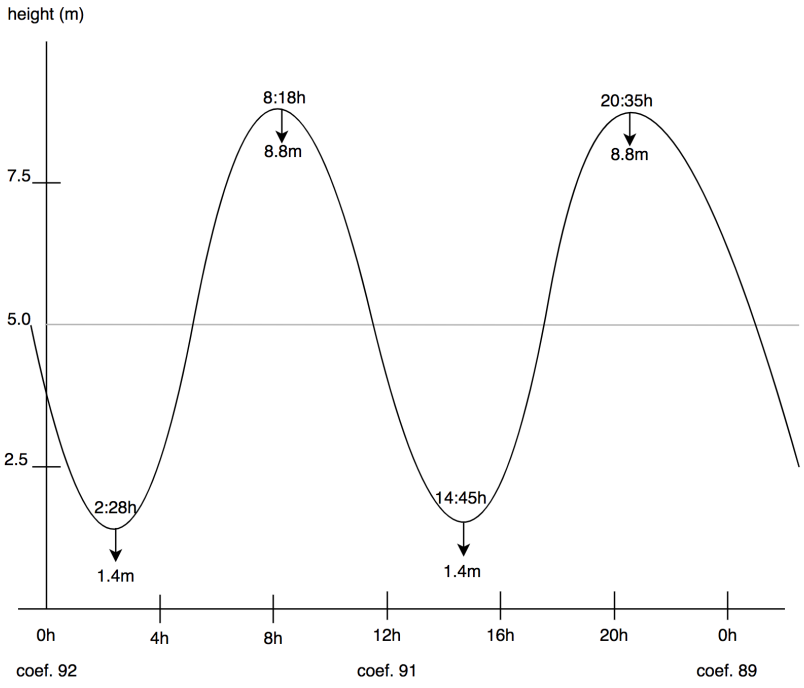
Tidal currents or periodic currents vary in a relatively predictable rhythm, making predictions for future dates possible. The tidal coefficient is the size of the tide in relation to its mean. Typically, it fluctuates between 20 and 120. The tidal range, or the variation in water height between high and low tide, increases with increasing tidal coefficient. As an example shown in Figure 5.3, the tidal range is 3.1m corresponding to the lower coefficients of around 40 and the tidal range is 7.4 corresponding to the higher coefficients of around 90. This indicates that the water level rises and retreats considerably. The mean water height is different between places. It is 5m in Roscoff, France whereas the Bay of Mont Saint-Michel has a mean water height of 7.5m.

Figure 5.4 shows the tidal estimation in the same place in April 2023. The strong tides, called spring tides, occur near the times of a new moon or full moon when the Earth, moon, and sun are all lined up so that the tidal bulges caused by the moon and by the sun add to each other. The weak tides, called neap tides, occur near times of first quarter and third quarter the tides when the moon and the sun work against each other. The mean coefficient of spring tides is 95 and neap tides is 45.

Two basic measurements are needed for the estimation of the tidal current velocity [43]. The first data is the spring water level that will be needed for the evaluation of the tidal coefficient. It represents the difference between high and low tides. Thanks to astronomic considerations, this coefficient can be easily calculated in advance for years by oceanographic services. The second needed data is the spring and neap water current velocities data, v_{sw} and v_{nw}, respectively. These current velocities are given by the Naval Hydrographic and Oceanographic Service (SHOM) [141].



(a)



(b)

Figure 5.3 – Water height in relation with tidal coefficients in Roscoff, France (a) 01 April 2023 (b) 08 April 2023. Larger coefficients: water level will increase and decrease in a large range indicating a strong flow of water. Smaller coefficients: water level will increase and decrease in a small range indicating a weak flow of water.

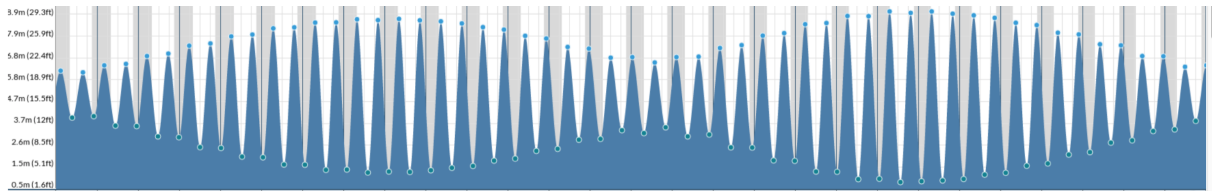


Figure 5.4 – Tidal estimation in Roscoff, France. The periodic rise and fall of sea water estimated in one-month period April 2023.
(https://www.tide-forecast.com/locations/Roscoff-France/tides/latest accessed on 01 April 2023)

SHOM operates for reference maritime and coastal geographic information. One of the products provided by SHOM is the tidal current prediction in CSV format including the components of surface tidal currents hour by hour and for two characteristic tidal coefficients of 45 and 95. The data come from calculations of finite element models as also developed using TELEMAC-2D, a computational system that calculates free surface flows in two dimensions of horizontal space. This is an open-source system, which is helpful in developing a model suitable for different situations. The system is of general application to hydraulic problems as it is able to take into account various phenomena. Awad and Darwich [9] used the TELEMAC-2D model to simulate the quality of seawater affected by wind and local currents. The aim was to determine how much the pollutants outlet into the sea with the sea current intensity and direction. Three scenarios were taken into consideration with the average daily spill of sewage produced by the inhabitants living in the river basin. Li et al. [99] studied the impact of building representations on urban flooding using the TELEMAC-2D model. By physical experiments and numerical modeling, the results showed that it is largely applicable to the simulation of urban flooding although there are some differences exist in the simulation results. The model was used in [52] to simulate the dynamics of the rivers during a flood period in Brazil. The results were validated using the city flood map provided by the government.

An example of the current data provided by SHOM is shown below. Each data file includes a header and three lines for each data point, in which the header is the port name. The first line represents the positions in geographical coordinates with latitude and longitude in degrees and minutes reported to the WGS84 geodesic system. The data points are presented as colored dots in Figure 5.5. The second and third lines represent the spring tide currents with coefficient 95 and the neap tide currents with coefficient 45, respectively, in 0.1 knots, hour by hour from -6 hours to +6 hours in relation to the high tides or low tides. The horizontal u and vertical v components are separated by an

asterisk. These components are positive towards the east and the north. An example of the data is shown in Table 5.1.

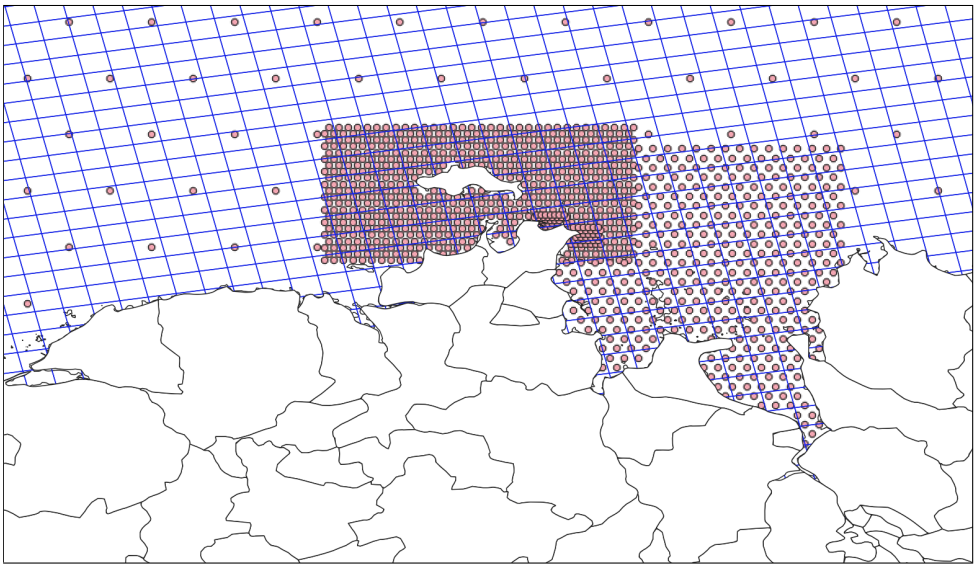


Figure 5.5 – Ocean tile versus ground layout: IRIS on land and grid of 1 km^2 in the ocean. Tidal currents are presented as data points depending on SHOM choice.

Roscoff
4843.628 -358.975
0 0 4 12 11 8 3 -6-11-11 -7 -1 0 * 0 0 0 -1 -1 -1 -1 0 0 1 -1 0 0
2 5 7 8 7 5 1 -5 -8 -7 -6 -4 -2 * 0 0 -1 -1 0 0 0 0 0 0 0 0 0 0

Table 5.1 – Tidal current in one data point from SHOM.

The current velocity is then calculated by equation 5.1 [141]. In which, a_0 is the tidal coefficient for neap tides, which is 45, and b_0 is the tidal coefficient for the spring tides, which is 95. The current direction is calculated by equation 5.2 also depending on the mean tidal coefficient of spring and neap tides.

$$v = v_{nw} + \frac{C - a_0}{b_0 - a_0}(v_{sw} - v_{nw}) \tag{5.1}$$

$$d = d_{nw} + \frac{C - a_0}{b_0 - a_0}(d_{sw} - d_{nw}) \tag{5.2}$$

where:

— v: velocity in a place
— $v_{nw} = \sqrt{u_{45}^2 + v_{45}^2}$: velocity in neap tides
— $v_{sw} = \sqrt{u_{95}^2 + v_{95}^2}$: velocity in spring tides
— C: coefficient in a time
— d: current direction in a place
— $d_{nw} = atan2(u_{45}, v_{45}) * \frac{180}{\pi} + 180$: direction in neap tides
— $d_{sw} = atan2(u_{45}, v_{45}) * \frac{180}{\pi} + 180$: direction in spring tides
— u_{45}, v_{45}: the horizontal and vertical components in coefficient 45 provided by SHOM
— u_{95}, v_{95}: the horizontal and vertical components in coefficient 95 provided by SHOM

5.3 Hybrid System: Regular in Combination with Irregular Cell Space

To model the impact of ground activities on marine pollution, we need to represent the two spaces in simulation systems. In this work, we segment the sea area into regular tiles and the ground area into irregular cell spaces. Doing this way, the impact of the ground to sea can be represented by interactions between neighboring cells. This work has been published in the International Journal of Environmental Science and Development (section 5.4 of the Appendix).

The simulation systems are generated and used in experiments as described below.

5.3.1 System Generation

On land, the IRIS is used to query local values monitored in the territories. As shown in Figure 5.6, the segmentation on the ocean is regular tiles with each cell having a size of 1 km^2. This is provided by the European Environment Agency [1]. For each country in the European Union, three types of vector polygon grid shape files, 1, 10, and 100 km2, are available covering at least the country borders plus a 15km buffer that is not reflecting the extent of the territorial waters. As can be seen in Figure 5.5, the 1 km^2 grid is different from the SHOM grid of currents. Thus, we use the nearest point interpolation method to estimate currents in the 1 km^2 grid. The ground and ocean segments are connected with channels carrying messages at an abstract level. Messages are the only way to exchange inside the neighborhood.

Simulation systems are generated as the cooperation between irregular polygons in

land and ocean tiles. This cooperation provides a new approach to environmental simulation, especially pollution on shores related to chemicals released from land use: a process system based on CA with a set of processes including IRISs and a grid of 1 km^2. Each process interacts with other processes within the adjacent neighborhood. The simulation in coastal areas strongly depends on tidal currents. The initial conditions are tidal currents in tiles and local data collected on land. QuickMap/PickShape allows to zoom and pan to select specific locations and PickShape generates cell systems with adjacent neighborhoods in the location depending on input shapefiles.

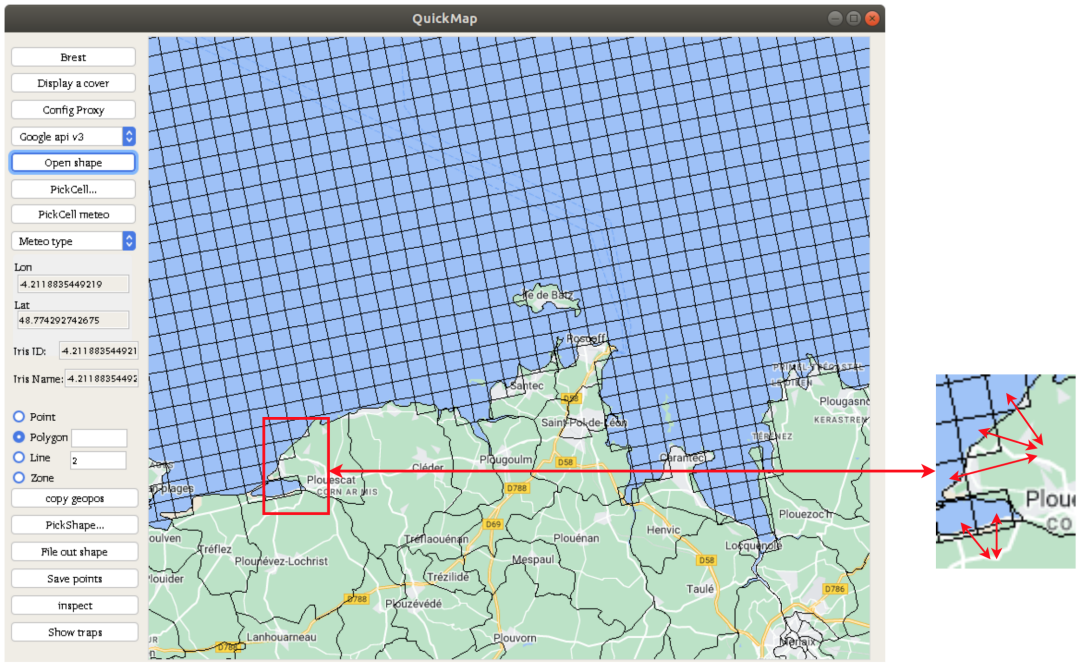


Figure 5.6 – Segmentation on ground and sea area. Irregular segments on ground as the geographic divisions by government and regular tiles on the ocean. QuickMap/PickShape allow to zoom and pan for zone selection and generate process systems in the selected zone.

5.3.2 Water Circulation with Currents

Experiments are presented with a system of 891 cells including 854 grid cells in 1 km^2. In Figure 5.7, current directions are represented by arrows and current strengths are the length of the arrows. The current direction and strength will be changed hourly with different tidal coefficients. In this case, we use tidal coefficients provided by SHOM

in Roscoff harbor [6] in a period of 2 months, June and July 2022.

Water circulation is described by currents in the sea area. This is useful in monitoring marine pollution since water flow transports and spreads pollution in aquatic environments. In areas with limited water circulation, pollutants tend to accumulate and persist, leading to serious environmental and ecological consequences. Lack of water circulation promotes the accumulation of pollution, such as the growth of harmful algal blooms as a case study will present in section 5.3.4.

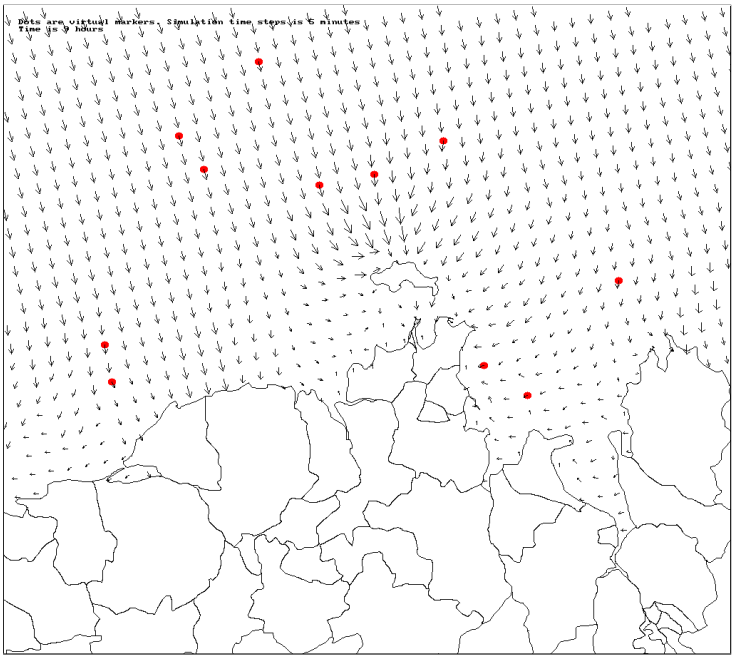
As shown in Figure 5.7a, the global direction of the currents is southeast with flood currents. After an hour, in Figure 5.7b, the global direction is changed along with the current strength. The figures reveal three retention zones with a lack of water circulation that can concentrate and retain pollution without being advected offshore by currents. These places show the need for regular monitoring to identify the potential sources of contamination. This is crucial to prevent the development of high pollutant concentrations in marine coastal waters.

5.3.3 Tracking Virtual Markers

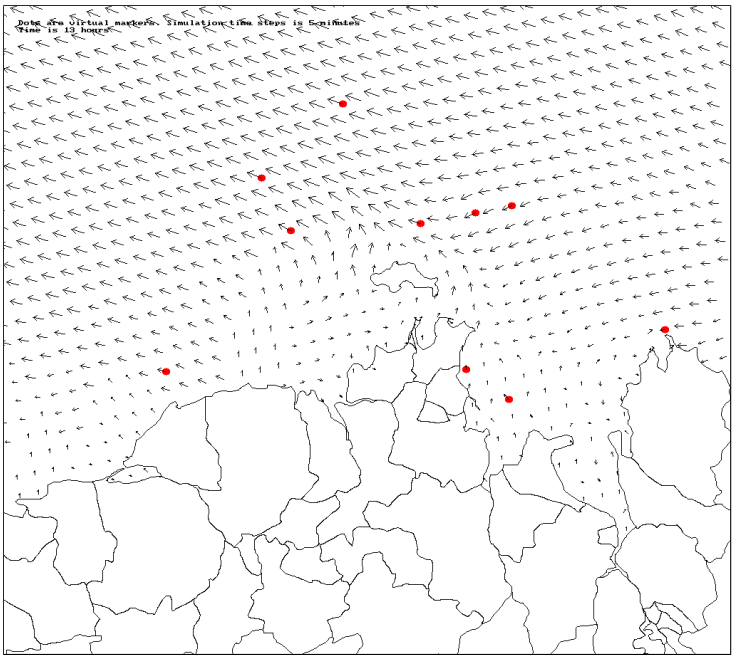
We monitor virtual markers moving in sea areas with currents. Following the CA approach, the simulation is based on synchronous processes communicating over channels. These processes execute a single program operating on local data and messages representing neighbor interactions. Data are tidal currents considering geographical locations. A simulator fires all the cells in parallel, taking care of transports, cell to cell, whatever the nature of the transport. Simulation steps follow the physical application rate, and the necessity to cover large time periods.

The maximum current strength is 7 knots (12.95 km/h) in Brittany beaches, France. Thus, to catch the movement between cells in 1 km^2 (move 1 cell at most by each iteration), the simulation time step is 5 minutes. We drop virtual markers in the ocean area and monitor the movement of markers as shown in Figure 5.7, where 13 markers are represented by red dots. The marker movement in grid cells is calculated by current speed and direction as shown in Algorithm 1. More figures showing the simulation results are available in the Appendix section 4.2.

6. https://maree.SHOM.fr/harbor/ROSCOFF/coeff?date=2022-07-01 accessed 18 April 2023



(a)



(b)

Figure 5.7 – Marker movement by currents. Current direction is represented by arrows and current strengths are the length of arrows. a) 01 June 2022 9AM. b) 01 June 2022 1PM. Notice the decreased current strength and the direction, reflecting orientation of the sea entering the English channel. Also notice the influence of Île de Batz on current direction.

Algorithm 1: Marker movement in grid cells

```
Initialize n markers in n random grid cells.
if This cell has a marker then
    Calculate current speed ;
    Calculate the marker position in this cell;
    if marker position > 1 km then
        Move marker to my neighbor;
foreach n: neighbors do
    if n move a marker to its neighbor then
        Calculate n current direction;
        Calculate n current speed;
        // check moving by current direction
        if n move a marker to this cell then
            Update marker position in this cell; // by current speed
            Update marker identity for tracking;
end
```

It is possible to keep a history of visited places of each marker to track biological or physical alterations. Figure 5.8 shows the positions of markers after 4 days. It can be seen that there are loops in the direction of the markers. Some markers close to the coastal line cannot escape to the sea affected by flood and ebb currents. A clearer view is provided in Figures 5.9 and 5.10. These figures show the marker movement in different places in the first 6 hours and the second 6 hours.

A subsystem in the simulation is that cells will have pipelines to slow down transfer objects internally. Transfers are achieved upon the agreement of the neighbors. Management of virtual markers in studied zones is similar to pollution accumulation on shore: with counting bags storing objects in excess. The local count represents the density of objects such as pollution. This gives a potential way of monitoring the spatial and temporal distribution of marine buoyant plastics by providing the dynamics and pathways of how plastic waste moves in the sea area and enters the shores.

5.3.4 Monitoring Pollution Coverage Area

In the same space, green algae development can be monitored with the effect of currents. Ground information is coming from administrative data collection, with known agricultural activities producing pollution. Colored IRIS represented the area with nitrate concentration on land as shown in Figure 5.11. These IRIS are sources of pollution

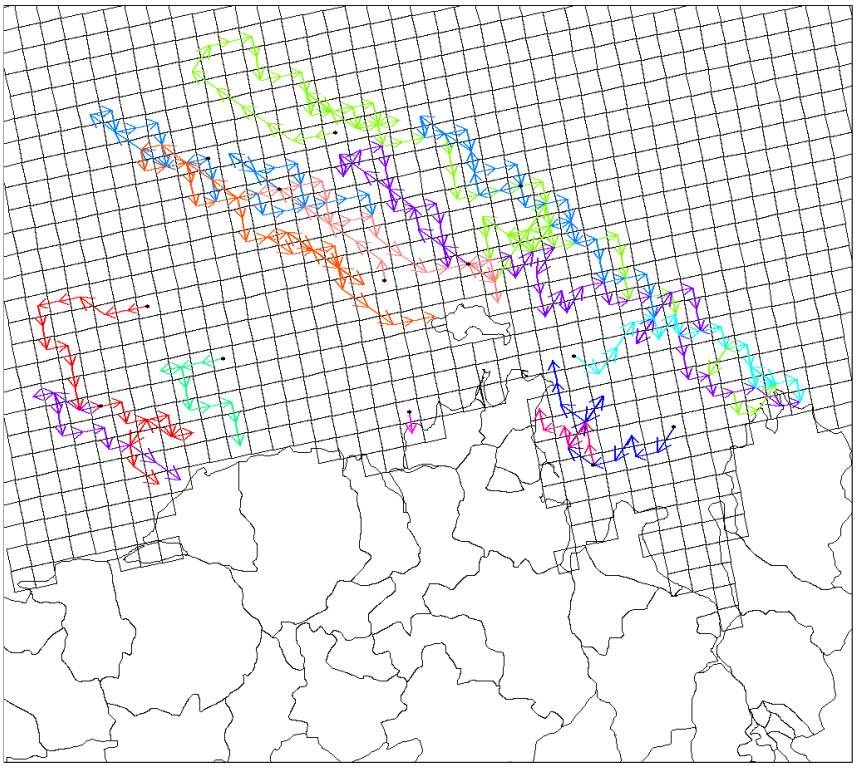


Figure 5.8 – Marker positions by 4 days showing loops in direction.

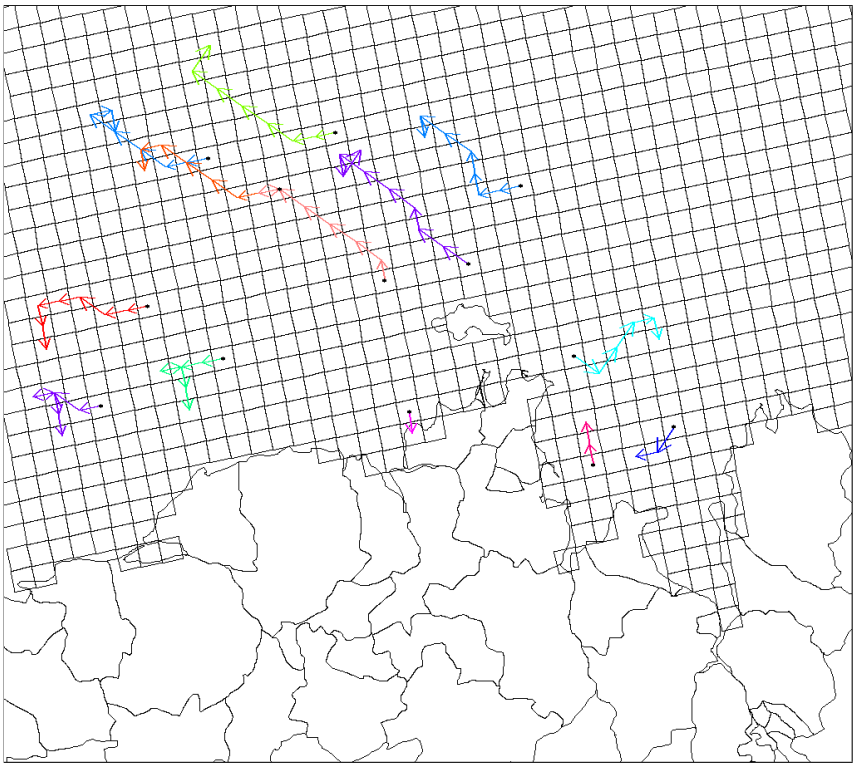


Figure 5.9 – Marker positions by first six hours.

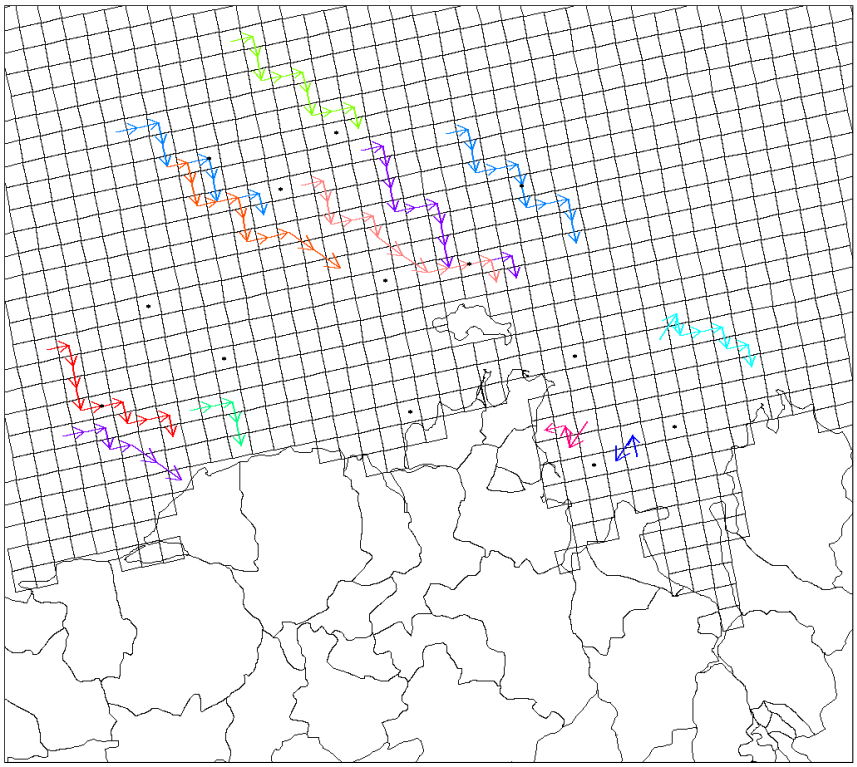


Figure 5.10 – Marker positions by last six hours.

that propagate nutrients to coastal areas.

Figure 5.12 shows the simulation of green algae development affected by currents. We initialize grid cells in the sea area with an algae density shown as green color in places close to colored IRIS (Figure 5.12a). The interaction between grid cells is followed by a Moore neighborhood with a simple transition rule defined as in equation 5.3. The algae density is the growth rate and the density sending/receiving to/from its neighbors. Density transmission between cells is affected by current strength and current direction.

$$D = D + D * r - Sum_{send} + Sum_{receive} \tag{5.3}$$

where:

— D: algae density in a cell
— r: growth rate per day, in this illustration, it is 75% meaning 0.0026% per iteration of 5 minutes
— Sum_{send}: algae density sending to its neighbors calculated by equation 5.4
— $Sum_{receive}$: algae density receiving from its neighbors calculated by equation 5.5

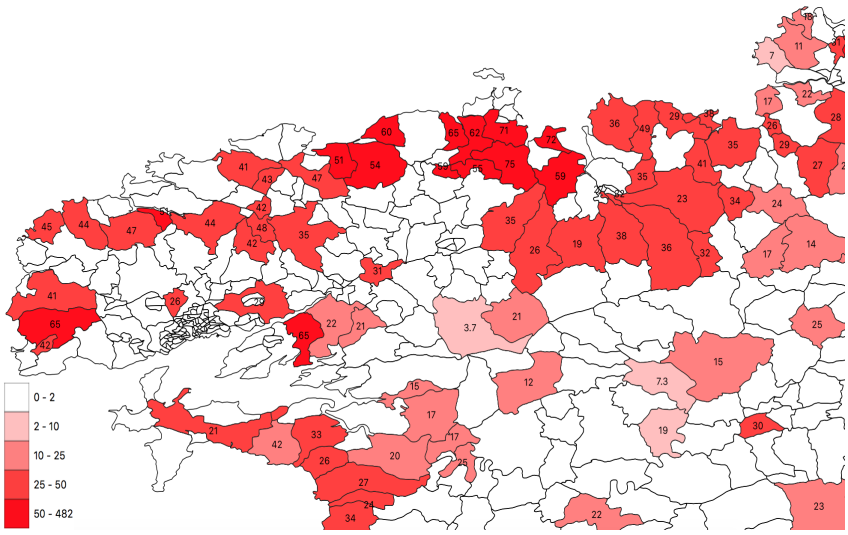


Figure 5.11 – Nitrate concentration values in North West Brittany, France in 2020. The values are presented inside IRIS in mg/l units.

$$Sum_{send} = \frac{D * \sqrt{u_C^2 + v_C^2}}{100} \tag{5.4}$$

where:

— u_C and v_C: tidal current of coefficient C represented by west-east and south-north components

— Sending a part of its density depending on its current strength

$$Sum_{receive} += D_i * \frac{v_i}{100} * \alpha \tag{5.5}$$

where:

— D_i: algae density in cell neighbor i

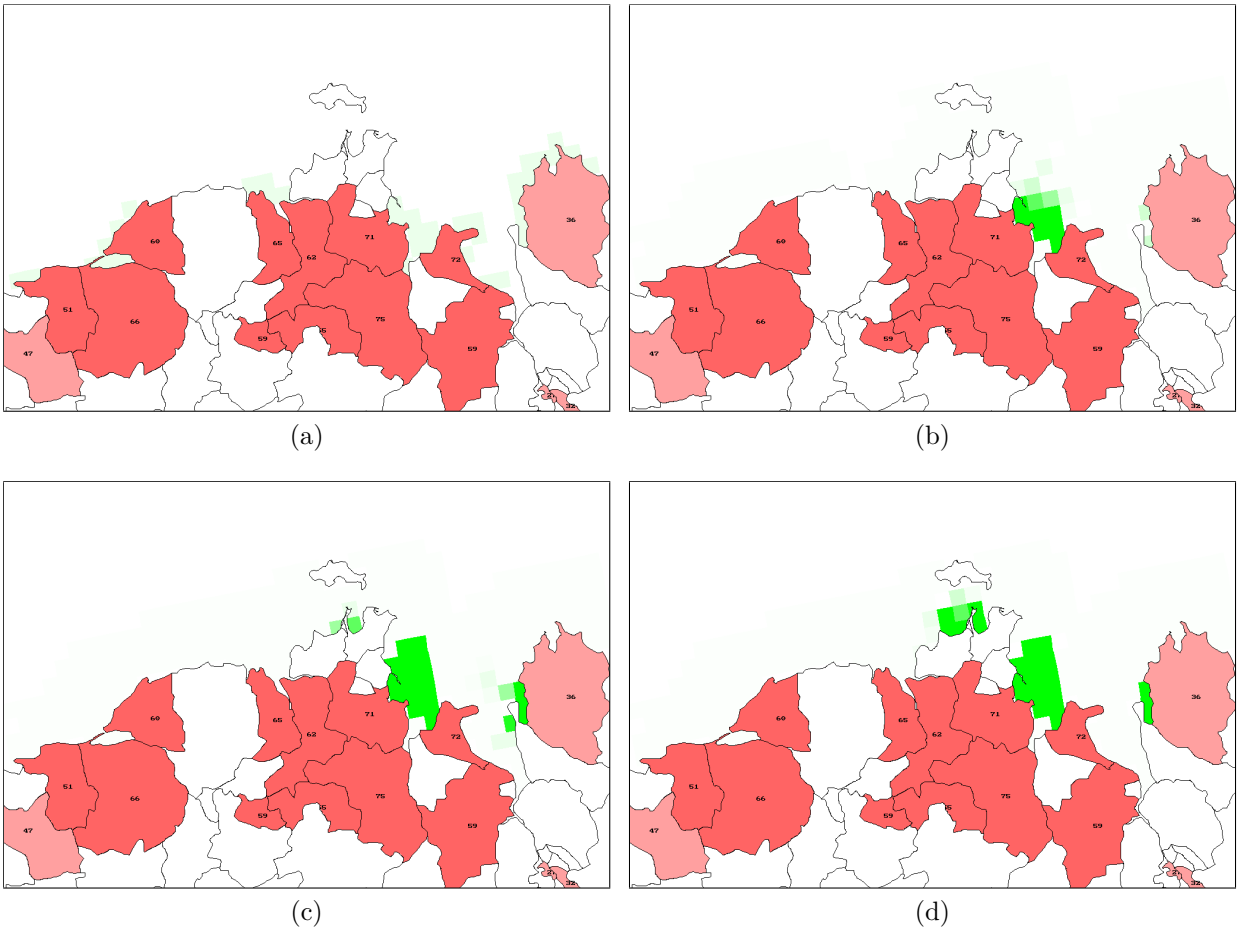
— v_i: current strength of cell neighbor i

— α: a fraction receiving from neighbor i depending on its current direction with west-east and south-north components

As shown in Figure 5.12b, after 15 days, the algae density develops in the area near Saint-Pol-de-Léon whereas other beaches show less chance of growing green tides due to water circulation. The sea currents spread progressively algae to neighbor cells, especially in retention zones. A cell will send its algae density to its neighbors with a percentage depending on its current strength. It can send to at most three neighbors depending on its current direction with west-east and south-north components. After 40 days (Figure

5.12c) and after 55 days (Figure 5.12d), the algae density develops in the area near Roscoff and Saint-Samson.

The eutrophication phenomena are generally confined to large enclosed water systems. The nutrients are entrapped by tidal currents causing eutrophication phenomena. In open areas with suitable dilution factors and widespread dispersal, they would not give rise to eutrophication phenomena. The simulation results show three places with the possibility of algae development problems. Especially, the polygons in Roscoff do not contain sources of pollution that propagate nutrients to the sea area. Thus, it is important to simulate the pollution propagation with water flow to examine the pollution sources to have the correct and appropriate solutions.



(a) (b) (c) (d)

Figure 5.12 – Green algae development affected by currents. Algae density is represented in green color. a) Initialize grid cells in sea area with algae density in places close to colored IRIS. b) Algae development affected by currents after 15 days. c) Algae development after 40 days. d) Algae development after 55 days.

The simulation mechanism reveals risk from ground activities, with spatial and time characteristics. It can be used for a wide variety of accidental or systematic activities. This provides a general approach to model pollution on shores with currents showing water circulation in places suspected of pollution, which can be employed as a quick assessment tool for studying marine pollution issues. It is possible to test different scenarios by modifying the initial conditions and the transition function. One could track the pollution propagation over time given a source of pollution. Additionally, making the simulation accessible to policymakers will help support them in making decisions related to marine pollution prevention. Specifically, the marine areas vulnerable to pollution will need to be monitored more closely. Different coastal localities may have different safety indicators for pollutants due to differences in water circulation. Overall, it is important to manage ground activities in a way that reduces the inputs of nutrients and other pollutants into coastal waters and promotes the health and resilience of coastal ecosystems. The limitation side is that tidal currents need to be re-computed in different places with respect to the geographical locations and bathymetry. In addition, tidal currents showing water circulation are important in monitoring the shores however it is also important to take into account the effects of other parameters such as wind, waves, and water surface temperature. The random behavior of spreading pollutants on the surface of the sea is highly affected by wind and waves.

Chapter summary:

This chapter shows the cooperation between regular and irregular cell spaces in modeling pollution on the shores. The green tides are discussed as the major reason for eutrophication accelerating by excess amounts of nitrates from agricultural runoff. Modeling ground activities can be done in two ways, simulation propagation of pollution sources on a 2D ground with elevation differences or querying observation data within administrative boundaries. Modeling ocean behaviors is done based on tidal currents in a regular grid of 1 km^2. An illustration of tidal currents is provided with marker movements in the coastal area near Roscoff, France. We also provide a simple model of green algae development in the effect of tidal currents with nitrate input sources in the same place. The next chapter will be the conclusion of the work and future direction related to this work.

Chapter 6

Conclusion and Perspectives

6.1 Conclusion

Open Data plays a vital role in the environmental and social simulation by providing the necessary input to model and simulate complex systems. The incorporation of real-world data into computer-based simulations helps to study and understand complex systems and their dynamics. Data can be obtained from various sources, including government agencies, research institutions, non-profit organizations, and citizen science initiatives. These data sources provide information about various factors, such as climate, land use, population demographics, health indicators, etc. By incorporating real-world data, we can create reliable models to study and forecast environmental phenomena. Simulations based on Open Data provide insights into the potential impacts of different policies or interventions on the environment or society. Policymakers can use these simulations to inform their decisions and assess the potential outcomes of different scenarios.

CA is a modeling approach used in environmental simulation to study the behaviors of complex systems, particularly those that involve spatial interactions. It provides a framework for representing the interactions of individual components, typically within a grid-like environment. Each cell has a state, which represents its current condition or characteristic. The state of a cell is updated simultaneously in discrete time steps based on the states of its neighboring cells. The neighborhood can be defined in different ways, such as the immediate neighboring cells or a larger surrounding area. The initial cell states are assigned by Open Data in specific locations. Once the simulation runs, the result analysis involves examining trends and spatial distributions of specific variables of interest within the simulated environment. Color-coded maps are used to represent the evolving states of the cells over time. This work is motivated by scientific models based on the CA approach with the availability of Open Data providing initial conditions to the cell processes. The objective of this work is to provide a general way of environmental and social simulation with geographical objects linking to government Open Data.

We introduce the use of geographic divisions as irregular cell space in modeling and simulation with the CA approach. A list of work has been done in the direction of CA approach using the UBO tool set (QuickMap/PickCell). QuickMap allows loading standard maps, zoom, and pan for selecting locations. PickCell helps to generate cell systems of regular tiles with Moore/von Neumann neighborhood augmented by elevation. In this work, a new module (PickShape) to the UBO tool set is provided as an alternative to regular tiles that helps to generate cell systems of irregular shapes with adjacent neighborhoods. The input data are collected and queried from the government Open Data portal in these geographic boundaries. Binding data into these cell processes is more precise than data estimation for regular grid cells as shown in section 3.2.2. In addition, a wide range of data can be passed into irregular cells since governments monitor and manage a country through different levels of geographic divisions. We explore parallel computation with process systems generated in Occam/C syntaxes. The C syntax is simple and suitable to be used with CUDA and MPI libraries. Moreover, this programming language gives more choices of the supported library to display the simulation results in comparison with Occam.

We observed the availability of geographical data, meteorological data, demographic data, and health statistics in the government's Open Data portal. These data are analyzed to provide a better understanding of the problems and effectively determine the cause of the problems. The Covid-19 epidemic outbreak is modeled on geographic divisions allowing spreading control before applying measures. The data types used in the modeling process are demographic data, mobility data, and weather data. The data analysis allows us to analyze the dependencies between these parameters with the Covid-19 incidence rate thus important parameters are figured out to model the epidemic propagation. Machine learning models are fit to synthesize the transition rules in a case study in Brittany, France. Data are collected in a one-year period from May 2020 to May 2021.

We introduce an approach of hybrid systems of regular tiles cooperating with irregular cells in modeling activities on the shores. The work mechanism can be used for a wide range of accidental or systematic activities. In modeling green tides, there are two important pieces of information in the hybrid systems. Nitrate concentration values are observed on-ground activities, especially on river water surfaces. The high nitrate concentration will accelerate eutrophication phenomena causing green tides. The second important piece of information is tidal currents, which affect water-related activities in coastal areas. Nutrients end up in rivers and later join in the sea or ocean and can be entrapped by

tidal currents causing eutrophication. A simple model of green algae development affected by currents is provided in the North West Brittany area near Roscoff, France.

6.2 Perspectives

The approach presented in this study uses Open Data from various sources in distributed processing systems for environmental and social simulation. This requires a wide range of Open Data to feed into the systems. However, Open Data is not always available, especially in developing countries. In Vietnam, the government Open Data portal [1] was started on 31 August 2020 to publish and provide Open Data by state agencies. The data provided is mostly decisions, regulations, and lists in pdf format. There is a lack of geographical data, no health statistics, no meteorological data, and no fertilizer data from agricultural use. Thus, it is difficult to collect data feeding the modeling systems.

In addition, geographic divisions are used as irregular cell spaces in modeling social and environmental phenomena. This helps in the case of correct binding data collected from administrative boundaries as compared to grid estimation. In this work, we use IRIS, the smallest level of administrative management, in collecting data and generating process systems. It is advantageous in simulating social activities associated with people. However, the average size of IRISs is 11.3 km^2, which is very large for environmental changes such as the spread of brown plant hoppers in the fields. In this case, land parcels may be a more reasonable choice for simulation. Land parcels are also a form of irregular cell spaces that can be generated by our systems.

In the case study of the Covid-19 pandemic, the natural way of spreading is to its adjacent neighbors. However, this is a simple representation assuming people can only move and contact with others living in the neighboring areas. In reality, people can move around by car, train, and flight to other regions or countries. Thus, the channel connections between cell processes need to consider this factor. In addition, the parameters used in modeling are only considered in relation to the incidence rate. There is also other important information such as the number of hospitalized, recovered, or died cases. To provide a better model, it is necessary to consider more factors in the modeling process.

The hybrid systems using regular and irregular cell processes provide an approach to model pollution on the shores. The tidal currents are used to represent water circulation in ocean areas. However, wind and waves are two other important factors that need to

1. https://data.gov.vn accessed on 28 April 2023

be considered in water movement. In the case of green tides, the growth rate of green algae blooms is affected by surface temperature, light, nutrients, and wind. These factors in optimal conditions will provide high growth rates causing green tides. Thus, the future direction is to take into account these factors for more accurate models.

The research opportunities opened by this work are in a wide range. Advanced deep-learning techniques have been used in epidemic simulation. A future direction is to use deep learning techniques in synthesizing the transition rules to consider more factors in the modeling process with high accuracy. The hybrid systems provide a starting point to monitor landslides and salinity threats. These two problems have occurred with a higher frequency in recent years in Vietnam. Since the beginning of 2023, Hau Giang Province, Vietnam has witnessed a total of 39 landslide locations, spanning a combined length of 902 meters and resulting in a loss of 5,031 square meters of land. This represents an increase of 11 landslide locations compared to the same period the year before. The tidal currents and soil characteristics are valuable information that can help to apply our approach to monitoring landslides and salinity threats.

Abbreviations

A-SIR (Asymptomatic-Susceptible-Infected-Recovered)

BUFR (Binary Universal Form for the Representation of meteorological data)

CA (Cellular Automata)

CPU (Central Processing Unit)

CUDA (Compute Unified Device Architecture)

INSEE (French National Institute of Statistics and Economic Studies)

GIS (Geographical Information Systems)

GPU (Graphics Processing Unit)

IGN (French National Geographic Institute)

LOD (Linked Open Data)

LAU (Local Administrative Unit)

MPI (Message Passing Interface)

NUTS (Nomenclature of Units for Territorial Statistics)

SHOM (Naval Hydrographic and Oceanographic Service)

SDAI (Standard Data Access Interface)

SEIR (Susceptible-Exposed-Infected-Recovered)

SHAP (Shapley Additive exPlanation)

SIR (Susceptible-Infected-Recovered)

WMO (World Meteorological Organization)

References

1. Agency, E. E., *EEA Reference Grid for France (1km), May 2013* accessed on 01 April 2023, `https://sdi.eea.europa.eu/catalogue/srv/api/records/ada072ce-a203-4e36-87f4-cbd021ab6435f`.
2. Al-Ahmadi, K., See, L., Heppenstall, A. & Hogg, J., Calibration of a Fuzzy Cellular Automata Model of Urban Dynamics in Saudi Arabia, *Ecological Complexity* **6,** 80–101 (2009).
3. Alamo, T., Reina, D. G., Mammarella, M. & Abella, A., Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic, *Electronics* **9,** 827 (2020).
4. Anderson, R., *in Population biology of infectious diseases* 149–176 (Springer, 1982).
5. Anderson, R. M. & May, R. M., Directly transmitted infections diseases: control by vaccination, *Science* **215,** 1053–1060 (1982).
6. Anis, A., The Effect of Temperature Upon Transmission of COVID-19: Australia And Egypt Case Study. *Available at SSRN 3567639* (2020).
7. Apley, D. W. & Zhu, J., Visualizing the effects of predictor variables in black box supervised learning models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82,** 1059–1086 (2020).
8. Aràndiga, F., Baeza, A., Cordero-Carrión, I., Donat, R., Martí, M. C., Mulet, P. & Yáñez, D. F., A Spatial-Temporal Model for the Evolution of the COVID-19 Pandemic in Spain Including Mobility, *Mathematics* **8,** 1677 (2020).
9. Awad, M. M. & Darwich, T., Evaluating sea water quality in the coastal zone of north Lebanon using Telemac-2D TM, *Lebanese science journal* **10,** 35 (2009).
10. Barnes, F. R., Blocking system calls in KRoC/Linux, *Communicating Process Architectures* **58,** 155–178, `https://kar.kent.ac.uk/21966/1/Blocking_System_Calls_in_KRoC-Linux.pdf` (2000).
11. Bastos, S. B. & Cajueiro, D. O., Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil, *Scientific Reports* **10,** 1–10 (2020).
12. Bays, C., *in Game of Life Cellular Automata* 1–7 (Springer, 2010).
13. Bennett, J., *OpenStreetMap* (Chapter 1. Making a free, editable map of the world) (Packt Publishing Ltd, 2010).

14. Bertuzzo, E., Mari, L., Pasetto, D., Miccoli, S., Casagrandi, R., Gatto, M. & Rinaldo, A., The geography of COVID-19 spread in Italy and implications for the relaxation of confinement measures, *Nature communications* **11,** 1–11 (2020).

15. Beyers, N., Gie, R., Zietsman, H., Kunneke, M., Hauman, J., Talley, M. & Donald, P., The Use of a Geographical Information System (GIS) to Evaluate The Distribution of Tuberculosis in a Highincidence Community, *South African Medical Journal* **86,** https://www.ajol.info/index.php/samj/article/view/155851/145478 (1996).

16. Blumberg, A. F. & Mellor, G. L., A description of a three-dimensional coastal ocean circulation model, *Three-dimensional coastal ocean models* **4,** 1–16 (1987).

17. Brandi, P., Ceppitelli, R. & Salvadori, A., Epidemic evolution models to the test of Covid-19, *Bollettino dell'Unione Matematica Italiana* **13,** 573–583 (2020).

18. Braunschweig, K., Eberius, J., Thiele, M. & Lehner, W., The State of Open Data, *Limits of Current Open Data Platforms,* https://wwwdb.inf.tu-dresden.de/opendatasurvey/www2012_short.pdf (2012).

19. Breiman, L., Random forests, *Machine learning* **45,** 5–32 (2001).

20. Brumley, B. H., Cabrera, R. G., Deines, K. L. & Terray, E. A., Performance of a broad-band acoustic Doppler current profiler, *IEEE Journal of Oceanic Engineering* **16,** 402–407 (1991).

21. Campbell, J. E. & Shin, M., *Essentials of Geographic Information Systems* (chapter 5. Geospatial Data Management), ISBN: 978-1-4533219-6-6, https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/index.html (Saylor Academy Open Textbooks, 2011).

22. Castilho, C., Gondim, J., Marchesin, M. & Sabeti, M., Assessing the efficiency of different control strategies for the COVID-19 epidemic, *Electron J Differ Equ* **2020,** 1–17 (2020).

23. Chandra, R., Dagum, L., Kohr, D., Menon, R., Maydan, D. & McDonald, J., *Parallel programming in OpenMP* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=45faf2f1b9119079beda1383ea78497f499a649a (Morgan kaufmann, 2001).

24. Charlier, R. H., Morand, P., Finkl, C. W. & Thys, A., *Green Tides on The Brittany Coasts in 2006 IEEE US/EU Baltic International Symposium* (2006), 1–13.

25. Chen, T. & Guestrin, C., *XGBoost: A Scalable Tree Boosting System in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, San Francisco, California, USA, 2016), 785–794, ISBN: 9781450342322.

26. Choi, J.-K., Yang, H., Han, H.-J., Ryu, J.-H. & Park, Y.-J., Quantitative estimation of suspended sediment movements in coastal region using GOCI, *Journal of Coastal Research* **65,** 1367–1372 (2013).

27. Clarke, K. C., Cellular automata and agent-based models, *Handbook of regional science,* 1217–1233 (2014).

28. Collar, P. & Griffiths, G., *in Encyclopedia of Ocean Sciences* (ed Steele, J. H.) 2796–2803 (Academic Press, Oxford, 2001), ISBN: 978-0-12-227430-5.

29. Costa, G. S., Cota, W. & Ferreira, S. C., Metapopulation Modeling of COVID-19 Advancing into The Countryside: An Analysis of Mitigation Strategies for Brazil, *medRxiv* (2020).

30. Dahal, K. R. & Chow, T. E., Characterization of neighborhood sensitivity of an irregular cellular automata model of urban growth, *International Journal of Geographical Information Science* **29,** 475–497 (2015).

31. Daniya, T., Geetha, M. & Kumar, K. S., Classification and regression trees with Gini index, *Advances in Mathematics: Scientific Journal* **9,** 8237–8247 (2020).

32. Davies, A. & Aldridge, J., A numerical model study of parameters influencing tidal currents in the Irish Sea, *Journal of Geophysical Research: Oceans* **98,** 7049–7067 (1993).

33. Defeo, O., McLachlan, A., Schoeman, D. S., Schlacher, T. A., Dugan, J., Jones, A., Lastra, M. & Scapini, F., Threats to sandy beach ecosystems: a review, *Estuarine, coastal and shelf science* **81,** 1–12 (2009).

34. Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H., Complexity of the basic reproduction number (R0), *Emerging infectious diseases* **25,** 1 (2019).

35. Desiderio, A., Salina, G. & Cimini, G., Multiplex mobility network and metapopulation epidemic simulations of Italy based on open data, *Journal of Physics: Complexity* **3,** 04LT01 (2022).

36. Desjardins, M. R., Hohl, A. & Delmelle, E. M., Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters, *Applied geography* **118,** 102202 (2020).

37. Deutsch, A. & Dormann, S., *in Cellular Automaton Modeling of Biological Pattern Formation: Characterization, Applications, and Analysis* 45–56 (Birkhauser Boston, Boston, MA, 2005), ISBN: 978-0-8176-4415-4.

38. Diaz, M., Darnhofer, I., Darrot, C. & Beuret, J.-E., Green Tides in Brittany: What Can We Learn About Niche–Regime Interaction?, *Environmental Innovation and Societal Transitions* **8,** 62–75 (2013).

39. Diekmann, O. & Heesterbeek, J. A. P., *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation* page 4 (John Wiley & Sons, 2000).

40. Doran, R. J. & Laffan, S. W., Simulating The Spatial Dynamics of Foot and Mouth Disease Outbreaks in Feral Pigs and Livestock in Queensland, Australia, Using a Susceptible-Infected-Recovered Cellular Automata Model, *Preventive veterinary medicine* **70,** 133–152 (2005).

41. Dormann, S. & Deutsch, A., Modeling of self-organized avascular tumor growth with a hybrid cellular automaton, *In silico biology* **2,** 393–406 (2002).

42. Du, M., Liu, N. & Hu, X., Techniques for Interpretable Machine Learning, *Communications of the ACM* **63,** 68–77 (2019).

43. El Tawil, T., Charpentier, J. F. & Benbouzid, M., Tidal energy site characterization for marine turbine optimal installation: Case of the Ouessant Island in France, *International journal of marine energy* **18,** 57–64 (2017).

44. Elshawi, R., Al-Mallah, M. H. & Sakr, S., On the interpretability of machine learning-based model for predicting hypertension, *BMC medical informatics and decision making* **19,** 146 (2019).

45. Emery, W. J., Thomas, A., Collins, M., Crawford, W. R. & Mackas, D., An objective method for computing advective surface velocities from sequential infrared satellite images, *Journal of Geophysical Research: Oceans* **91,** 12865–12878 (1986).

46. Ermentrout, G. B. & Edelstein-Keshet, L., Cellular automata approaches to biological modeling, *Journal of theoretical Biology* **160,** 97–133 (1993).

47. Fanelli, D. & Piazza, F., Analysis and forecast of COVID-19 spreading in China, Italy and France, *Chaos, Solitons & Fractals* **134,** 109761 (2020).

48. Ferretti, M., Musci, M. & Santangelo, L., *A hybrid OpenMP and OpenMPI approach to geometrical motif search in proteins in 2014 IEEE International Conference on Cluster Computing (CLUSTER)* (2014), 298–304.

49. Filipe, J. & Gibson, G., Studying and Approximating Spatio–temporal Models for Epidemic Spread and Control, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **353,** 2153–2162 (1998).

50. Fisher, A., Rudin, C. & Dominici, F., Model class reliance: Variable importance measures for any machine learning model class, from the Rashomon perspective, *arXiv preprint arXiv:1801.01489* **68,** `https://arxiv.org/pdf/1801.01489v2.pdf` (2018).

51. Fletcher, R., The occurrence of "green tides"—a review, *Marine benthic vegetation: recent changes and the effects of eutrophication,* 7–43 (1996).

52. Forster, A., Costi, J., Marques, W. C., Wormsbecher, A. G. & Bendo, A. R. R., *Application of the TELEMAC-2D Model in the fluvial hydrodynamics simulation and reproduction of flood patterns in Defect and Diffusion Forum* **396** (Trans Tech Publ, 2019), 187–196.

53. Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F. & Billa, L., Spatial analysis and GIS in the study of COVID-19. A review, *Science of the total environment* **739,** 140033 (2020).

54. Frisch, U., d'Humières, D., Hasslacher, B., Lallemand, P., Pomeau, Y. & Rivet, J.-P., *in Lattice Gas Methods for Partial Differential Equations* 77–136 (CRC Press, 2019), `http://photocityland.free.fr/archives/prepa/tipe/20072008/documents/lattice-gas-hydrodynamics.pdf`.

55. Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R. H., Daniel, D. J., Graham, R. L. & Woodall, T. S., *Open MPI: Goals, concept, and design of a next generation MPI implementation in Recent Advances in Parallel Virtual Machine and Message Passing Interface: 11th European PVM/MPI Users' Group Meeting Budapest, Hungary, September 19-22, 2004. Proceedings 11* (2004), 97–104.

56. Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R. & Rinaldo, A., Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures, *Proceedings of the National Academy of Sciences* **117,** 10484–10491 (2020).

57. Glibert, P. M., Seitzinger, S., Heil, C. A., Burkholder, J. M., Parrow, M. W., Codispoti, L. A. & Kelly, V., Eutrophication, *Oceanography* **18,** 198, `http://chnep.wateratlas.usf.edu/upload/documents/EutrophicationAndHABs.pdf` (2005).

58. Google, *New Climate Simulation Data Models now available in Google Cloud* accessed on 13 June 2023, Dec. 2019, `https://cloud.google.com/blog/products/data-analytics/new-climate-model-data-now-google-public-datasets`.

59. Gravier, D., Wulff, A. & Torstensson, A., Monitoring of Green Tides on The Brittany Coasts (France), *Primary producers of the sea Bio458* **2012,** 1–9, `https://dorian-gravier.com/files/pdf/gravier_-_2012_-_monitoring_of_green_tides_on_the_brittany_coasts_france.pdf` (2011).

60. Gropp, W. & Lusk, E., *User's Guide for mpich, a Portable Implementation of MPI* 1996, `https://www.researchgate.net/profile/Ewing-Lusk/publication/2807985_User's%5C_Guide%5C_for%5C_mpich%5C_a%5C_Portable%5C_Implementation%5C_of%5C_MPI/links/00463528aa8f280476000000/Users-Guide-for-mpich-a-Portable-Implementation-of-MPI.pdf`.

61. Guan, W.-J., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-X., Liu, L., Shan, H., Lei, C.-L., Hui, D. S., Du, B., Li, L.-J., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y. & Zhong, N.-s., Clinical Characteristics of Coronavirus Disease 2019 in China, *New England Journal of Medicine* **382,** 1708–1720 (2020).

62. Gudbjartsson, D. F., Helgason, A., Jonsson, H., Magnusson, O. T., Melsted, P., Norddahl, G. L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A. B., *et al.,* Spread of SARS-CoV-2 in the Icelandic population, *New England Journal of Medicine* **382,** 2302–2315 (2020).

63. Gupta, A., Watson, S. & Yin, H., Deep learning-based aerial image segmentation with open data for disaster impact assessment, *Neurocomputing* **439,** 22–33 (2021).

64. Hassen, H. B., Elaoud, A., Salah, N. B. & Masmoudi, A., A SIR-Poisson Model for COVID-19: Evolution and Transmission Inference in the Maghreb Central Regions, *Arabian Journal for Science and Engineering* **46,** 93–102 (2021).

65. Hernandez-Ceron, N., Feng, Z. & Castillo-Chavez, C., Discrete epidemic models with arbitrary stage distributions and applications to disease control, *Bulletin of mathematical biology* **75,** 1716–1746 (2013).

66. Hickey, K., Khan, R. & Walsh, J., Parametric estimation of ocean surface currents with HF radar, *IEEE journal of oceanic engineering* **20,** 139–144 (1995).

67. Hiraoka, M., Kinoshita, Y., Higa, M., Tsubaki, S., Monotilla, A. P., Onda, A. & Dan, A., Fourfold daily growth rate in multicellular marine alga Ulva meridionalis, *Scientific reports* **10,** 12606 (2020).

68. Howe, J., The Rise of Crowdsourcing, *Wired magazine* **14,** 1–4, `https://sistemas-humano-computacionais.wdfiles.com/local--files/capitulo%3Aredes-sociais/Howe_The_Rise_of_Crowdsourcing.pdf` (2006).

69. Hu, Z., Wang, D.-P., Pan, D., He, X., Miyazawa, Y., Bai, Y., Wang, D. & Gong, F., Mapping surface tidal currents and Changjiang plume in the East China Sea from Geostationary Ocean Color Imager, *Journal of Geophysical Research: Oceans* **121,** 1563–1572 (2016).

70. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Li, G., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J. & Cao, B., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, *The lancet* **395,** 497–506 (2020).

71. Hunter, E., Mac Namee, B. & Kelleher, J., An open-data-driven agent-based model to simulate infectious disease outbreaks, *PloS one* **13,** e0208775 (2018).

72. Hunter, E., Mac Namee, B. & Kelleher, J. D., *An Open Data Driven Epidemiological Agent-Based Model for Irish Towns. in AICS* (2016), 92–103.

73. INSEE, *2016 population census: sub-municipal databases - IRIS* Distributed by ADISP, 2016, `http://www.progedo-adisp.fr/enquetes/XML/lil.php?lil=lil-1369`.

74. Isern-Fontanet, J., Ballabrera-Poy, J., Turiel, A. & García-Ladona, E., Remote sensing of ocean surface currents: A review of what is being observed and what is being assimilated, *Nonlinear Processes in Geophysics* **24,** 613–643 (2017).

75. Jacob, B. & Stanev, E. V., Understanding the impact of bathymetric changes in the German bight on coastal hydrodynamics: one step toward realistic morphodynamic Modeling, *Frontiers in Marine Science* **8,** 640214 (2021).

76. Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L. & Daszak, P., Global trends in emerging infectious diseases, *Nature* **451,** 990–993 (2008).

77. Kang, D., Choi, H., Kim, J.-H. & Choi, J., Spatial epidemic dynamics of the COVID-19 outbreak in China, *International Journal of Infectious Diseases* **94,** 96–102 (2020).

78. Kang, S. K., Lee, S.-R. & Yum, K.-D., *in Elsevier oceanography series* 25–48 (Elsevier, 1991).

79. Karafyllidis, I. & Thanailakis, A., A Model for Predicting Forest Fire Spreading Using Cellular Automata, *Ecological Modelling* **99,** 87–97 (1997).

80. Kassem, A. Z. E., Does temperature affect COVID-19 transmission?, *Frontiers in public health* **8,** 554964 (2020).

81. Keita, E. B., *Modèles physiques et perception, contributions à l'analyse du milieu sonore urbain* PhD thesis (University of Brest, Brest, France, 2015).

82. Keita, E., Monthé, V. & Pottier, B., Discrete Simulation of Sound Propagation in the City Based on Cellular Automaton, *DEStech Transactions on Computer Science and Engineering* (2017).

83. Keshava, A., *Climate model simulation crashes data set* accessed on 13 June 2023, May 2020, `https://www.kaggle.com/datasets/analakeshava/climate-model-simulation-crashes-data-set`.

84. Kitchin, R., *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (pages 1-10, Conceptualising Data) (Sage, 2014).

85. Kostkova, P., Brewer, H., De Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., Koczan, P., Knight, P., Marsolier, C., McKendry, R. A., Ross, E., Sasse, A., Sullivan, R., Chaytor, S., Stevenson, O., Velho, R. & Tooke, J., Who owns the data? Open data for healthcare, *Frontiers in public health* **4,** 7 (2016).

86. Kowaliw, T., Grogono, P. & Kharma, N., *Bluenome: A Novel Developmental Model of Artificial Morphogenesis in Genetic and Evolutionary Computation Conference* (2004), 93–104.

87. Kremer, C., Torneri, A., Boesmans, S., Meuwissen, H., Verdonschot, S., Driessche, K. V., Althaus, C. L., Faes, C. & Hens, N., Quantifying superspreading for COVID-19 using Poisson mixture distributions, *Scientific Reports* **11,** 1–11 (2021).

88. Lacey, S. W., Cholera: Calamitous Past, Ominous Future, *Clinical Infectious Diseases* **20,** 1409–1419 (1995).

89. Lam, B. H., *Réseaux de capteurs sans fil pour l'observation du climat et de la biologie dans une région tropicale d'agriculture intensive: méthodes, outils et applications pour le cas du Delta du Mékong, Vietnam* PhD thesis (University of Brest, Brest, France, 2018).

90. Lam, B. H., Truong, T. P., Nguyen, K. M., Huynh, H. X. & Pottier, B., *An Hierarchical Scheduled Algorithm for Data Dissemination in a Brown Planthopper Surveillance Network in Nature of Computation and Communication: Second International Conference, ICTCC 2016, Rach Gia, Vietnam, March 17-18, 2016, Revised Selected Papers 2* (2016), 246–263.

91. Lamport, L. & Lynch, N., *Chapter on distributed computing* tech. rep. (Massachusetts Inst of Tech Cambridge Lab for Computer Science, 1989), `https://groups.csail.mit.edu/tds/papers/Lynch/handbook-chapter.pdf`.

92. Laskowski, M., Mostaço-Guidolin, L. C., Greer, A. L., Wu, J. & Moghadas, S. M., The impact of demographic variables on disease spread: influenza in remote communities, *Scientific reports* **1,** 105 (2011).

93. Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., Rossi, L., Manganelli, R., Loregian, A., Navarin, N., *et al.,* Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo', *Nature* **584,** 425–429 (2020).

94. Le Gendre, R., Morin, J., Maheux, F., Fournier, F., Simon, B., Cochard, M., Pierre-Duplessix, O., Dumas, F., Harmel, B., Paul, C., *et al.,* DILEMES-DIspersion LarvairE de Mytilus Edulis en baie de Seine, *Rapport IFREMER et CRPBN, France,* `https://archimer.ifremer.fr/doc/00188/29916/` (2014).

95. Lee, S., Oh, K.-H., Jang, S.-T., You, H. Y., Park, J. & Song, K.-M., M2 Tidal Current Estimation from One-day Observation Data off the Western and Southern Coasts of Korea, *Ocean Science Journal* **54,** 39–50 (2019).

96. Lewis, B. & Berg, D. J., *Multithreaded programming with Pthreads* (Prentice-Hall, Inc., 1998).

97. Li, R., Richmond, P. & Roehner, B. M., Effect of population density on epidemics, *Physica A: Statistical Mechanics and its Applications* **510,** 713–724 (2018).

98. Li, X., Zhang, Y., Liu, X. & Chen, Y., Assimilating Process Context Information of Cellular Automata into Change Detection for Monitoring Land Use Changes, *International Journal of Geographical Information Science* **26,** 1667–1687 (2012).

99. Li, Z., Liu, J., Mei, C., Shao, W., Wang, H. & Yan, D., Comparative analysis of building representations in TELEMAC-2D for flood inundation in idealized urban districts, *Water* **11,** 1840 (2019).

100. Liaw, A. & Wiener, M., randomForest: Breiman and Cutler's random forests for classification and regression, *R package version* **4,** 14, `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf` (2015).

101. Liew, A., Understanding Data, Information, Knowledge and Their Inter-relationships, *Journal of Knowledge Management Practice* **8,** 1–16, `http://www.tlainc.com/articl134.htm` (2007).

102. Lim, C. C., Yoon, J., Reynolds, K., Gerald, L. B., Ault, A. P., Heo, S. & Bell, M. L., Harmful algal bloom aerosols and human health, *EBioMedicine* **93,** 1–23 (2023).

103. Lin, T., Liu, X., Song, J., Zhang, G., Jia, Y., Tu, Z., Zheng, Z. & Liu, C., Urban waterlogging risk assessment based on internet open data: A case study in China, *Habitat International* **71,** 88–96 (2018).

104. Lizier, J. T., *The local information dynamics of distributed computation in complex systems* (Chapter 2. Computation in Complex Systems), PhD thesis (The University of Sydney, 2012).

105. Long, Y. & Liu, L., Transformations of urban studies and planning in the big/open data era: A review, *International Journal of Image and Data Fusion* **7,** 295–308 (2016).

106. Lucas, P.-Y., *Modélisations, Simulations, Synthèses pour des réseaux dynamiques de capteurs sans fil* PhD thesis (University of Brest, Brest, France, Dec. 2016).

107. Lundberg, S. M. & Lee, S.-I., *A Unified Approach to Interpreting Model Predictions in Advances in Neural Information Processing Systems* (eds Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R.) **30** (Curran Associates, Inc., 2017), https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

108. Lynch, N. A., *Distributed algorithms* Chapter 2, pages 17-23, https://dl.acm.org/doi/pdf/10.5555/2821576 (Elsevier, San Francisco, California, 1996).

109. Mecenas, P., Bastos, R. T. d. R. M., Vallinoto, A. C. R. & Normando, D., Effects of temperature and humidity on the spread of COVID-19: A systematic review, *PLoS one* **15,** e0238339 (2020).

110. Michailidis, P. D. & Margaritis, K. G., Scientific computations on multi-core systems using different programming frameworks, *Applied Numerical Mathematics* **104,** 62–80 (2016).

111. Miller, J. F., *Evolving Developmental Programs for Adaptation, Morphogenesis, and Self-repair in European Conference on Artificial Life* (2003), 256–265.

112. Miller, T., Explanation in artificial intelligence: Insights from the social sciences, *Artificial intelligence* **267,** 1–38 (2019).

113. Molnar, C., *Interpretable machine learning* (Lulu.com, 2020).

114. Molnar, C., Casalicchio, G. & Bischl, B., iml: An R package for interpretable machine learning, *Journal of Open Source Software* **3,** 786 (2018).

115. Moore, E. F., *Machine models of self-reproduction in Proceedings of symposia in applied mathematics* **14** (1962), 17–33.

116. Moreno, N., Wang, F. & Marceau, D. J., Implementation of a dynamic neighborhood in a land-use vector-based cellular automata model, *Computers, Environment and Urban Systems* **33,** 44–54 (2009).

117. Mount, D. M., *Geometric Intersection in Handbook of Discrete and Computational Geometry, chapter 33* (1997), https://www2.cs.sfu.ca/~binay/2014/813/Mount-GoodmanORourkeHandbookOfDiscreteMath.pdf.

118. Murray-Rust, P., Open data in science, *Nature Precedings,* 1–1 (2008).

119. O'Sullivan, D., Exploring spatial process dynamics using irregular cellular automaton models, *Geographical Analysis* **33,** 1–18 (2001).

120. Okabe, A., Boots, B. & Sugihara, K., Nearest Neighbourhood Operations with Generalized Voronoi Diagrams: A Review, *International Journal of Geographical Information Systems* **8,** 43–71 (1994).

121. Olami, Z., Feder, H. J. S. & Christensen, K., Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes, *Physical review letters* **68,** 1244 (1992).

122. Ortmann, J., Limbu, M., Wang, D. & Kauppinen, T., *Crowdsourcing linked open data for disaster management in Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web in conjunction with the ISWC* (2011), 11–22.

123. Packard, N. H. & Wolfram, S., Two-dimensional Cellular Automata, *Journal of Statistical physics* **38,** 901–946 (1985).

124. Pan, Z. & Reggia, J. A., Computational Discovery of Instructionless Self-replicating Structures in Cellular Automata, *Artificial Life* **16,** 39–63 (2010).

125. Pavlík, J., Hrnčírová, M., Stočes, M., Masner, J. & Vaněk, J., Usability of IoT and open data repositories for analyzing water pollution. A case study in the Czech Republic, *ISPRS International Journal of Geo-Information* **9,** 591 (2020).

126. Perrot, T., Rossi, N., Ménesguen, A. & Dumas, F., Modelling green macroalgal blooms on the coasts of Brittany, France to enhance water quality management, *Journal of Marine Systems* **132,** 38–53 (2014).

127. Piret, J. & Boivin, G., Pandemics throughout history, *Frontiers in microbiology* **11,** 631736 (2021).

128. Prehofer, C. & Bettstetter, C., Self-organization in communication networks: principles and design paradigms, *IEEE Communications magazine* **43,** 78–85 (2005).

129. Preparata, F. P. & Shamos, M. I., *Computational Geometry: An Introduction* Chapter 7, `http://www.cs.kent.edu/~dragan/CG/CG-Book.pdf` (Springer Science & Business Media, 2012).

130. Radočaj, D., Obhođaš, J., Jurišić, M. & Gašparović, M., Global open data remote sensing satellite missions for land monitoring and conservation: A review, *Land* **9,** 402 (2020).

131. Rezník, T., Charvát, K., Lukas, V., Charvát Jr, K., Horáková, Š. & Kepka, M., *Open data model for (precision) agriculture applications and agricultural pollution monitoring in EnviroInfo and ICT for Sustainability 2015* (2015), 97–107.

132. Rouen, A., Adda, J., Roy, O., Rogers, E. & Lévy, P., COVID-19: relationship between atmospheric temperature and daily new cases growth rate, *Epidemiology & Infection* **148,** e184 (2020).

133. Roy, I., The role of temperature on the global spread of COVID-19 and urgent solutions, `https://discovery.ucl.ac.uk/id/eprint/10140885/1/Roy_COVID_IJEST.pdf` (2020).

134. Salcido, A., A lattice gas model for infection spreading: Application to the COVID-19 pandemic in the Mexico City Metropolitan Area, *Results in Physics* **20,** 103758 (2021).

135. Sankaran, M., 275. note: The discrete poisson-lindley distribution, *Biometrics,* 145–149 (1970).

136. Saunders-Hastings, P. R. & Krewski, D., Reviewing The History of Pandemic Influenza: Understanding Patterns of Emergence and Transmission, *Pathogens* **5,** 66 (2016).

137. Schimit, P. H., A model based on cellular automata to estimate the social isolation impact on COVID-19 spreading in Brazil, *Computer methods and programs in biomedicine* **200,** 105832 (2021).

138. SGS-THOMSON Microelectronics Limited, U., *Occam 2.1 reference manual* tech. rep. (SGS-THOMSON Microelectronics, May 1995), `https://www.wotug.org/occam/documentation/oc21refman.pdf`.

139. Shamos, M. I. & Hoey, D., *Geometric Intersection Problems in 17th Annual Symposium on Foundations of Computer Science (sfcs 1976)* (1976), 208–215.

140. Shapley, L. S., A value for n-person games, *Contributions to the Theory of Games* **2,** 307–317, `https://apps.dtic.mil/sti/pdfs/AD0604084.pdf` (1953).

141. SHOM, *Courants de Marée des Côtes de France (Manche/Atlantique)* accessed on 01 April 2023, `https://diffusion.shom.fr/media/wysiwyg/pdf/courants%5C_2d%5C_notice.pdf`.

142. Silva, T., Wuwongse, V. & Sharma, H. N., Disaster mitigation and preparedness using linked open data, *Journal of Ambient Intelligence and Humanized Computing* **4,** 591–602 (2013).

143. Sirakoulis, G. C., Karafyllidis, I. & Thanailakis, A., A Cellular Automaton Model for The Effects of Population Movement and Vaccination on Epidemic Propagation, *Ecological Modelling* **133,** 209–223 (2000).

144. Snow, J., Report on The Cholera Outbreak in The Parish of St James, Westminster During The Autumn of 1854, *Medical Times,* 39–54, `http://resource.nlm.nih.gov/34721190R` (1854).

145. Söderman, D. & Gibson, J., *A Binary Universal Form for The Representation of Meteorological Data - an Introduction to FM 94 BUFR in ECMWF/WMO Workshop on Radiosonde Data Quality and Monitoring, 14-16 December 1987* (ECMWF, Shinfield Park, Reading, 1987), 249–260, `https://www.ecmwf.int/node/12522`.

146. Stevens, D. & Dragićević, S., A GIS-based Irregular Cellular Automata Model of Land-use Change, *Environment and Planning B: Planning and Design* **34,** 708–724 (2007).

147. Sun, S., Wang, F., Li, C., Qin, S., Zhou, M., Ding, L., Pang, S., Duan, D., Wang, G., Yin, B., *et al.,* Emerging challenges: Massive green algae blooms in the Yellow Sea, *Nature Precedings,* 1–1 (2008).

148. Suri, S., Hubbard, P. M. & Hughes, J. F., Analyzing Bounding Boxes for Object Intersection, *ACM Transactions on Graphics (TOG)* **18,** 257–277 (1999).

149. *Sustainability of Digital Formats: Planning for Library of Congress Collections* accessed on 08 December 2022, Dec. 2016, `https://www.loc.gov/%20preservation/digital/formats/fdd/fdd000449.shtml`.

150. Sy, K. T. L., White, L. F. & Nichols, B. E., Population density and basic reproductive number of COVID-19 across United States counties, *PloS one* **16,** e0249271 (2021).

151. Tatem, A. J., WorldPop, open data for spatial demography, *Scientific data* **4,** 1–4 (2017).

152. Thorpe, W., A Guide to The WMO Code Form FM 94 BUFR, `http://dss.ucar.edu/docs/formats/bufr/bufr.pdf` (1995).

153. Toffoli, T. & Margolus, N., *Cellular automata machines: a new environment for modeling* Chapter 11, pages 109-114, https://people.csail.mit.edu/nhm/cam-book.pdf (MIT press, 1987).

154. Truong, M. T. T., *Modeling and simulation for the surveillance of water systems, application of cellular automata to environment monitoring and control* PhD thesis (University of Brest, Brest, France, 2020).

155. Truong, M. T. T., Yazdani, S. S., Pottier, B., Rodin, V. & Huynh, X. H., *Multiscale Geographic Exploration, Observation, Simulation, and Representation in 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (2019), 1–8.

156. Truong, T. P., *Simulation and compiler support for communication and mobility for environment sensing* PhD thesis (University of Brest, Brest, France, 2018).

157. Truong, T. P., Pottier, B. & Huynh, H. X., Cellular Simulation for Distributed Sensing over Complex Terrains, *Sensors* **18,** 2323 (2018).

158. Uhlir, P. F. & Schröder, P., Open data for global science, *Data Science Journal* **6,** OD36–OD53 (2007).

159. Varlık, N., Rethinking the history of plague in the time of COVID-19, *Centaurus* **62,** 285–293 (2020).

160. Vellido, A., Martín-Guerrero, J. D. & Lisboa, P. J., *Making machine learning models interpretable in European Symposium on Artificial Neural Networks* **12** (2012), 163–172.

161. Von Neumann, J., *Theory of Self-reproducing Automata* (ed Burks, A. W.) https://cdn.patentlyo.com/media/docs/2012/04/VonNeumann.pdf (University of Illinois Press, 1966).

162. Von Neumann, J., *in Systems Research for Behavioral Science Research* 97–107 (Routledge, 2017), https://www.eecs.ucf.edu/~dcm/Teaching/COP5611-Spring2013/Papers/Old/vonNeumannSelfReproducingAutomata.pdf.

163. Voronoi, G., Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelles Journal)* **1908,** 198–287 (1908).

164. Wang, S., Mu, L., Qi, M., Yu, Z., Yao, Z. & Zhao, E., Quantitative risk assessment of storm surge using GIS techniques and open data: A case study of Daya Bay Zone, China, *Journal of environmental management* **289,** 112514 (2021).

165. White, R. & Engelen, G., Cellular Automata and Fractal Urban Form: A Cellular Modelling Approach to The Evolution of Urban Land-use Patterns, *Environment and planning A* **25,** 1175–1199 (1993).

166. White, S. H., Del Rey, A. M. & Sánchez, G. R., Modeling epidemics using cellular automata, *Applied mathematics and computation* **186,** 193–202 (2007).

167. Windom, H. L., Contamination of the marine environment from land-based sources, *Marine Pollution Bulletin* **25,** 32–36 (1992).

168. Wolfram, S., *Cellular Automata as Simple Self-organizing Systems* tech. rep. (Calif. Inst. Technol., Pasadena, CA, 1982), `http://cds.cern.ch/record/140047`.

169. Wolfram, S., Cellular Automata as Models of Complexity, *Nature* **311,** 419–424 (1984).

170. Wolfram, S., Computation Theory of Cellular Automata, *Communications in Mathematical Physics* **96,** 15–57 (1984).

171. Wolfram, S., Universality and Complexity in Cellular Automata, *Physica D: Nonlinear Phenomena* **10,** 1–35 (1984).

172. Wolfram, S., Cellular automaton fluids 1: Basic theory, *Journal of Statistical Physics* **45,** 471–526 (1986).

173. Wu, J., Chen, H., Orlandi, F., Lee, Y. H., O'Sullivan, D. & Dev, S., *An interoperable open data portal for climate analysis in 2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)* (2021), 104–105.

174. Yan, P., *in Mathematical epidemiology* 229–293 (Springer, 2008).

175. Yang, H., Choi, J.-K., Park, Y.-J., Han, H.-J. & Ryu, J.-H., Application of the Geostationary Ocean Color Imager (GOCI) to estimates of ocean surface currents, *Journal of Geophysical Research: Oceans* **119,** 3988–4000 (2014).

176. Yassemi, S., Dragićević, S. & Schmidt, M., Design and Implementation of an Integrated GIS-based Cellular Automata Model to Characterize Forest Fire Behaviour, *Ecological Modelling* **210,** 71–84 (2008).

177. Ye, N.-h., Zhang, X.-w., Mao, Y.-z., Liang, C.-w., Xu, D., Zou, J., Zhuang, Z.-m. & Wang, Q.-y., 'Green tides' are overwhelming the coastline of our blue planet: taking the world's largest example, *Ecological Research* **26,** 477–485 (2011).

178. Zalesny, V. B., Gusev, A. V., Lukyanova, A. N. & Fomin, V. V., Numerical modelling of sea currents and tidal waves, *Russian Journal of Numerical Analysis and Mathematical Modelling* **31,** 115–125 (2016).

179. Zhang, Y., Sensitivity Analysis in Simulating Oil Slick Using CA Model, *Ocean Engineering* **218,** 108216 (2020).

180. Zhang, Y., Li, X., Liu, X. & Qiao, J., Self-modifying CA Model Using Dual Ensemble Kalman Filter for Simulating Urban Land-use Changes, *International Journal of Geographical Information Science* **29,** 1612–1631 (2015).

181. Zhang, Y., Qiao, J., Wu, B., Jiang, W., Xu, X. & Hu, G., *Simulation of Oil Spill Using ANN and CA Models in 2015 23rd International Conference on Geoinformatics* (2015), 1–5.

182. Zhang, Y., Qiao, J., Wu, B., Jiang, W., Xu, X. & Hu, G., *Simulation of Oil Spill Using Logistic-regression CA Model in 2015 23rd International Conference on Geoinformatics* (2015), 1–6.

183. Zhong, W., Altun, G., Tian, X., Harrison, R., Tai, P. C. & Pan, Y., Parallel protein secondary structure prediction schemes using Pthread and OpenMP over hyperthreading technology, *The Journal of Supercomputing* **41,** 1–16 (2007).

184. Zhou, Y. & Suri, S., Analysis of a Bounding Box Heuristic for Object Intersection, *Journal of the ACM (JACM)* **46,** 833–857, `http://euro.ecom.cmu.edu/people/faculty/mshamos/1976GeometricIntersection.pdf` (1999).

185. Zietz, B. P. & Dunkelberg, H., The history of the plague and the research on the causative agent Yersinia pestis, *International journal of hygiene and environmental health* **207,** 165–178 (2004).

Appendices

1 Storing and processing geographical data

1.1 POSTGIS

PostGIS is a spatial database extender for PostgreSQL object-relational database. It adds support for geographical objects allowing location queries to be run in SQL. PostGIS supports all the objects and functions specified in the OpenGIS Consortium's [2] "Simple Features for SQL" specification. Examples of the text representations of the spatial objects of the features are as follows.

- POINT(0 0)
- LINESTRING(0 0,1 1,1 2)
- POLYGON((0 0,4 0,4 4,0 4,0 0),(1 1, 2 1, 2 2, 1 2,1 1))
- MULTIPOINT((0 0),(1 2))
- MULTILINESTRING((0 0,1 1,1 2),(2 3,3 2,5 4))
- MULTIPOLYGON(((0 0,4 0,4 4,0 4,0 0),(1 1,2 1,2 2,1 2,1 1)), ((-1 -1,-1 -2,-2 -2,-2 -1,-1 -1)))
- GEOMETRYCOLLECTION(POINT(2 3),LINESTRING(2 3,3 4)

The shape files of French administrative divisions are downloaded and stored in local PostGIS database. There are three main attributes that appear in all the shapefiles, which are code_insee, nom, and geom. Code_insee is an identification provided by INSEE, nom is the name of the division, and geom is the geographical data of the division. The first thing to do is to create a spatial table to store data using "CREATE TABLE" SQL statement with a column of type geometry. The following example creates a table for storing department information in France (in Listing 1).

Listing 1 – SQL statement to create a table storing department information in France.

```
1 CREATE TABLE fra_dept(
2             gid       serial PRIMARY KEY,
3             code_insee        VARCHAR(80),
4             nom       VARCHAR(80),
5             nut3      VARCHAR(80),
6             wikipedia         VARCHAR(80),
7             surf_km2          NUMERIC,
8             latitude          REAL,
```

2. https://www.ogc.org accessed on 01 November 2022

```
 9            longitude          REAL,
10            neighbors          VARCHAR(80),
11            geom       geometry(MULTIPOLYGON,4326)
12  );
```

```
gisdata=# \d fra_dept
                                  Table "public.fra_dept"
   Column   |           Type            | Collation | Nullable |               Default
------------+---------------------------+-----------+----------+--------------------------------------
 gid        | integer                   |           | not null | nextval('fra_dept_gid_seq'::regclass)
 code_insee | character varying(80)     |           |          |
 nom        | character varying(80)     |           |          |
 nuts3      | character varying(80)     |           |          |
 wikipedia  | character varying(80)     |           |          |
 surf_km2   | numeric                   |           |          |
 latitude   | double precision          |           |          |
 longitude  | double precision          |           |          |
 neighbors  | character varying(80)     |           |          |
 geom       | geometry(MultiPolygon,4326) |           |          |
Indexes:
    "fra_dept_pkey" PRIMARY KEY, btree (gid)
    "fra_dept_geom_idx" gist (geom)
```

Figure 1 – Created fra_dept table in PostGIS database.

Once the spatial table has been created (shown in Figure 1), we can load data into the database using the shapefile loader. The **shp2pgsql** data loader converts Shapefiles into SQL suitable for insertion into a PostGIS/PostgreSQL database with several operating modes selected by command line flags (-I: create a GiST index on the geometry column; -s: creates and populates the geometry tables with the specified SRID). Then, data can be bulk-loaded into database by loading a text file of SQL INSERT statements using the **psql** SQL utility (-d: database; -U: users; -f: sql file).

```
shp2psql -I -s 4326 /home/ubuntu/shape/departements-20180101.shp fra-dept
> dept.sql
psql -h localhost -d gisdata -U postgres -f dept.sql
```

Spatial data can be extracted from the database using either SQL or the Shapefile dumper. The **pgsql2shp** table dumper connects to the database and converts a table (possibly defined by a query) into a shape file.

```
pgsql2shp [option] database tablename -f tofilename.shp
```

The most straightforward way is to use a SQL SELECT query to define the data set to be extracted and dump the resulting columns into a parsable text file. The **psql** is a terminal-based front-end to PostgreSQL enabling to query interactively. The command 'c' is used to connect to an existing database providing the database name.

```
psql -U postgres
\c gisdata
```

Some examples of SQL queries using the fra_dept table are shown in Listing 2.

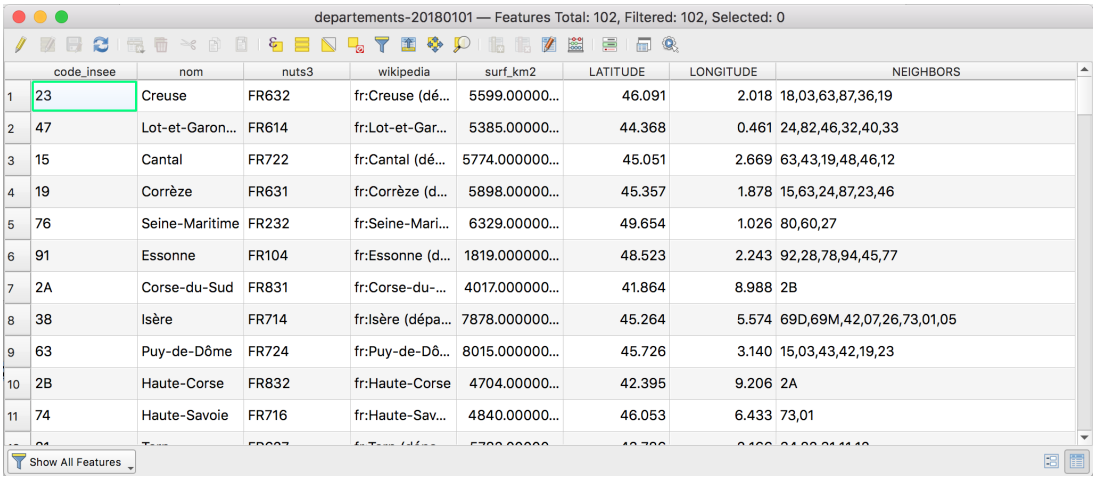
Listing 2 – Some examples of SQL queries in PostGIS database.

```
1 SELECT nom, ST_Area(geom, true) sqm FROM fra_dept;
2 SELECT nom, ST_AsText(ST_Centroid(geom,true)) centroid
3          FROM fra_dept;
4 SELECT code_insee, nom FROM fra_dept
5          WHERE ST_Intersects(geom,
6          (SELECT geom FROM fra_dept WHERE code_insee='83'));
```

The **ST_Intersects** function tells whether two geometries share any space. This function is used for calculating the neighborhood of a division in PostGIS database.

1.2 QGIS

QGIS is a free and open-source GIS application that supports viewing, editing, printing, and analysis of geospatial data. To import shapefile into QGIS, we need to do several steps.



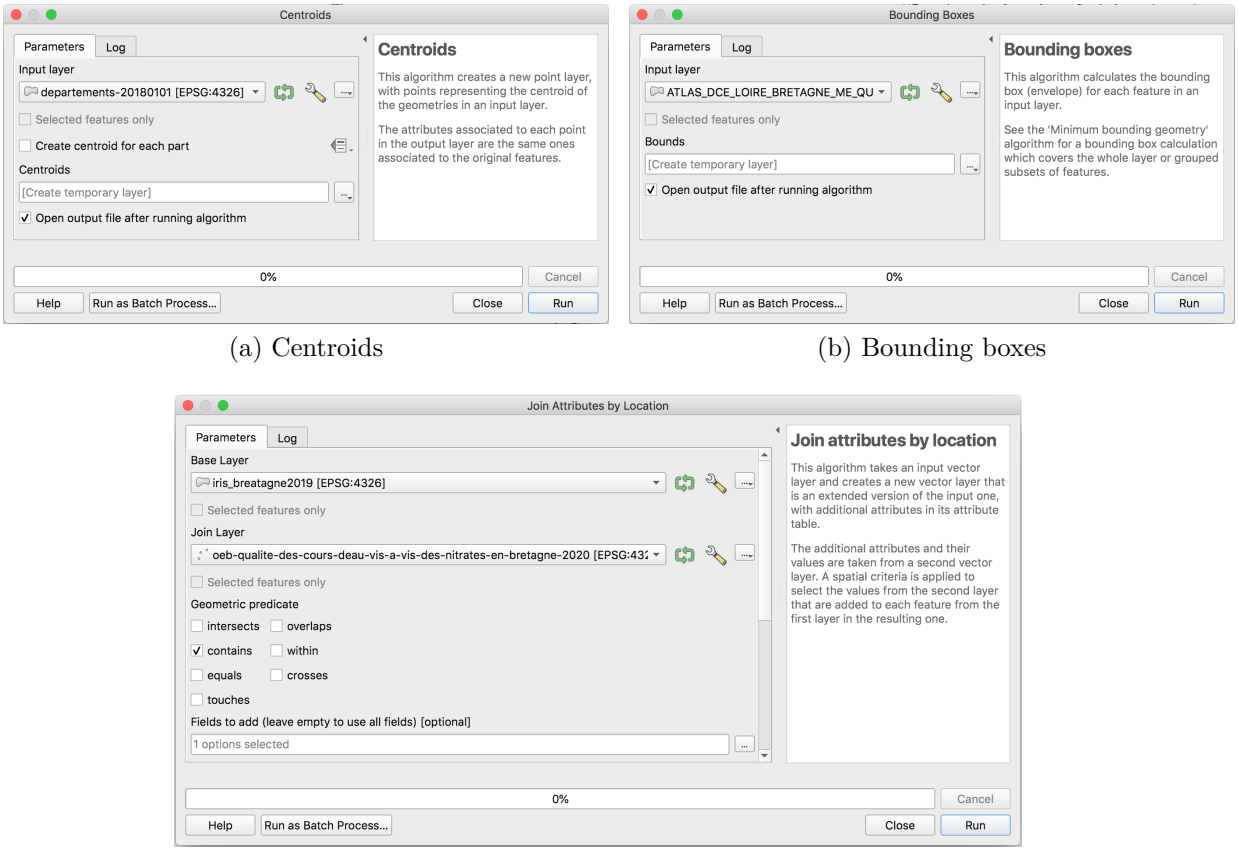
	code_insee	nom	nuts3	wikipedia	surf_km2	LATITUDE	LONGITUDE	NEIGHBORS
1	23	Creuse	FR632	fr:Creuse (dé...	5599.00000...	46.091	2.018	18,03,63,87,36,19
2	47	Lot-et-Garon...	FR614	fr:Lot-et-Gar...	5385.00000...	44.368	0.461	24,82,46,32,40,33
3	15	Cantal	FR722	fr:Cantal (dé...	5774.000000...	45.051	2.669	63,43,19,48,46,12
4	19	Corrèze	FR631	fr:Corrèze (d...	5898.00000...	45.357	1.878	15,63,24,87,23,46
5	76	Seine-Maritime	FR232	fr:Seine-Mari...	6329.00000...	49.654	1.026	80,60,27
6	91	Essonne	FR104	fr:Essonne (d...	1819.000000...	48.523	2.243	92,28,78,94,45,77
7	2A	Corse-du-Sud	FR831	fr:Corse-du-...	4017.000000...	41.864	8.988	2B
8	38	Isère	FR714	fr:Isère (dépa...	7878.000000...	45.264	5.574	69D,69M,42,07,26,73,01,05
9	63	Puy-de-Dôme	FR724	fr:Puy-de-Dô...	8015.000000...	45.726	3.140	15,03,43,42,19,23
10	2B	Haute-Corse	FR832	fr:Haute-Corse	4704.00000...	42.395	9.206	2A
11	74	Haute-Savoie	FR716	fr:Haute-Sav...	4840.00000...	46.053	6.433	73,01

Figure 2 – French departments shapefiles shown as QGIS table.

1. Click on the “Layer” menu, click the “Add Vector Layer” button

2. Click on the “Browse” button and navigate to the folder of shapefile (need to be sure the data type is selected to shapefile)
3. Click “Open”

Once the vector layer is added, we can view, zoom, and edit the data in QGIS. Figure 2 shows the department information in France as QGIS table. The attributes with values are shown in the table with scrolling option to view all data in the shapefile. The only information cannot be shown in the table is the geometry column, which is visualized as map in the main canvas.



(a) Centroids

(b) Bounding boxes

(c) Join attributes by location

Figure 3 – Examples of vector processing tools in QGIS.

Geographical data processing is quite simple with QGIS since it provides a bunch of processing tools on the shapefile. For example, the Centroids tool takes a shapefile as an input layer and creates new layer with points representing the centroid of the geometries in the input layer (Figure 3a). The Bounding boxes tool calculates the bounding box of

each geometry in the input layer (Figure 3b). The Join attributes by location tool receives two input layers. The first one is the base layer and the second one is the join layer. It will calculate the relation between the layers based on several geometric predicates such as 'intersects' or 'contains'. Any feature in the base layer meets the given conditions will be extended with the attributes provided in the join layer.

Especially, QGIS has been designed with a plugin architecture. Plugins can be written in Python, a common language in the geospatial world, to let users interact with geographical objects. Besides, QGIS also provides the Python Console, which is an interactive shell for the python command executions. Listing 3 shows python code to calculate the neighborhood for each geometry in the shapefile and append a new field (namely neighbors) into the attribute table. At the same time, it also counts the number of neighbors for each geometry and put into a new field (namely sum).

Listing 3 – A python module can be used in QGIS to calculate adjacent neighbors for each object in a shapefile and write information back to the shapefile.

```
1  from qgis.utils import iface
2  from PyQt5.QtCore import QVariant
3  _NAME_FIELD = 'code_insee'
4  _SUM_FIELD = 'code_insee'
5  _NEW_NEIGHBORS_FIELD = 'neighbors'
6  _NEW_SUM_FIELD = 'sum'
7  layer = iface.activeLayer()
8  layer.startEditing()
9  layer.dataProvider().addAttributes(
10           [QgsField(_NEW_NEIGHBORS_FIELD, QVariant.String)
11           , QgsField(_NEW_SUM_FIELD, QVariant.Int)])
12 layer.updateFields()
13 feature_dict = {f.id(): f for f in layer.getFeatures()}
14 index = QgsSpatialIndex()
15 for f in feature_dict.values():
16   index.insertFeature(f)
17 for f in feature_dict.values():
18   print ('Working on '+ str(f[_NAME_FIELD]))
19   geom = f.geometry()
20   intersecting_ids = index.intersects(geom.boundingBox())
```

```
21      neighbors = []
22      neighbors_sum = 0
23      for intersecting_id in intersecting_ids:
24        intersecting_f = feature_dict[intersecting_id]
25        if (f != intersecting_f and
26          not intersecting_f.geometry().disjoint(geom)):
27          neighbors.append(intersecting_f[_NAME_FIELD])
28          neighbors_sum += 1
29      f[_NEW_NEIGHBORS_FIELD] = ','.join(neighbors)
30      f[_NEW_SUM_FIELD] = neighbors_sum
31      layer.updateFeature(f)
32  layer.commitChanges()
```

2 Decomposition of BUFR messages

BUFR is a strong and complex format with self-descriptive nature. There are description tables (as shown in Figure 4) to know which kind of data to be read, the units, and the length for that data. The descriptors appear in section 3 of a BUFR message composed of thee parts - F (2 bits), X (6 bits), and Y (8 bits). A data modeling process is necessary to define and analyze BUFR data.

Table Reference F X Y	Element Name	Units	Scale	Reference Value	Data Width (Bits)
0 01 001	WMO block number	numeric	0	0	7
0 01 002	WMO station number	numeric	0	0	10
0 02 001	Type of station	code table	0	0	2
0 04 001	Year	Year	0	0	12
0 04 002	Month	Month	0	0	4
0 04 003	Day	Day	0	0	6
0 04 004	Hour	Hour	0	0	5
0 04 005	Minute	Minute	0	0	6
0 05 002	Latitude (coarse accuracy)	Degree	2	-9000	15
0 06 002	Longitude (coarse accuracy)	Degree	2	-18000	16

Figure 4 – Example of BUFR descriptive tables.

In this work, we used Express modeling language to model the data so that it can be easy to transform into other machine-readable formats. Express is a standard data modeling language which is formalized in the ISO standard. It helps define data objects and the relationships between the objects. The data specification is clear, and it has the connections with both data modeling and object-oriented systems. A data model for BUFR messages can be defined as in Listing 4.

Listing 4 – Textual representation of BUFR schema used as input of some application-programming interfaces for reading, compiling, and generating data defined by the model.

```
SCHEMA bufr_schema;
        ENTITY bufr;
                s0 : section0;
                s1 : section1;
                s2 : section2;
                s3 : section3;
                s4 : section4;
                s5 : section5;
        END_ENTITY;
        ENTITY indication;
                numoctet : INTEGER;
                val : STRING;
                meaning : STRING;
        END_ENTITY;
        ENTITY detail;
                numoctet : INTEGER;
                val : INTEGER;
                description : STRING;
        END_ENTITY;
        ENTITY datetime;
                year : detail;
                month : detail;
                day : detail;
                hour : detail;
                minute : detail;
        END_ENTITY;
```

```
27          ENTITY section0;
28                  start : indication;
29                  len : detail;
30                  edition : detail;
31          END_ENTITY;
32          ENTITY section2;
33                  len : detail;
34                  reserved : detail;
35                  reservedLocal : detail;
36          END_ENTITY;
37          ENTITY section5;
38                  end_bufr : detail;
39          END_ENTITY;
40          ENTITY section1;
41                  len : detail;
42                  mstable : detail;
43                  center : detail;
44                  subcenter : detail;
45                  sequence : detail;
46                  section2flag : detail;
47                  category : detail;
48                  intercategory : detail;
49                  localcategory : detail;
50                  mstableversion : detail;
51                  localtableversion : detail;
52                  time : datetime;
53                  reserved : detail;
54          END_ENTITY;
55          ENTITY section3;
56                  len : detail;
57                  reserved : detail;
58                  subsets : detail;
59                  flagObCom : detail;
60                  desc : SET [1:?] OF descriptor;
```

```
61          END_ENTITY;
62          ENTITY section4;
63                  len : detail;
64                  reserved : detail;
65                  data : SET [1:?] OF realdata;
66          END_ENTITY;
67          ENTITY descriptor;
68                  fxy : STRING;
69                  f,x,y : INTEGER;
70                  element_name : STRING;
71                  note : STRING;
72                  unit : STRING;
73                  scale : STRING;
74                  refer : STRING;
75                  bitwidth : INTEGER;
76                  fxy1 : STRING;
77                  title : STRING;
78          END_ENTITY;
79          ENTITY realdata;
80                  name : STRING;
81                  val : STRING;
82          END_ENTITY;
83  END_SCHEMA;
```

The textual representation of this schema is important as an input for SDAI (Standard Data Access Interface), which is an abstract specification on how to deal with Express schema and can be mapped to various programming languages, which means that the programming code will be automatically generated. We use JSDAI[3], which is a Java application-programming interface for reading, compiling, and writing object-oriented data defined by an Express model. The java code with classes of objects is auto-generated with JSDAI API. It is further added some functions for reading raw binary data and export the result into commonly used and machine readable-format - CSV. The graphical representation of the schema is presented in Figure 5.

3. https://www.jsdai.net/ accessed on 08 December 2022

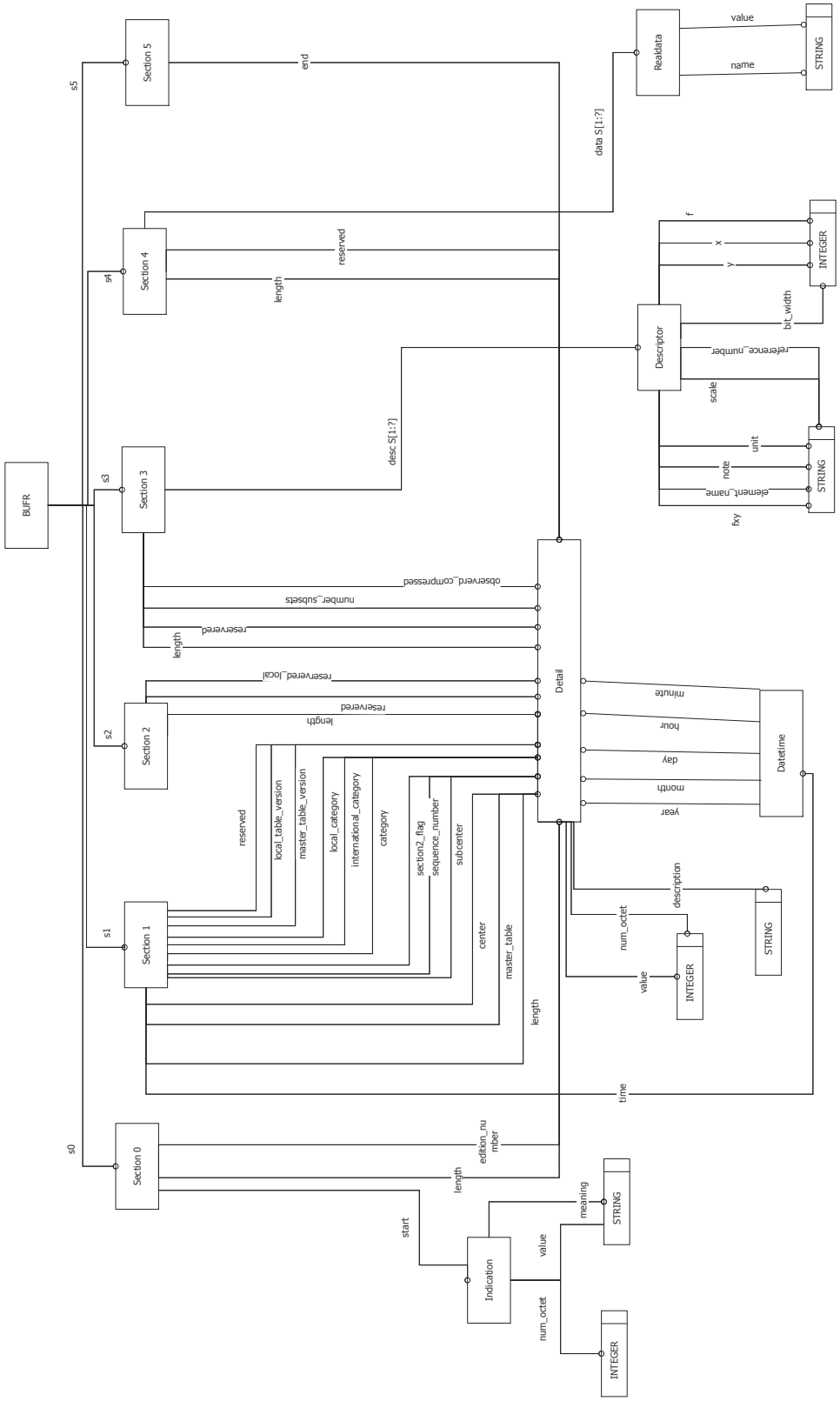


Figure 5 – The graphical representation of BUFR schema.

3 Process system generation using irregular cell space

QuickMap/PickCell/PickShape is the tool set developed by our research team that is built on Smalltalk programming language in VisualWorks [4]. This is a cross-platform implementation suitable for software developers who need to build applications quickly and efficiently with Smalltalk. Objects are employed to represent everything in this programming language, including all the conventional data types that may be found in any programming language such as integers and booleans.

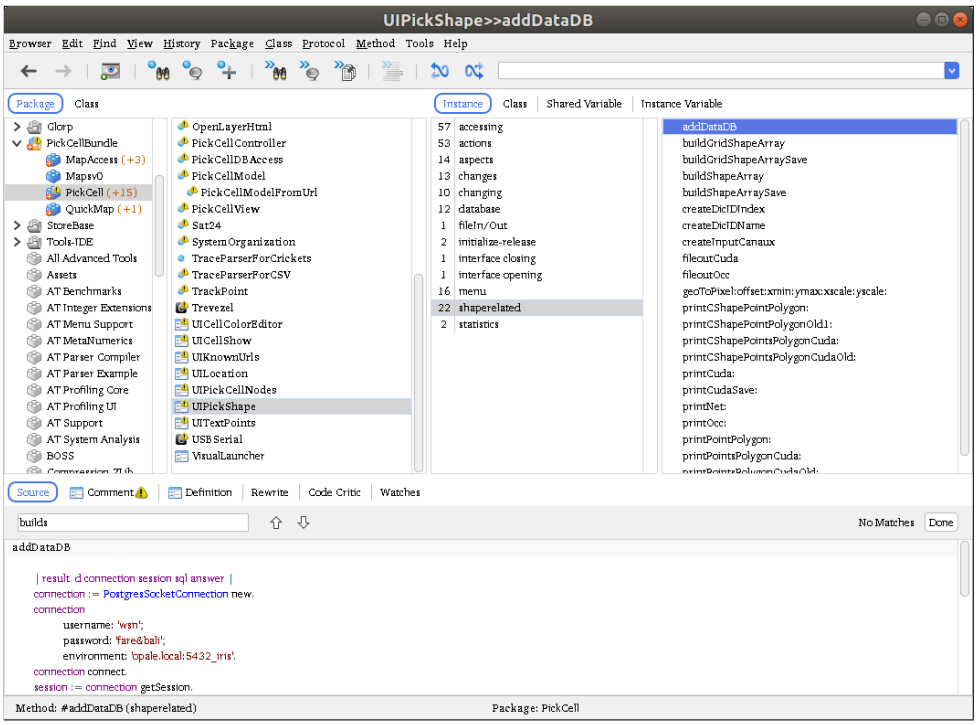


Figure 6 – PickShape tool implementation on VisualWorks with Smalltalk.

Shapefiles are read and visualized on QuickMap allowing zone selection. The Shapefile C Library [5] provides the ability to write simple C programs for reading, writing and updating (to a limited extent) ESRI Shapefiles, and the associated attribute file (.dbf).This library is used from Smalltalk to provide functions for shape readers. A function to read shape objects in Smalltalk is presented in Listing 5. The 'open' and 'openDBF' are called to open the .shp file and .dbf file. The identification of each shape is read from 'CODE_IRIS' field and nVertices contains all points made up the shape polygon.

4. https://www.cincomsmalltalk.com/main/products/visualworks/ accessed on 04 February 2023
5. http://shapelib.maptools.org accessed on 04 February 2023

Listing 5 – A function to read shape objects in Smalltalk using C Shapefile library.

```
readAllObjectsFromShapefile: aFilenameOrString

    | sr nb diviCol newDivi iCodeField |
    sr := self file: aFilenameOrString asString.
    sr open.
    sr openDBF. "put .dbf file in the same folder"
    iCodeField := sr getFieldIndexwithName: 'CODE_IRIS'.
    sr getInfo.
    nb := sr entities.
    diviCol := Dictionary new.
    (0 to: nb - 1) do: [:i |
  | shape positions nVertices key id longitude latitude |
      shape := sr readObject: i.
      nVertices := shape memberAt: #nVertices.
      key := shape memberAt: #nShapeId.
      id := sr readStringFieldshape: key iField: iCodeField.
      longitude := shape memberAt: #padfX.
      latitude := shape memberAt: #padfY.
      nVertices > 0
    ifTrue:
        [positions := OrderedCollection new: nVertices.
        (0 to: nVertices - 1) do:
      [:index |
          positions add: (GeoPosition lon: (longitude at: index)
                  lat: (latitude at: index)).
      ].
        newDivi := Batiment withPositions: positions id_division: id.
        diviCol at: id put: newDivi].
  ].
    sr close.
    \↑diviCol
```

After reading a shapefile and visualizing shape objects on QuickMap, users can zoom

and pan to select an area of interesting. PickShape provides PostGIS database connection to query all polygons with their centroids inside the selected frame based on the geographical coordinates (Listing 6). At the same time, all local data related to these geographical objects can be queried from database to include for the system generation.

Listing 6 – Database connection from Smalltalk to query all administrative divisions in selected frame.

```
connection := PostgresSocketConnection new.
connection
    username: '...';
    password: '...';
    environment: 'wsn.univ-brest.fr:8080_iris'.
connection connect.
session := connection getSession.

sql := 'select code_insee,nom,ST_Y(ST_Centroid(geom)),
  ST_X(ST_Centroid(geom)) from fra_iris where sum>0 and
  ST_Y(ST_Centroid(geom)) <= ',((self geoorigin y) asFloat)
  printString,' and ST_Y(ST_Centroid(geom)) >= ',
  ((self geocorner y) asFloat) printString,' and
  ST_X(ST_Centroid(geom)) <= ',((self geocorner x) asFloat)
  printString,' and ST_X(ST_Centroid(geom)) >= ',
  ((self geoorigin x) asFloat) printString.

temp1 := OrderedCollection new.
1 to: (collection size) do: [:i |
  d := collection at: i.
  sql := 'select population,area from fra_irisdata where
    code_insee=''',d id_division,''' '.
  session
    prepare: sql;
    execute.
  result1 := session answer.
  [result1 atEnd]
    whileFalse: [
```

```
29          temp:= result1 next.
30          d population: (temp at:1).
31          d area: (temp at: 2).
32        ].
33    sql := 'select neighbors from fra_iris where code_insee=''',
34              d id_division, ''' '.
35    session
36      prepare: sql;
37      execute.
38    result1 := session answer.
39    [result1 atEnd]
40      whileFalse: [
41        temp := result1 next.
42        (temp at: 1) ifNil: [temp1 add: d.]
43    ifNotNil: [
44      neighbors:= ( temp at: 1) asArrayOfSubstrings: ,.
45      neighbors do:[:each| (set includes: each)
46        ifTrue: [d neighborsAdd: each].
47      ].
48    ].
49    ].
50  ].
51  collection removeAll: temp1.
```

Another important function in PickShape is to generate process systems in Occam/C syntaxes. The generated codes will be used later for simulation. Developers only need to add functions to describe the data send/receive between processes and the state changes over time steps. Listing 7 shows a fragment code to write file in Occam syntax.

Listing 7 – Write a process system to file in Occam syntax

```
1  | fileName stream d arr index |
2  self createDicIDName.
3  fileName := (Filename named: aFileName).
4  stream :=  fileName newReadAppendStream.
5  stream nextPutAll: 'DATA TYPE Location'; cr.
```

```
 6 stream nextPutAll: '  RECORD'; cr.
 7 stream nextPutAll: '    REAL32 latitude:'; cr.
 8 stream nextPutAll: '    REAL32 longitude:'; cr.
 9 stream nextPutAll: '    REAL32 population:'; cr.
10 stream nextPutAll: '    REAL32 area:'; cr.
11 stream nextPutAll: '    REAL32 txstd:'; cr.
12 stream nextPutAll: ':'; cr; cr.
13 "... ... ..."
14 stream cr; nextPutAll: '#USE "course.lib"'; cr.
15 stream nextPutAll: 'VAL INT MaxFanOut IS ',maxfan printString,':';cr.
16 stream nextPutAll: 'VAL INT MaxNodes IS ',maxnode printString,':';cr.
17 stream cr; nextPutAll: '#INCLUDE "nodes-test-include.occ"'; cr.
18 "..."
19 stream cr; nextPutAll: '  --Channel table declarations'; cr.
20 1 to: (collection size) do: [:i |
21   d := collection at: i.
22   arr := d neighbors.
23     stream nextPutAll: '  ', (d process_name),'.out IS ['.
24     1 to: (arr size) do: [:j |
25       j=(arr size)
26         ifTrue: [stream nextPutAll: (d process_name), '.' ,
27           (dic at: (arr at: (arr size)) ).]
28         ifFalse: [stream nextPutAll: (d process_name), '.' ,
29           (dic at: (arr at: j) ), ', '.].
30     ].
31     stream nextPutAll: '] :'.
32     stream cr.
33     stream nextPutAll: '  ', (d process_name),'.in IS ['.
34     1 to: (arr size) do: [:j |
35       j=(arr size)
36         ifTrue: [stream nextPutAll: (dic at: (arr at: (arr size)) ),
37           '.' , (d process_name).]
38         ifFalse: [stream nextPutAll: (dic at: (arr at: j) ), '.' ,
39           (d process_name), ', '.].
```

```
40          ].
41          stream nextPutAll: '] :'.
42          stream cr.
43      ].
44      stream cr;nextPutAll: '  --Program body'; cr.
45      stream cr;nextPutAll: '  [MaxNodes]CHAN OF BYTE toMux:'; cr.
46      stream nextPutAll: '  PAR' ; cr.
47      index := 0.
48      1 to: (collection size) do: [:i |
49        d := collection at: i.
50          stream nextPutAll: '    Node( '.
51          stream nextPutAll: (d process_name), '.in, ',
52              (d process_name), '.out, '.
53          stream nextPutAll: (index printString), ', toMux[',
54              (index printString) , '] )'.
55          index := index + 1.
56          stream cr.
57      ].
58      stream nextPutAll: '    Mux( toMux, stdout)' ; cr.
59      stream nextPutAll: ':'.
60      stream close.
```

4 Simulation results

4.1 Covid-19 spreading in Brittany, France

Experiments were conducted in modeling Covid-19 spreading in Brittany, France. The incidence rate is showed in Figures 7 and 8.

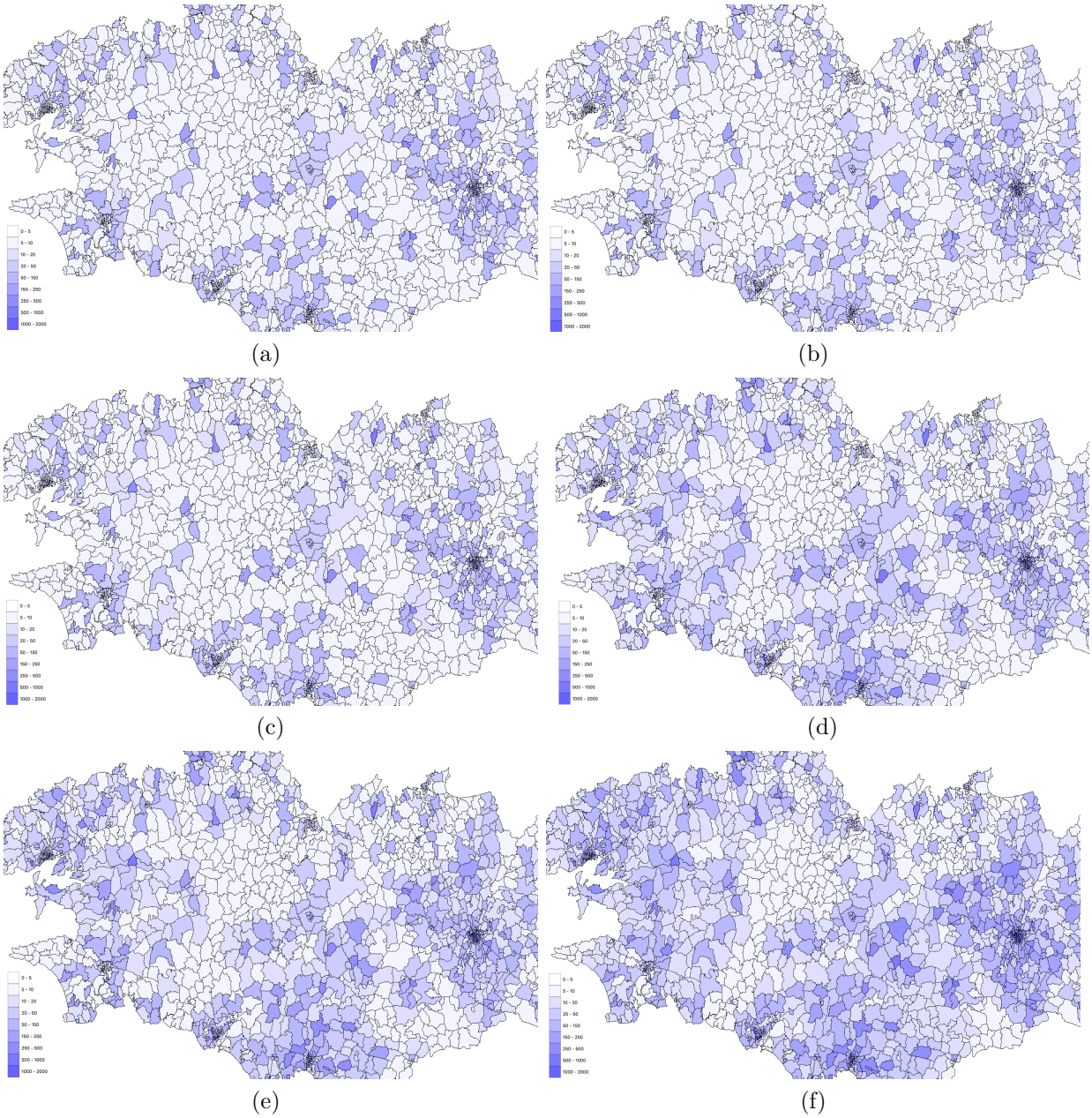


Figure 7 – Covid-19 incidence rate of Brittany, France over time. a) Incidence rate in 01 September 2021. b) Simulation results after 10 days. c) Simulation results after 20 days. d) Simulation results after 30 days. e) Simulation results after 40 days. f) Simulation results after 50 days.

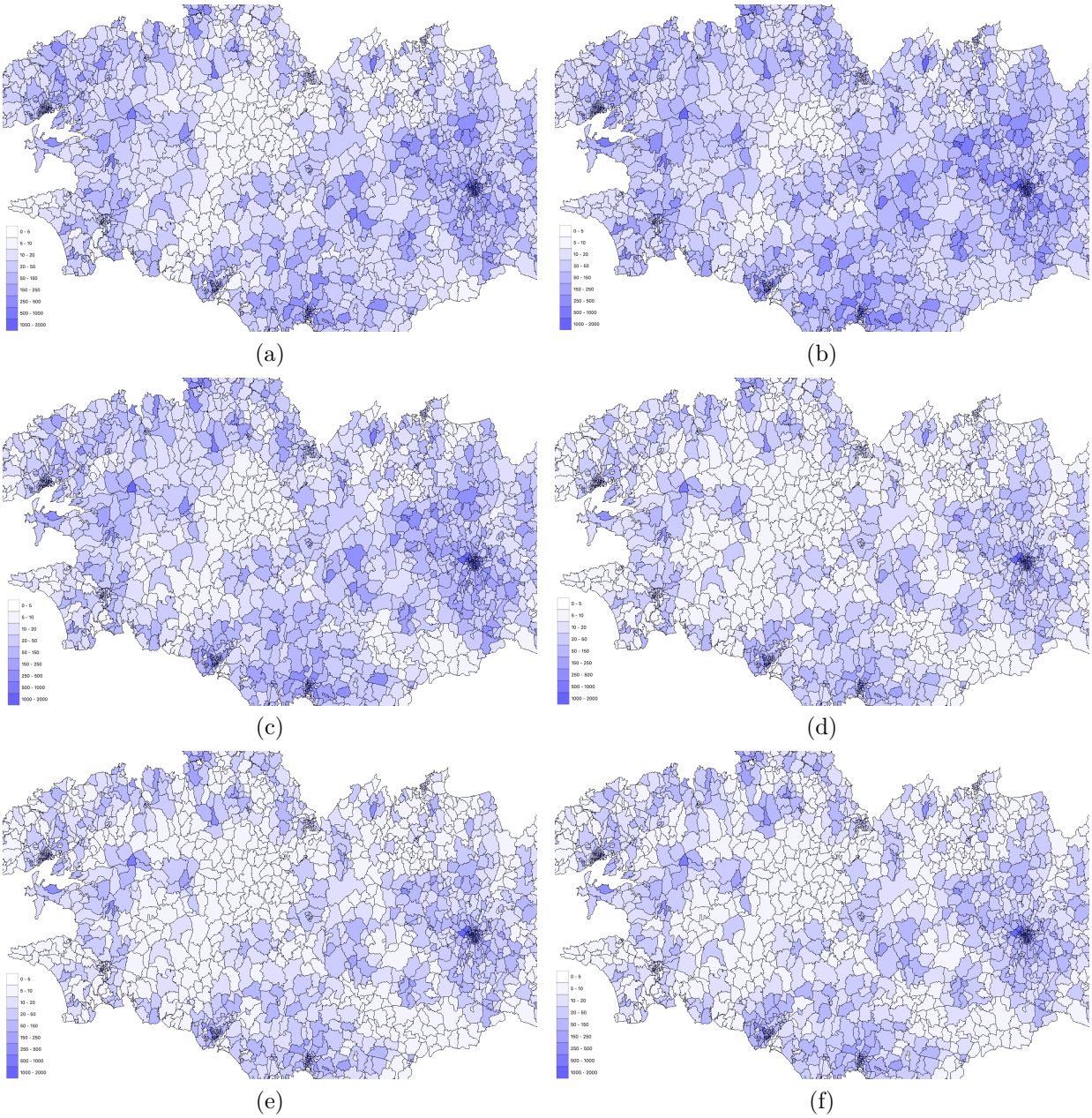
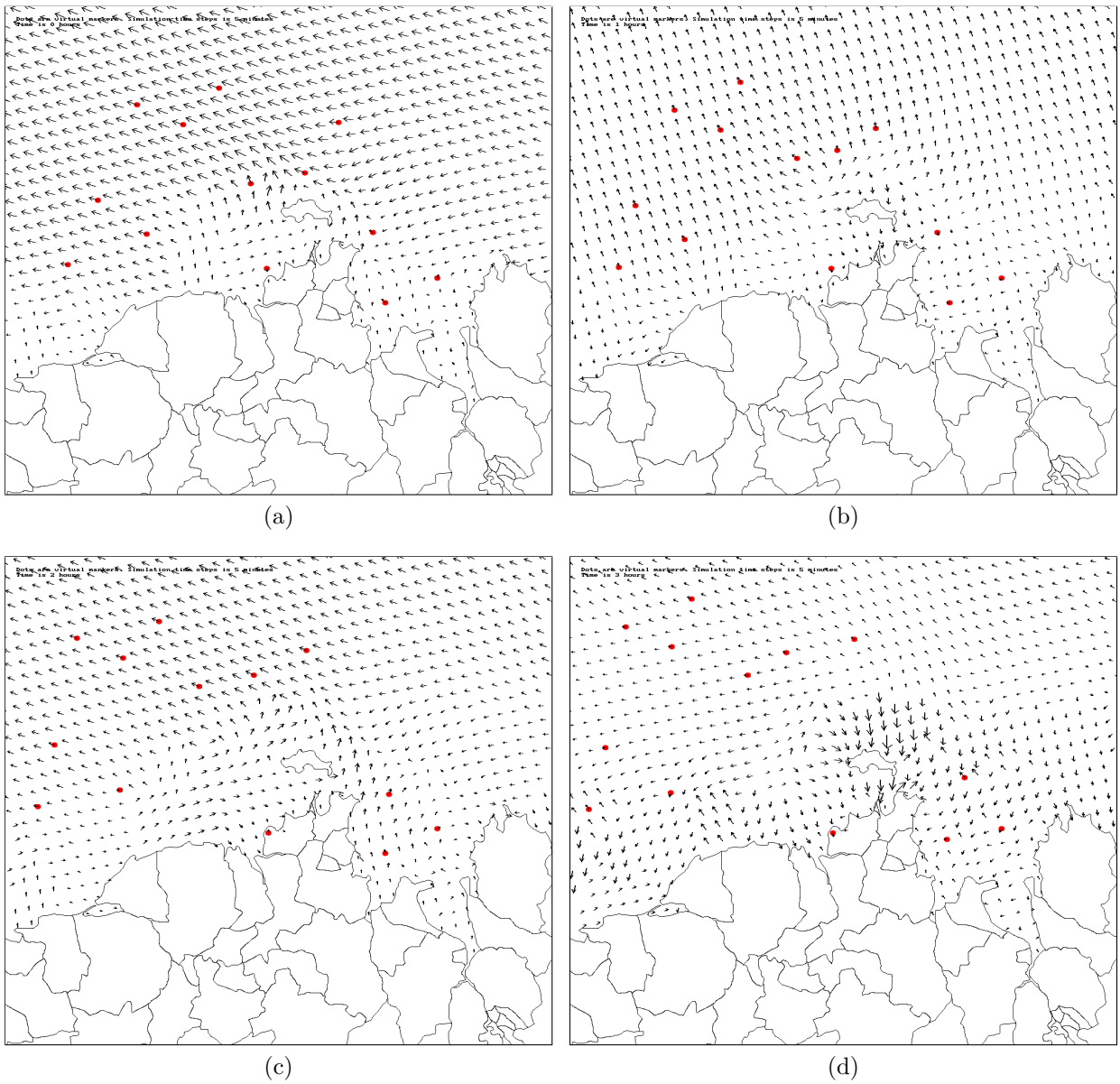


Figure 8 – Covid-19 incidence rate of Brittany, France over time. a) Simulation results after 60 days. b) Simulation results after 70 days. c) Simulation results after 80 days. d) Simulation results after 90 days. e) Simulation results after 100 days. f) Simulation results after 110 days.

4.2 Marker movement with currents

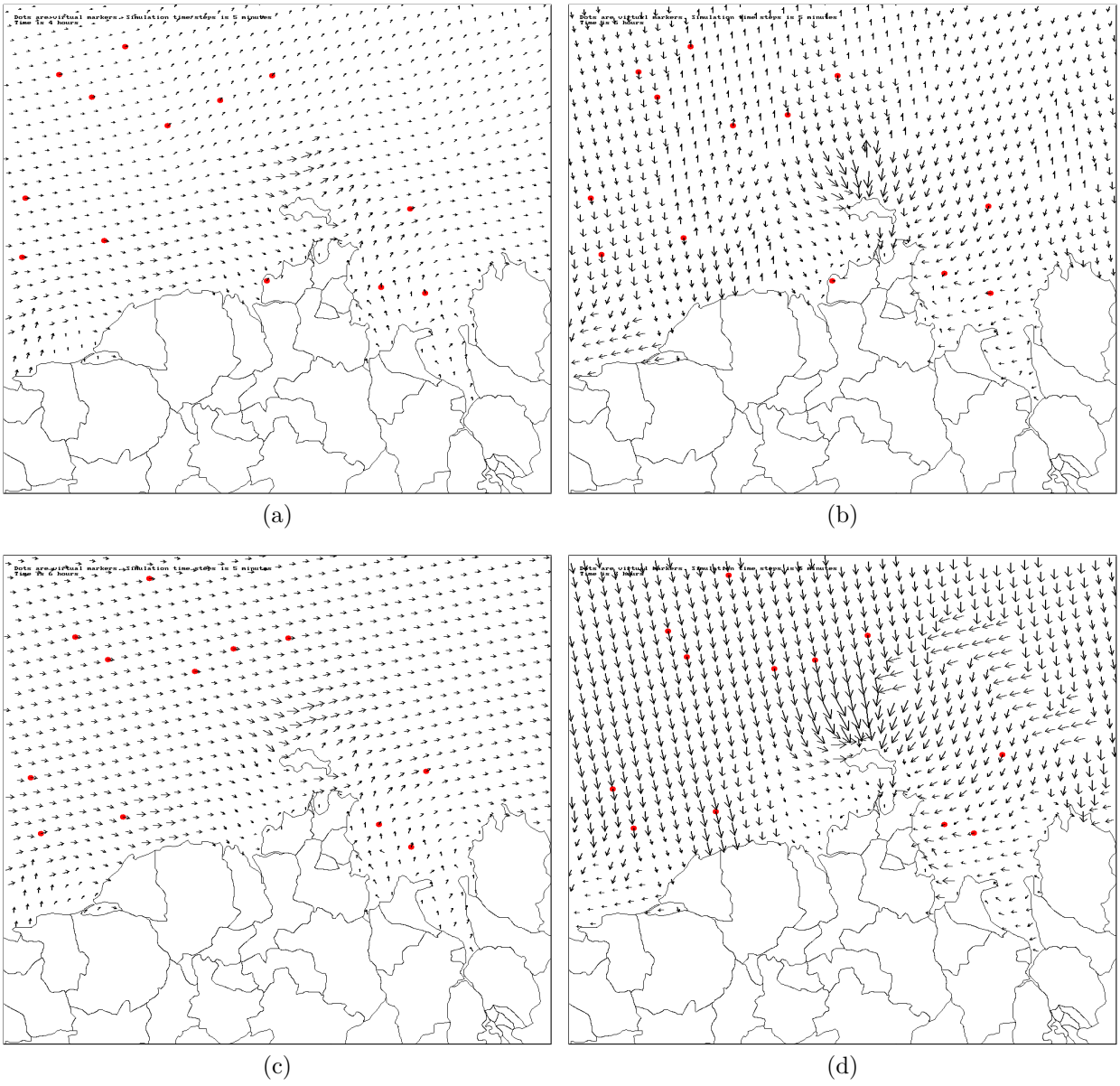
Markers are dropped randomly in sea area and currents cause the movement of markers in the region. The marker movements in first twelfth hours are showed in Figures 9, 10, and 11.



(a) (b)

(c) (d)

Figure 9 – Marker movement by currents. a) 01 June 2022 12AM. b) 01 June 2022 1AM. c) 01 June 2022 2AM. d) 01 June 2022 3AM.



(a) (b)

(c) (d)

Figure 10 – Marker movement by currents. a) 01 June 2022 4AM. b) 01 June 2022 5AM. c) 01 June 2022 6AM. d) 01 June 2022 7AM.

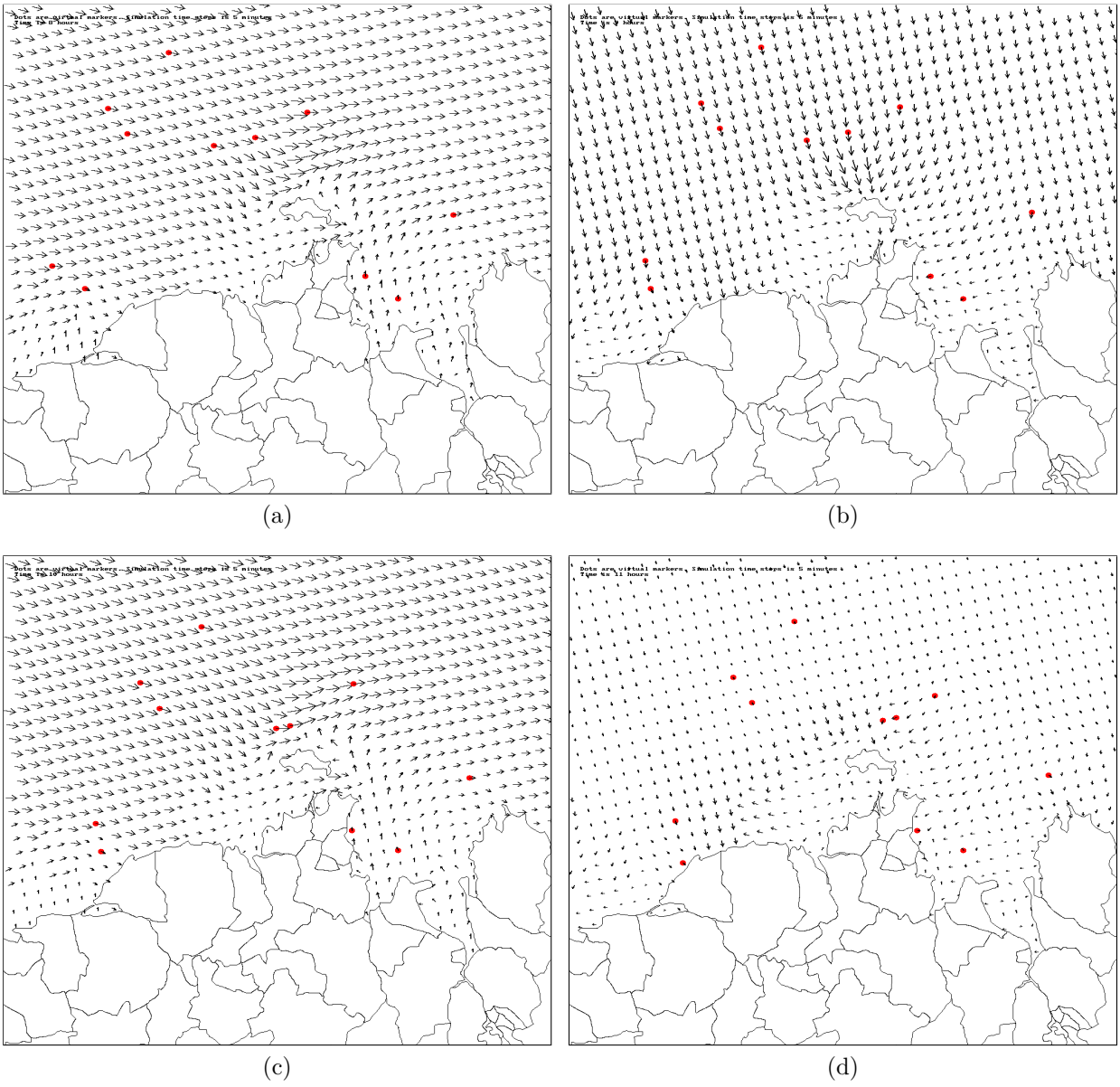


Figure 11 – Marker movement by currents. a) 01 June 2022 8AM. b) 01 June 2022 9AM. c) 01 June 2022 10AM. d) 01 June 2022 11AM.

5 Publications

In the scope of this thesis, we have several publications as follows:

1. **Conference**: Trieu, T. N., Pottier, B., Rodin, V. & Huynh, X. H., *Interpretable Machine Learning for Meteorological Data in ICMLSC 2021 The 5th International Conference on Machine Learning and Soft Computing* https://dx.doi.org/10.1145/3453800.3453803 (Sanya, China, 29-31, January 2021), 11–17
2. **Journal:** Trieu, T. N., Williams, Z., Dorville, J.-F., Huynh, X. H., Rodin, V. & Pottier, B., Open Data for Environment Sensing: Crowdsourcing Geolocation Data, *EAI Endorsed Transactions on Context-aware Systems and Applications (CASA)* **7,** European Union Digital Library (EUDL), https://dx.doi.org/10.4108/ eai.12-5-2020.164496 (May 2020)
3. **Conference:** Trieu, T. N., Huynh, X. H., Pottier, B. & Rodin, V., *Epidemic Spreading Simulation on Distributed Process Systems in CAIT 2023 The 4th International Conference on Computers and Artificial Intelligence Technology* https://doi.org/10.1109/CAIT59945.2023.10469367 (Macau, China, 13-15, December 2023), 76–83
4. **Journal:** Trieu, T. N., Huynh, X. H., Rodin, V. & Pottier, B., Shore Pollution Simulation Based on Tidal Currents and Ground Effects, *International Journal of Environmental Science and Development (IJESD)* **15,** https://dx.doi.org/10.18178/ijesd.2024.15.3.1477, 122–129 (2024)

5.1 Open Data for Environment Sensing: Crowdsourcing Geolocation Data

EAI Endorsed Transactions
on Context-aware Systems and Applications Research Article EAI.EU

Open Data for Environment Sensing: Crowdsourcing Geolocation Data

Ngoan Thanh Trieu[1], Zachary E. S. Williams[2], Jean-François M. Dorville[3], Hiep Xuan Huynh[1,*], Vincent Rodin[4], Bernard Pottier[4]

[1] Can Tho University, Can Tho, Vietnam
[2] University of The West Indies, Kingston, Jamaica
[3] The Caribbean Geophysical and Numerical Research Group, Baie-Mahault, Guadeloupe
[4] Université de Bretagne Occidentale, Brest, France

Abstract

There are numerous situations where the digital representation of the environment appears critical for understanding and decision-making: threats on soils, water, seashores, risk of fires, pollutions are evident applications. If spatial cellular decomposition is evidence in the more common applications, there remains a large field for environment and activities modelling. The integration and composition of several information sources is perhaps the main difficulty with the need to deal with data interpretation and semantics inside concurrent simulators. Besides, the data on population, people's behaviours, people's perceptions are essential in environmental assessments, where the technical aspect is not counted as much as the common acceptance of impact technology. We provide a model for building environmental services with open data systems. A case study is given for getting information from the public about their relationship with freshwater and its scarcity in Jamaica.

Keywords: Open Data, Web Semantic, Environment Sensing, Geolocation Data, And Environmental Simulation.

Received on 29 February 2020, accepted on 09 May 2020, published on 12 May 2020

doi: 10.4108/eai.12-5-2020.164496

*Corresponding author. Email: hxhiep@ctu.edu.vn

1. Introduction

With the Internet development, the connection between people is better supported. The data generated from this connection can be up to the volume of exabytes or even zettabytes. Many challenges arise with this large amount of data such as data storage, processing, and leveraging the value of the data [1]. The data can include geographic information, environmental information, public health, education, statistics, etc. These data are stored under different formats and are kept in separate storage of organizations. There are almost no links between these data that allow the aggregation of different data sources.

Open data is referred to as a solution to this problem. Open data [2, 3] is the data that anyone can use and redistribute. The most recent research [9] points out the usefulness of combining crowdsourcing [26] (a large number of users in data creation) and sensing for a smart city. Data is collected from sensors, bus operating companies, and users to provide complete paths information according to individual's needs. This research provides applications that support people moving in Smart City by equipping them with accessible and personalized paths.

The urbanization process is causing negative impacts on the environment [4], contributing to pollution and harming human health. In this paper, we will give a model for building an open data system of environment sensing. The aim is to have a full path from collecting environmental data, creating open databases, and using the data for environmental simulations to provide warnings for specific issues. We give a case study that has

been done to know how a population can adapt to new freshwater resources related to climate change, natural resource variation, and impact on the marine environment.

The rest of this paper is organized as follows. Section 2 presents the background knowledge related to open data and linked open data. Section 3 shows the environment simulation methodology and our development efforts. Section 4 will be a case study where we have collected data from the public questioning freshwater issues. The conclusion of the paper is presented in section 5.

2. BACKGROUND

2.1 Semantic Web

The Semantic Web [14] is a collaborative movement led by the W3C organization. This is a standard for the development of common data formats on the World Wide Web to archive the goal of making machine-understandable Internet data.

Web Ontology Language

Ontology [15] is a way of describing the concepts and relationships. It is basically to define the knowledge structure for different fields: nouns represent object classes and verbs denote the relationship between objects.

Web Ontology Language (OWL) [16, 17] is a semantic web language that is designed to express rich and complex knowledge about things, groups of things, and the relationships among things. Figure 1 shows an example of OWL structure defining specific objects.

```
<owl:Class rdf:ID="Wine">
    <rdfs:subClassOf
    rdf:resource="&food;PotableLiquid"/>
    <rdfs:label xml:lang="en">wine</rdfs:label>
    <rdfs:label xml:lang="fr">vin</rdfs:label>
    ...
</owl:Class>
<owl:Class rdf:ID="Pasta">
    <rdfs:subClassOf
    rdf:resource="#EdibleThing"/>
    ...
</owl:Class>
```

Figure 1. Example of OWL

Resource Description Framework

Resource Description Framework (RDF) [18] is a general method for describing data by defining relationships between data objects. It is a directed, labelled graph data format for the information on the Web. RDF will split the information into three parts: subject, predicate, and object. The subject is a resource that can be identified by a Uniform Resource Identifier (URI), the predicate is the relationship specification, and the object is a resource or a literal.

An example of an RDF link that connects a DBPedia URI identifying the city of Brest, France (`<http://dbpedia.org/resource/Brest, France>`) with its extra information provided by Geonames server (`<https://www.geonames.org/3030300/>`) is presented in figure 2.

```
<http://dbpedia.org/resource/Brest, France>
owl:sameAs <https://www.geonames.org/3030300/>
```

Figure 2. Example of RDF link

SPARQL

SPARQL [19] is a RDF query language that can retrieve and manipulate data stored in RDF format. The example of a query in SPARQL is shown in figure 3. The result is the names and the email addresses of every person *x* in the database.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE{
    ?x  a  foaf:Person .
    ?x foaf:name ?name .
    ?x foaf:mbox ?mbox .
 }
```

Figure 3. Example of SPARQL

2.2 Open Data

When data can be freely used, reused, and redistributed by anyone, it is open data [1, 2]. The term is introduced when people have problems accessing and using data that is commercially valuable. In fact, data is considered as a new kind of resources, which has its intrinsic value. It is necessary to transform or to refine the data to take full advantages of its internal value.

Open data aims to build a technology platform and technical standards to ensure that individuals and groups of the social community can access and freely use the data without any special restrictions or licenses. An open data system can be conceived as a unified portal, where there is a complete catalogue of all the different open data repositories. The data will be systematically organized and are regularly updated and supplemented. This is an important step in exploiting the value of data by providing a convenient mechanism for users to develop applications based on multiple data sources.

Recently, the term open data has become popular and becoming a trend in developed countries. Many governments are interested in open data since it is an indicator of the United Nations level of e-government development. Many countries have set up dedicated portals for sharing open data. The United States is the first country to publish Government Open Data through the government data portal data.gov. The portal was opened on May 21, 2009, at the initiative of President Barack Obama. Along with providing data, description data is also added to provide more information about each data set such as data content, origin, and update time. After the United States published Government Open Data, the provision of Government Open Data quickly became the

goal of information and data transparency commitments of many countries.

The main principles when considering open data are:

- Accessibility: Data must be available to a wide range of users and a variety of purposes. Protocols and formats of data delivery must be standard.
- Processability: The data provided must be organized so that it is convenient for automatic processing. The usability of the data is influenced by properly encoding the data.
- Globality: People must be able to use data without distinction between groups or domains.

The OpenSense Project [8] aims to provide the most convenient and efficient mechanism for monitoring air pollution. This is an important issue because it directly affects human health, especially in big cities where air pollution is getting worse. This project attaches sensors on public transport systems to collect data everywhere quickly and reduce the cost of installing sensors in multiple places. The large-scale environmental monitoring has posed many challenges for real-time handling of large data.

Open data is often associated with crowdsourcing data production [26], which means the involvement of a large number of users in data creation. With the participation of many users, the tasks will be done quickly and at a lower cost. An example is Wikipedia[1], an international online project for creating a free encyclopedia in multiple languages. Another example similar to Wikipedia is OpenStreetMap[2], the goal is to create a set of map data to freely use and edit. Users can download portions of OpenStreetMap information in vector or raster formats for later processing.

2.3 Linked Open Data

Linked data [5][6] is an important term in the concept of the Semantic Web. It means to create databases that can be understood by human and machine. In other words, this is the creation of a set of design principles for sharing machine-readable linked data on the Web. Machine-readable data [7] can be RDF, XML, and JSON.

Tim Berners-Lee outlines the five-star principles of Linked Data:

- Making data available on the Web
- Making data available as structured data
- Making data in a non-proprietary format
- Use URI to identify things, so that people can point at the data
- Link the data to other data to provide context

The Linking Open Data project developed by the W3C community[3] has put a lot of effort to enrich the linked open data cloud. This project has published various open datasets (such as DBPedia[4], Musicbrainz[5], DBLP[6], and Geonames[7]) as RDF on the Web. By interlinking, the user can navigate between DBPedia data to extra information provided by many different sources. Data is interconnected on a large scale allowing users to get more useful information from external databases when developing applications.

3. Environmental Simulation

3.1 Real-time monitoring

Many changes are appearing in climate, life, and economy balances. Fortunately, scientific activities brought knowledge and methods that give the hope to find solutions to rising problems. Domains such as meteorology, atmosphere studies, oceanography, agriculture, and biology are efficient and sometimes well organized.

It is known that some changes are very difficult to measure and monitor. Biodiversity and density of species are examples of the difficulties rising for measuring wide and sparse phenomena. Mekong Delta is infested by billions of insects that can destroy rice production and water salinity is invading the land putting even more pressure on agriculture. But there is no immediate way to classify and count insects, and for the physical underground water penetration, it is the same.

The core of research-oriented to climate change needs elaborated tools and techniques to collect physical information, to process this information and synthesize scientific facts accurately. Sensing is one part of the problem and deduction of distributed behaviour from local measures is another part. From an understanding of a physical, biological, or social status, it becomes an obvious issue to deduce possible evolutions and the effectiveness of counter-actions.

Previous research efforts associating these aspects can be mentioned for insect monitoring [21], building context-aware communication systems, and simulating physical phenomena [20]. From an understanding of physical, biological or social status, it becomes an obvious issue to deduce possible evolutions and the effectiveness of counter-actions. These efforts are currently improved using highly parallel computations [22] over a wide area and fine resolutions.

3.2 Environmental simulation

[1]https://www.wikipedia.org/

[2]https://www.openstreetmap.org

[3]https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[4]https://wiki.dbpedia.org

[5]https://musicbrainz.org

[6]https://dblp.org

[7]https://www.geonames.org

The methodology is based on a cellular decomposition of geography. Practically, cells will embed information extracted from a database, completed by other geo-localized data coming from different sources. It is currently the case for elevations used to model radio signal propagation or rain flooding simulation. It will be the case for other information coming from sensor fields, satellite image analysis, and feedback information from the public.

Current tools are presented in [20], they address geographic position, sensor network abstraction, and physical representation based on cell systems. The tools enable fast production of high-performance simulators yet ready for concurrent process networks, and graphic processing units, and soon supercomputing with scales of millions of cells and hundred of squared kilometres. The systems are animated using a computing method called "Cellular Automata". We will keep these core functionalities, opening the input data integration, and producing result publications as web services.

The current development efforts[8] include:

- Database storage based on Postgis support, and OpenStreetMap
- Serving tiles for local (Quickmap) and remote browsers (OpenLayers)
- Generation of high-performance concurrent simulators (Multicores, GPUs, MPI)
- Service software architecture for remote end-users (Seaside).
- External data integration in database: meteorological radar map, elevations, sensor fields

The core objective is environment sensing and simulation, in evolving aspects and larger information fields. This includes the support for open data integration, production and publication of predictions coming from simulations, direct interaction with engineers, specialists, and in some place interested publics.

We propose a model for building environmental services with open data (figure 4). The whole process is combining of environmental data collection, open databases creation, and environmental services formation.

4. Case study: Crowdsourcing geolocation data

How, in a real case, useful data can be generated, linked, and made accessible to the worldwide population. This section presents an example of crowdsourcing data collection, a new trend of data collection with help of a large group of people.

4.1 Environmental context and needs of data

Freshwater is a vital element for human beings but since the industrial revolution is becoming one of the most endangered resources [11]. Population growth but also the increase of the need in agriculture and industry in a climate change context-induced more scarcity of the 3% freshwater available at earth's surface. Solutions are available to face these new needs of freshwater particularly during the dry season and drought events. Desalination is based on the usage of the main water resource available at the earth's surface (i.e. salted water) to produce drinkable water [24]. Water molecules are segregated from dissolved salt by thermal, chemical or mechanical methods. Most of the eight main methods of desalination use large amounts of energy to produce freshwater and a highly salted waste. The brine produced can be converted in salt but is more frequently released at the coastline area with negative effects on the environment [10].

Quality and taste of the freshwater produced by the desalination process are not the same as springs, rivers, or well water. A shift from conventional freshwater procurement to desalination cannot be done without a large amount of energy, full access to high seawater quality, and a population ready to change its water usage habits. That change involves data to design the industrial plant, determine its best location, ease water resource management, and evaluate how people will be able to adapt or not.

As part of a research project to design a desalination plant powered by mix renewable energy i.e. wind-solar-

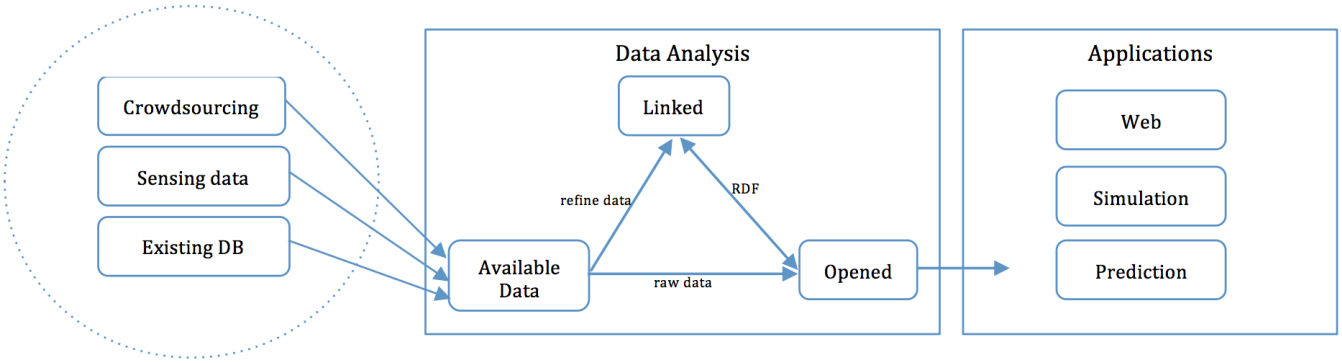


Figure 4. Model for Environmental services with Open

[8]http://sames.univ-brest.fr/sameswp/

wave for the island of Jamaica (\#JamGeenDesal), the needs of appropriate data was highlighted. Official open

data are available from government websites (e.g. NWC [https://www.nwcjamaica.com/Physical_Facili_Ops\#1]) but also from international organizations (e.g. FAO dataset [http://www.fao.org/faostat/en/\#home]). Those data are for the most part not structured and/or not linked. A large part of the data analysis in this project was based on pre-processing to ease the reading and correlation of information.

Analysis of Geographic Information System (then after GIS) of distribution of freshwater resources such as rivers, wells or lakes indicate that the current situation is unsuitable to face climate change previsions and population growth [13]. A large number of freshwater sources gives a wrong indication on the availability of the resource; indeed most of them are non-renewable. In Jamaica, the growth of population (12% in 20 years) and the needs in agriculture and industry in a global warming context imbalance the local water cycle. The needs of new freshwater resources are justified despite its reputation on the island of wood and water.

Data compiled with the renewable energy resource of the last 20 years allow the determination of the best location for a desalination plant but also the potential production of freshwater and volume of waste.

A large amount of data are available on the population as distribution in space, gender, age, etc. no any are related to their behaviours with freshwater, energy, environment, and their point of view on climate change and impacts of the desalination process. Those data are essential in this kind of project where the technical aspect does not count as much as the popular acceptance of impacting technology. Those data must be created or retrieved at the source. A data collection process has been conveyed to the citizens.

4.2 Design of the data collection process

Five main classes of information are needed to define the behaviour of citizens with water, acceptability of the new source of water, and the management of new waste and their impacts on the marine environment (Figure 5). The aim of this process, some can call it a collection of information, is to determine the understanding, level of awareness, and habits of the population.

The structure of the process (accessible TCGNRG website [http://www.tcgnrg.com]) is based on the five

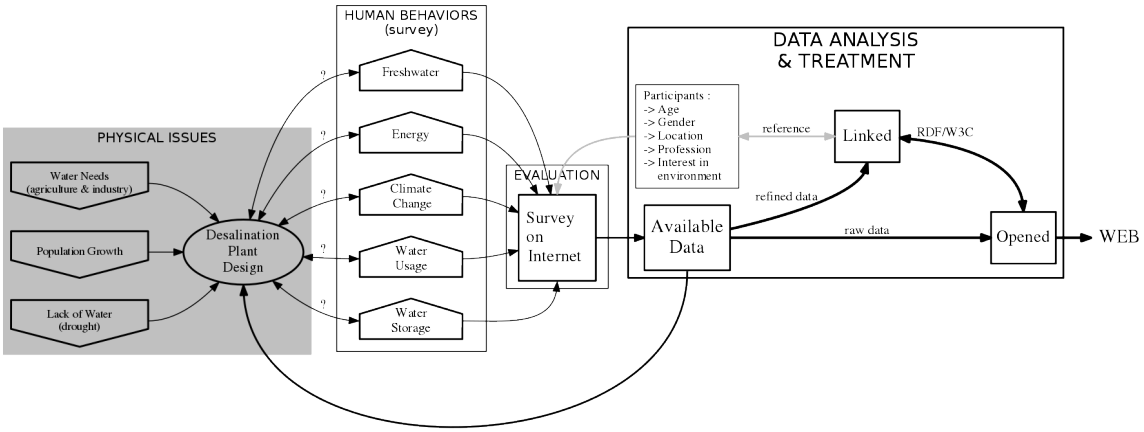


Figure 5. Chart flow of crowdsourcing geolocation data for desalination plant design

classes, which can be listed as Freshwater, Energy, Global Warming, Water consumption, Water storage. Those five classes are related to a desalination plant powered by renewable energy. Where freshwater is the final product which should be used by the population, Energy will be the most costly part (energy used to run a Desktop PC during 12 hours serves to produce 1m3 of freshwater), Global Warming due to climate change is the main key of sustainability of the production with extreme dry condition pushing the development of alternative solutions; Water consumption amount and habits determine the current and future needs and water storage gives some indication of planning of freshwater usage.

All those data should be linked with at least the five parameters allowing a good contextualization of information. In that case, the determination of the best location for a desalination plant is based on the main administrative units of Jamaica: the parishes, the country is divided into fourteen parishes. Nature of the population i.e. age, gender, and profession of the participant, indicates how acceptance of new freshwater resources can be facilitated. Engagement or sensibility of persons to the environment should be taken into account to validate their answers. So, the five classes of questions must be linked to the age, gender, profession, residence, and interest to the environment of the responder (Figure 5).

Questions and answers proposed in this process depend on the method of data harvesting and the capacities of the participant to understand and respond to the questions. The choice was made in this first phase to use individual questioning with simple meaning [12]. Reading and analysis of the answers are facilitated by usage of ranked answer system of three, five, or ten levels as "yes, no, I don't know" or "strongly agree, agree, no opinion, disagree, strongly disagree". That method allows attribution of a value of -1, 0, and 1 for the first case or 0 to 5 for the second. Digits facilitate data manipulation and the generation of an index of concern. To respect the

privacy and anonymity of the participants no email addresses or IP addresses were retrieved or recorded.

4.3 Results

Over the period of June 16, 2019 to September 15, 2019 the crowdsourcing data collection process fully running on the Internet using Google Form was conducted and conveyed to the Jamaican public. A communication campaign to push participants to answer was based on bulk emailing, social network dissemination, and personal network outreach. Only 211 responses were validated.

This poor result can be attributed to a lack of funding to launch a targeted communication campaign but also the length of the survey, the 30 questions asking 2 to 5 minutes to answer. The period of the process, summer holiday could also account for the mediocre rate of response. This small number of participants does not allow the validation of the results but only the retrieving of the trend. The graphical representation of the main outputs is presented in Figure 6. The participants presented more Female (69%) than Male (29%) more aged persons of more than 26-35 years old, which is not representative of the Jamaican population [23]. More persons concerned by the environment have taken the time to answer the questions in the form. The greater concentration of answers comes from Kingston (Capital of the country) where access to the Internet and computer equipment is best.

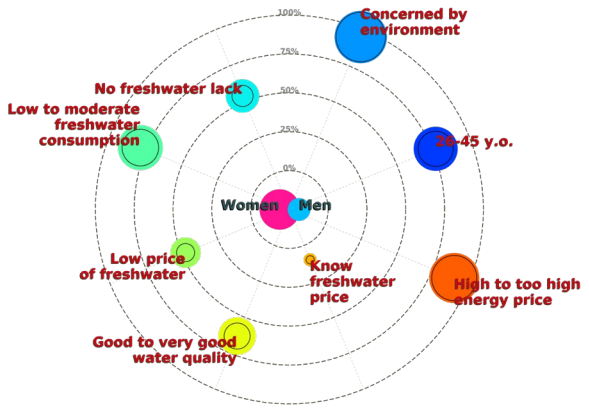


Figure 6. The highlight of the data collection main output linked with participant gender, concentric circles indicate the percentage of women agree with that assertion

By order of concern, participants indicated that climate is first followed by energy and freshwater. Up to 79% of the participants consider that their water consumption is moderate to low, 63% think that water quality is "good" to "very good" but only 10% of the participants gave an acceptable range of freshwater price with 60% who clearly said they do not know. Energy seems more important than water for most of the participants; they have a better understanding of energy price and usage. That point of view can be explained by the large needs of energy and particularly electricity due to modern living although the fact stands that only freshwater is indispensable to life. Participants are more able to give their energy consumption than the volume of freshwater used.

The Jamaican population is aware of climate change impacts but cannot link fossil fuel energy and freshwater distribution. Both aspects of freshwater and energy are not linked together despite the main part of water treatment and distribution (pumping) is based on the usage of fossil fuel.

Water storage class of questions indicates that in a country with frequent water shortage 64% of the householder has permanent water reserves using both non-permanent in bottle and permanent storage systems (e.g. a water tank). They are a bit more concerned about the amount of water stored than the quality.

Data retrieved are analysed in the case of this study. They allowed to obtain an indication of perception and usage of freshwater but can be easily reused in another context.

4.4 Data analysis and data conversion

Analysis of this collection process is based on the digitalization of the answer to retrieve exploitable data (Figure 6). The data can be an integer or a float number. They can represent a statistical value, a count of the item, or a constructed index but they must be digitalized to ease exploitation.

The questions in the Google Form can be on two main forms. The first form is an evaluation of opinion on a subject where the participant says if she/he is agreed or not with a sentence or a concept. In case of evaluation of an opinion a level of agreement is estimate indicating how close the participant is to the opposite hypotheses, **H0** totally agrees with this point of view and **H1** totality disagrees with this point of view [25]. In the middle "I don't know" or "no opinion" means not agree with the two hypotheses.

The second form of the questions is a choice of one or several items in a list or a collection. We ask the participants to select an item, a value, a number, a colour, a word, a sentence, a location, or a country or several of them. The values can be directly used after the computation of the average or determination of the more frequent answer. The other elements selected (i.e. colour, word, etc.) can be manipulated through the percentage of occurrence in the set of answers. But to ease the manipulation they can also be converted to a digit using a dictionary or a mathematical formula, in that case, the value obtain is related to a psychological or perception parameter (Figure 6).

Data retrieved from the process related to environmental issues can be used as a parameter of environmental simulation as described in section 3.

4.5 Generation of open data

The structure of the collection process can push to use a relational database where each question class is stored in a dedicated table or sub-table related by an index. Access to the database can be a limit to the concept of Open Data.

Another organization of the results can be chosen to ease the dissemination of the information, it is base on RDF format [23] through a dedicated XML file or web page using RDF format (Figure 5). This web page or XML file will summarize the results in human-readable format, with information linked to main features of the participants i.e. age, gender, location, profession and interest to environmental issues.

4.6 Perspectives

A second phase of the process will be launched soon with a large target audience with a better selection of the questions and modes of answers. The second phase will include a Geographic Information Systems (GIS) tool to get information with lower space units: at the scale of a city or of a district or even smaller. That small unit size will be close to the cells used for environmental simulations and ease the integration of human behaviour in modelling. It will also take into account time and integrate meteorological seasons in questions/answers. Data analysis will be designed to be used as fully open and linked data.

5. Conclusion

This study reveals the major interests of environmental sensing and simulation in prediction physical issues. A clear model is given showing how we collect environmental information and create open data for building environmental services. Cellular automata with transition rules between cells are the core concept in this work for simulations. Open data is hoped to give the vision of ambitious information systems covering the environment in the neighbourhood.

Acknowledgements.

This work is supported by a Brest Metropole initiative related to Open Data for environment. It is rooted in SAMES project (Ministère des Affaires Etrangeres 2016-2018) grouping mainly researchers from UBO, Can Tho University, ... and now from places in West Indies, Asia, and Africa around software developments for environment modelling and simulation.

References

[1] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. Big data: Issues and challenges moving forward. *46th Hawaii International Conference on System Sciences.* January 2013. IEEE. p. 995-1004.

[2] Evans, J. A., & Reimer, J. Open access and global participation in science. *Science.* 2009; *323*(5917): 1025-1025.

[3] Xu, G. H. Open access to scientific data: promoting science and innovation. *Data Science Journal. 2007; 6*: OD21-OD25.

[4] Uttara, S., Bhuvandas, N., & Aggarwal, V. Impacts of urbanization on environment. *International Journal of Research in Engineering and Applied Sciences*. 2012; *2*(2): 1637-1645.

[5] Bizer, C., Heath, T., & Berners-Lee, T. *Semantic Services, Interoperability and Web Applications: Emerging Concepts.* IGI Global; 2011. Linked data: The story so far; p. [205-227].

[6] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. Linked data on the web (LDOW2008). *Proceedings of the 17th International Conference on World Wide Web.* April 2008. p. 1265-1266.

[7] Barometer, O. D. Open data barometer. *World Wide Web Foundation.* 2015; 1-60.

[8] Aberer, K., Sathe, S., Chakraborty, D., Martinoli, A., Barrenetxea, G., Faltings, B., & Thiele, L. OpenSense: Open community driven sensing of environment. *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming.* November 2010. p. 39-42.

[9] Mirri, S., Prandi, C., Salomoni, P., Callegati, F., & Campi, A. On combining crowdsourcing, sensing and open data for an accessible smart city. *Eighth International Conference on Next Generation Mobile Apps, Services and Technologies.* IEEE; September 2014. p. 294-299.

[10] Jones, E., Qadir, M., van Vliet, M. T., Smakhtin, V., & Kang, S. M. The state of desalination and brine production: A global outlook. *Science of the Total Environment*. 2019; *657*: 1343-1356.

[11] Oki, T., & Kanae, S. Global hydrological cycles and world water resources. *Science*. 2006; *313*(5790): 1068-1072.

[12] Roßmann, J., Gummer, T., & Silber, H. Mitigating satisficing in cognitively demanding grid questions: evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*. 2018; *6*(3): 376-400.

[13] Z. E. S Williams. Renewable Energy for Desalination Process: Efficiency and Environmental Impacts in a Tropical Island Using Digital Tools. Master thesis, Jamaica*: Univ. The West Indies*; 2019.

[14] Berners-Lee, T., Hendler, J., & Lassila, O. The semantic web. *Scientific American*. 2001; *284*(5): 34-43.

[15] Euzenat, J., & Shvaiko, P. *Ontology Matching*. Volume 18. Heidelberg, Berlin: Springer; 2007.

[16] McGuinness, D. L., & Van Harmelen, F. OWL web ontology language overview. *W3C Recommendation*. 2004; *10*(10): 2004.

[17] Antoniou, G., & Van Harmelen, F. *Handbook on Ontologies*. Heidelberg, Berlin: Springer; 2004. Web ontology language: Owl; p. [67-92].

[18] Miller, E. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*. 1998; *25*(1): 15-19.

[19] Schmidt, M., Meier, M., & Lausen, G. Foundations of SPARQL query optimization. *Proceedings of the 13th International Conference on Database Theory.* March 2010. p. 4-33.

[20] Truong, T. P., Pottier, B., & Huynh, H. X. Cellular Simulation for Distributed Sensing over Complex Terrains. *Sensors*. 2018; *18*(7): 2323.

[21] Lam, B. H., Huynh, H. X., & Pottier, B. Synchronous networks for bioenvironmental surveillance based on cellular automata. *EAI Endorsed Transactions on Context-Aware Systems and Applications*. 2016; *3*(8).

[22] Truong, M. T. T., Samar, S. Y., Pottier, B., Rodin, V., & Huynh, X. H. Multiscale Geographic Exploration, Observation, Simulation, and Representation. *13th International Conference-Mathematics, Actuarial, Computer Science & Statistics (MACS 13)*. IEEE; December 2019. p. 1-8.
[23] Coy, C. et al. *The Jamaica Labour Force 2017: annual review.* Kingston, Jamaica: The Statistic Institute Of Jamaica: 2017.
[24] Buros, O. K. *The ABCs of Desalting*. Topsfield, MA: International Desalination Association; 2000.
[25] Shaffer, J. P. Multiple hypothesis testing. *Annual Review of Psychology*. 1995; *46*(1): 561-584.
[26] Howe Jeff. The rise of crowdsourcing. *Wired magazine.* 2006; 14(6): 1-4.

5.2 Interpretable Machine Learning for Meteorological Data

Interpretable Machine Learning for Meteorological Data

Ngoan thanh trieu*
LabSTICC UMR CNRS 6285 Université de Bretagne Occidentale, Brest France and Can Tho University, Can Tho Vietnam

Hiep Xuan Huynh
Can Tho University, Can Tho Vietnam

Vincent Rodin
LabSTICC UMR CNRS 6285 Université de Bretagne Occidentale, Brest France

Bernard Pottier
LabSTICC UMR CNRS 6285 Université de Bretagne Occidentale, Brest France

ABSTRACT

Weather forecasting is the task to predict the state of the atmosphere in a given location. In the past, the weather forecast has been done through physical models of the atmosphere as a fluid. It becomes the problem of solving sophisticated equations of fluid dynamics. In recent years, machine learning algorithms have been used to speed up weather data modeling, a computationally intensive task. Machine learning algorithms learn from data and produce relevant predictions. In addition to prediction, there is a need of providing knowledge about domain relationships inside the data. This paper provides a new approach using interpretable machine learning for explaining the characteristic variables of meteorological data. Interpretable machine learning is the use of machine learning models for the extraction of knowledge in the data. An illustration is shown on characteristic variables of meteorological data.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**;

KEYWORDS

Environment Simulations, Weather Data, BUFR/Express, Interpretable Machine Learning

ACM Reference Format:
Ngoan thanh trieu, Hiep Xuan Huynh, Vincent Rodin, and Bernard Pottier. 2021. Interpretable Machine Learning for Meteorological Data. In *2021 The 5th International Conference on Machine Learning and Soft Computing (ICMLSC'21), January 29–31, 2021, Virtual Event, Vietnam.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3453800.3453803

1 INTRODUCTION

Climate change causes serious impacts on human life in an increasingly severe way. The last decade has witnessed a large number of rising problems, such as floods, droughts, and rising sea levels. Human is facing with the environmental challenges that need scientific activities to bring knowledge for adaption with natural environment changes. Simulation and prediction on weather data can help to explain and predict the occurring problems and deduce possible solutions. Traditionally, the weather forecast has been done through physical simulations and sophisticated mathematics. The inaccuracy of forecasting is due to the dynamic nature of the atmosphere and incomplete understanding of complex atmospheric processes. In recent years, machine learning algorithms have been widely used for intelligent weather prediction since it is not necessary to have a complete understanding of complex processes that govern the atmosphere. In [15] [12], researchers used neural networks for weather forecasting because this popular machine learning model has a capacity of capturing non-linear dependencies of future conditions and past weather trends. Neural network techniques were found to be more suitable for non-linear problems compared to traditional techniques. However, there were problems of local minima and model over-fitting in systems using neural networks. The study [22] used support vector machine (SVM) to apply classifiers directly for weather prediction. The study has shown that the performance of support vector machine is better than multi-layer perceptron trained with back propagation algorithm. The study [10] used linear regression and a variation of functional regression to predict the temperatures based on historical patterns. The experiments proved that linear regression is low bias and high variance whereas functional regression is high bias and low variance. Recently, the study [16] applied convolutional neural networks [13], which has been proved to be effective on image classification, to automatically generate local weather forecasts.

There are available problems about the interpretability of machine learning models for weather prediction. Accuracy is no longer enough since predictions need to be explainable for deducing further solutions. Another problem is the chaotic nature of the atmosphere with many characteristic variables, such as precipitation, temperature, and humidity. It is necessary to know the dependencies between the attributes. In other words, there is a need to know which attribute is more important than others that can be used to represent for the prediction.

In this paper, we provide a new approach to using interpretable machine learning for meteorological data. The whole interpretability process of weather data is shown in Figure 1. The black box machine learning model is analyzed with interpretable machine learning on meteorological data to see which feature is important and how does it affect a weather prediction. This research also

*ttngoan@cit.ctu.edu.vn

ICMLSC'21, January 29–31, 2021, Virtual Event, Vietnam

ACM ISBN 978-1-4503-8761-3/21/01…$15.00
https://doi.org/10.1145/3453800.3453803

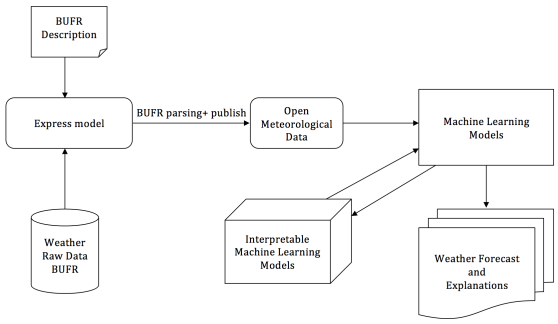


Figure 1: Interpretability process of meteorological data

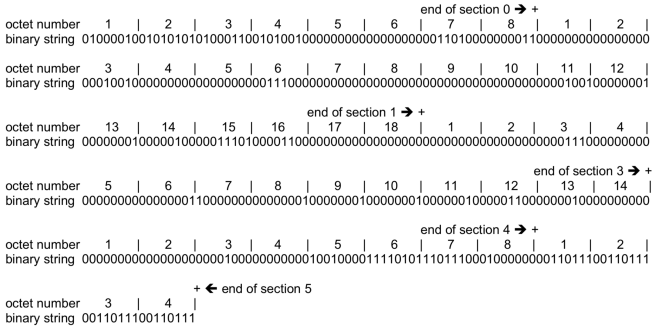


Figure 2: Example of a BUFR message [28]

defines a data model for meteorological data to transfer weather data from BUFR standard to a machine-readable format (such as CSV, JSON, and XML). BUFR [28] [1] is the common format that is used by almost all meteorological services in the world. Thus people can exploit the existing weather data sources for simulation and prediction toward finding solutions to the rising problems.

The structure of this paper is organized as follows. Section 2 provides a data model for meteorological data with BUFR standard and Express modeling language. Section 3 presents an approach on interpretability process of weather prediction. Section 4 will be the experiments processing weather data for prediction and explanation purposes. The conclusion of the paper is presented in section 5.

2 METEOROLOGICAL DATA

Weather data is particularly important, as everyone from individuals to huge companies can benefit from it. Applications that make use of open weather data can therefore potentially have a huge impact. The weather data collected every day can have a range of benefits if it is made open and interpreted in the right way. The crucial measurements taken every day can be used in different ways by people from all walks of life if the right tools are provided to them. In this way, simple information can create new benefits for everyone.

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,1} & \cdots & a_{m,n} \end{pmatrix}$$

Weather data include any facts or numbers about the state of the atmosphere, including temperature, humidity, precipitation, wind speed, and pressure. Weather data can be represented as matrix $A_{m,n}$, in which m is the number of observations recorded and n is the number of features (temperature, humidity, etc.).

Each observation is the information recorded about the state of the weather and the place that records the information. In other words, each line of a weather data file after the header row represents one observation of the weather by one weather station at a specific time. Each column represents a feature of the atmosphere.

In the last few years of twenty century, weather radar became a highly important tool for meteorology, especially with regards to short term forecasting. Weather radars were quite expensive tools thus people try to obtain good coverage of the area of interest at minimum costs, which means having as few radars as possible. It raised the problems of transmission for sharing radar data. The BUFR was the result of expert meetings and periods of experimental usage by several meteorological data processing centers. Lately, it is used as a standard for radar data representation of sharing meteorological data.

2.1 BUFR standard

BUFR (Binary Universal Form for the Representation of meteorological data) is the WMO code form that is designed to represent a continuous binary stream of any meteorological data. The core concept of BUFR standard is self-descriptive nature, which helps this standard in accommodating changes. It only needs to have additional data description tables when there is new observation data. The data description is a major part of the BUFR standard documentation.

A BUFR message contains any kind of observational data and a complete data description of the data. The term "message" means that the BUFR standard is used as a data transmission format. A BUFR message is a continuous binary stream comprising of six sections (Example as in Figure 2). Each section is made up of a series of octets, coined to qualify one byte as an 8-bit sequence. In theory, there is no upper limit to the size of a BUFR message. However, by convention, BUFR messages are restricted to around 15,000 octets. The BUFR is a strong and complex binary format with self-descriptive nature. Thus there is a need of a well-designed program for parsing the descriptors, matching the descriptors with the bit stream, and extracting the values out of the stream.

2.2 Express modeling language

Express is a standard data modeling language, which is formalized in the ISO standard 10303-11 [11] (within the STEP - Standard for The Exchange of Product model data). It is an object-flavored lexical language that is firstly designed to represent the models of industrial products. The data modeling language helps define data objects and the relationships between the objects and enabling the exchange of data between the objects. A data model can be defined in two ways, textual or graphical. In a textual form, a SCHEMA is clarified in which various data types and the structural constraints

and algorithmic rules can be defined. The Express-G is the graphical representation for all details formulated in the textual form. The advantage of using Express-G is that the information can be presented more understandably.

Express provides a series of data types for building blocks in a schema. The most important data type in Express is the entity data type. Entity attributes can relate an entity with other entities. In brief, Express can be used to model data and data relationships with a general inheritance mechanism and can be used as a procedural programming language to specify constraints on data instances.

2.3 BUFR/Express meteorological data modeling

A data modeling process is necessary to define and analyze the BUFR data. It not only defines the data elements, but it also defines the structures and relationships between the elements. This study uses Express as a test-case for meteorological data modeling. Express modeling language will help to analyze the meteorological data in BUFR messages and provide an accurate BUFR parser. A data model is defined for BUFR messages using Express modeling language. As an example, the descriptor entity in our model is represented as follows:

```
ENTITY descriptor;
        fxy : STRING;
        f,x,y : INTEGER;
        element\_name : STRING;
        unit : STRING;
        reference\_number : STRING;
        bitwidth : INTEGER;
END ENTITY;
```

The textual representation of this schema is important as an input for SDAI (Standard data access interface). SDAI is an abstract specification on how to deal with Express schema and can be mapped to various programming languages. This paper uses JSDAI [17], which is a Java application-programming interface for reading, compiling, and writing object-oriented data defined by an Express model.

The java code with classes of objects is auto-generated with JSDAI API. It is further added some functions for reading raw binary data and export the result into commonly used and machine readable-format - CSV.

3 PREDICTION AND EXPLANATION

3.1 Random forests

Breiman proposed random forests [2], a classifier consisting of a collection of tree-structured classifiers $\{\mathbf{h}(\mathbf{x}, \Theta_k), \mathbf{k} = \mathbf{1}, \ldots\}$ where the $\{\Theta_\mathbf{k}\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x*. The notation $\Theta_\mathbf{k}$ was introduced that represented all random choices made when fitting the $\mathbf{K^{th}}$ tree. Decision tree is the basic building blocks of random forests. A heuristic approach to build a decision tree is to choose attribute that will produce a tree with "purest" nodes. There are two common methods for measuring the impurity degree, i.e., Information gain and Gini index (Gini impurity).

Information gain is based on the concept of Entropy [23] in Information theory. It measures how well a given attribute separates the training examples according to their target classification. The best splitting attribute is the one that provides the highest information gain.

$$E(S) = \sum_{i=1}^{n} -p_i * \log_2 p_i \tag{1}$$

$$Gain(S) = E(S) - \sum P(S|A) * E(S|A) \tag{2}$$

where: **E(S)**: Entropy of collection S; $\mathbf{p_i}$: proportion of class **i** that add up to 1; **Gain(S, A)**: Information gain of collection **S; P(S |A)**: proportion of **S** given value **A**; **E(S|A)**: Entropy of collection **S** given value **A**

Gini index [4] is used by CART (Classification and regression tree) algorithm. This method is a measurement of the probability of classifying a data point incorrect. Given the probabilities for each class pi , Gini impurity of collection S is calculated as in equation 3.

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2 \tag{3}$$

3.2 Interpretable machine learning

Machine learning algorithms are mostly considered as a black box, which provides the prediction without explanations. There is a trade-off between accuracy and interpretability. In some applications, it is necessary to have interpretable machine learning models [20] [29] to gain insight into the data. There are several techniques for making machine learning models interpretative [6] [5], such as *global* and *local* techniques. In principle, global interpretability enables users to understand the entire model by its structures and parameters. In contrast, local interpretability examines the reasons for a specific decision is made.

Feature importance [7] is a global interpretation method, which is shuffling the values of the features and measures the drop of performances. A feature is considered to be important as if permuting its values will increase the model error. In other words, it is the amount of information gained by the prediction made on the values of this feature.

Feature interaction [9] is another global interpretation method that can measure how strongly features interact with each other. H-statistic [8] is one way to measure the interaction strength by measuring how much the variation of the predicted outcome depends on the interaction of the features.

Shapley value [25] [26] is a local interpretation method from coalition game theory. This method assumes that each feature is a player in a game and the prediction is the payout. For each feature x , it will evaluate the model of every combination of the features with and without x. The Shapley value of feature x is calculated as equation 4 [25].

$$\varphi_x(v) = \sum_{S \subseteq N\{x\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(v(S \cup \{x\}) - v(S)\right) \tag{4}$$

where: **n** is the total number of features; **S** is a subset of the features; ***v(S)*** is the prediction for feature values in set **S**

This computation is very expensive since the number of coalitions exponentially increases with the number of features. In [27], Strumbelj and Kononenko presented an approximation of the Shapley value algorithm using Monte-Carlo [24] sampling technique.

Lundberg and Lee presented SHAP [18], an efficient estimation Shapley value approach.

4 EXPERIMENTS

4.1 Data description

The experiments use data collected from Meteomanz.com [19], an FTP server of the National Oceanic and Atmospheric Administration (NOAA). This database consists of 13.000 weather stations all over the world specifying by geographical locations. It is noted that the meteorological stations provided are all registered with WMO and each will receive a 5-digit WMO index for identification. The weather data in this database is from the year 2000 until the present.

The data used are hourly collected in a year (from June 2019 to May 2020) in three different characteristic regions. The first dataset is retrieved from Rachgia Vietnam (station number 48907), which belongs to a tropical region. The second dataset is from Brest France (station number 07110), which is inside the temperate zone. The last one is data from Kemi Finland (station number 02864), which has a continental climate with freezing winters and mild summers. Each record is a BUFR message with a list of descriptors and the data section corresponding with the descriptors. The BUFR decoder will generate the CSV files corresponding to the datasets.

4.2 Preprocessing

The CSV file after decoding BUFR data is as follows:

- Rachgia, Vietnam: 1.271 records of 121 columns. There is a lack of data recorded in this city, especially with no data for the whole month in October 2019.
- Brest, France: 7.813 records of 191 columns. The weather data in this city is more complex that has different data descriptors from one day to another day.
- Kemi, Finland: 8530 records of 109 columns. Data recorded in this city is simple with fewer descriptors.

Data cleaning is a critically important step for any machine learning algorithm. In the dataset, there are columns without data (indicate as missing value). Columns that contain a single value or columns with "missing" indication are referred to as zero-variance predictors [14]. It means that measuring the variance of these predictors will return a zero. These columns do not contain any information for modeling. It is simply to remove the zero-variance predictors from the dataset.

4.3 Implementation tools

IML [21] is a R package that provides methods making machine learning models interpretable. Most methods in the package have been implemented in other packages, such as Feature importance [7], Interaction effects [9], and Shapley value [27]. However, this IML package puts all methods in one place thus it is more convenient with the same syntax and consistent functionality.

We use random forest approach, a combination of a large number of decision trees, for interpreting machine learning models. A randomForest [3] is a R package that provides methods to create models. The most common outcome of the decision trees is used as the final output.

4.4 Results

4.4.1 Global interpretation. Measuring how the importance of each feature for the predictions will help us to understand how the models make predictions based on the features and their influence ton the underlying model structure. As an example, we measure the importance of features for predicting Horizontal visibility (Figure 3). In the three places, Relative humidity is an important feature affecting the predictions. The Cloud amount and Height of the cloud are also important features for estimating the Horizontal visibility.

The interactions between the features (Figure 4), how strongly features interact with each other, can be measured. It is the change in the prediction on Horizontal visibility that occurs by varying other features.

4.4.2 Local interpretation. Shapley value is used to explain a specific prediction, which is the contribution of each feature to the difference between actual prediction and average prediction. In Figure 5a, the actual prediction value is 40.47, which is 3.18 below the average prediction of 43.65. The air temperature has the most positive contribution and the cloud amount has the most negative contribution. The sum of Shapley values yields the difference between actual and average prediction. Figure 5b is the scatter plot for feature importance scores with Shapley values of 100 random predictions on Horizontal Visibility in Kemi, Finland. In the whole feature space, it is clear that the Relative Humidity is a confusing feature since its values have the most positive and negative contributions to the prediction outcomes.

Figure 6 is the scatter plot for the Shapley values of 100 random predictions on Horizontal visibility in Brest, France. The Relative humidity and Height of base of cloud are the two confusing features with the positive and negative contributions to the predictions.

5 CONCLUSION

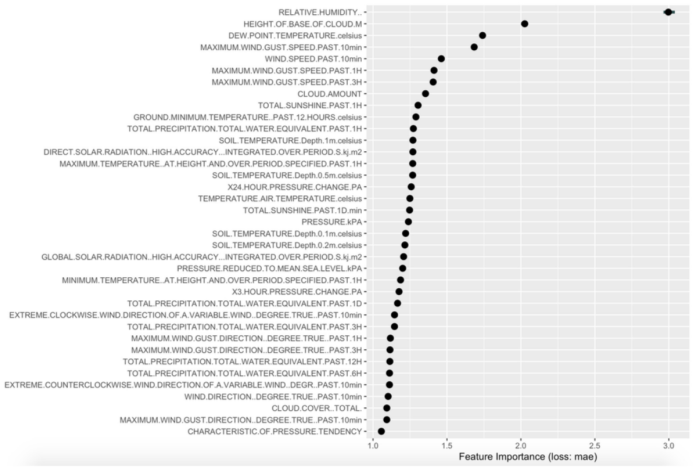
Many meteorological stations are provided for collecting and sharing weather data. It provides a good basis for weather simulations and disaster forecasting based on multiple data sources. This study provides a data model for weather data with BUFR/Express and interpreting meteorological data with interpretable machine learning models. The experiments have been conducted with our proposed model. These results can be used as a scientific basis for improving meteorological data understanding and deciding weather feature weights for the more accurate predictions. The ultimate goal of all researches is to find which counteractions need to be done for the best effects.

ACKNOWLEDGMENTS

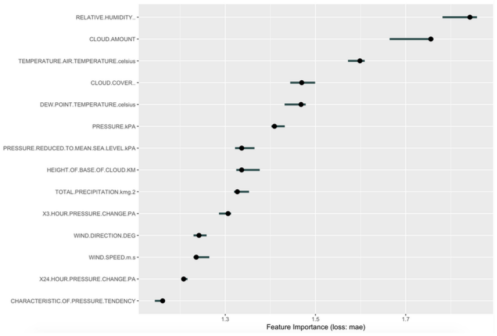
This work is supported by a Brest Metropole initiative related to Open Data for environment simulation. It is rooted in SAMES project (Ministère des Affaires Etrangeres 2016- 2018) grouping mainly researchers from University de Bretagne Occidentale, Can Tho University, and now from places in West Indies, Asia, and Africa around software developments for environment modeling and simulation.

REFERENCES

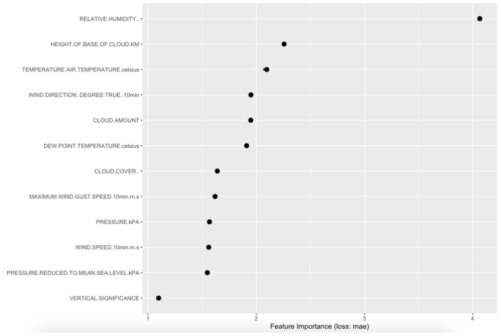
[1] Jean Claude Bergès. 2002. Support of WMO binary format (BUFR and GRIB). In Proceedings of the Open source GIS-GRASS user conference, Trento, Italy,



(a) Brest, France



(b) Rachgia, Vietnam



(c) Kemi, Finland

Figure 3: Feature Importance for predicting Horizontal Visibility in different places

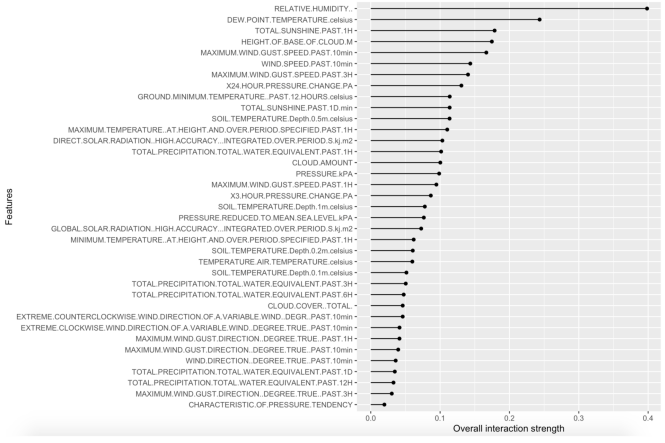


Figure 4: Interactions between features with Horizontal Visibility, Brest, France

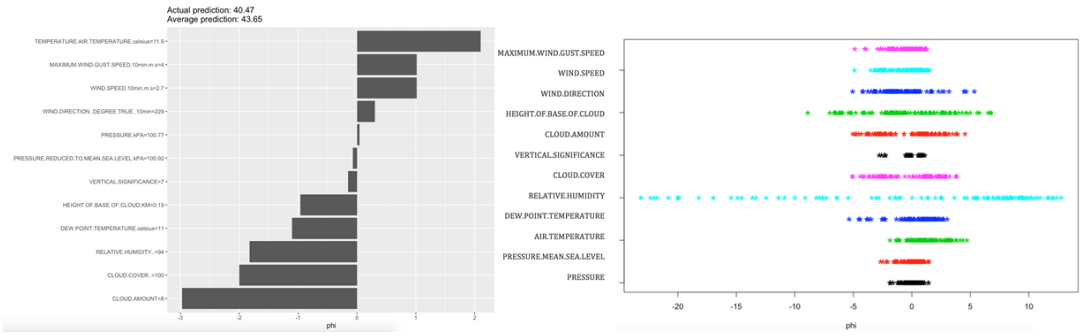


Figure 5: Shapley values of features on predictions in Kemi, Finland

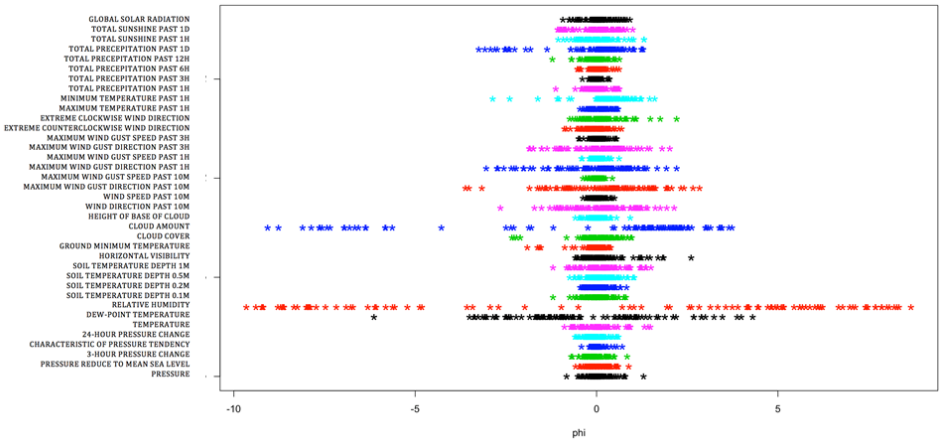


Figure 6: Shapley values of 100 random predictions Brest France

Universitadègli studi di Trento (Ed.). Universitadègli studi di Trento, 11–13.
[2] Leo Breiman. 2001. Random Forests. Mach. Learn. 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324
[3] Leo Breiman *et al.* 2015. randomForest: Breiman and Cutler's random forests for classification and regression. R package version 4 (2015), 6–12.
[4] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. Classification and regression trees. CRC press.
[5] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for Interpretable Machine Learning. Commun. ACM 63, 1 (2019), 68–77.
[6] Radwa Elshawi, Mouaz H Al-Mallah, and Sherif Sakr. 2019. On the interpretability of machine learning-based model for predicting hypertension. BMC medical informatics and decision making 19, 1 (2019), 146.
[7] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. arXiv preprint arXiv:1801.01489 68 (2018).
[8] Jerome H Friedman and Bogdan E Popescu. 2008. Predictive learning via rule ensembles. The Annals of Applied Statistics 2, 3 (2008), 916–954.
[9] Jerome H Friedman, Bogdan E Popescu, *et al.* 2008. Predictive learning via rule ensembles. The Annals of Applied Statistics 2, 3 (2008), 916–954.
[10] Mark Holmstrom, Dylan Liu, and Christopher Vo. 2016.Machine learning applied to weather forecasting. Stanford University (2016), 2–4.
[11] ISO 10303-11:1994(E) 1994. Industrial automation systems and integration—Product data representation and exchange—Part 11: Description methods: The EXPRESS language reference manual. Standard. International Organization for Standardization.
[12] Vladimir M Krasnopolsky and Michael S Fox-Rabinovitz. 2005. Complex hybrid models combining deterministic and machine learning components as a new synergetic paradigm in numerical climate modeling and weather prediction. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Vol. 3. IEEE, 1615–1620.
[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.
[14] Max Kuhn and Kjell Johnson. 2019. Feature engineering and selection: A practical approach for predictive models. CRC Press.
[15] Loi Lei Lai, H Braun, QP Zhang, Q Wu, YN Ma, WC Sun, and L Yang. 2004. Intelligent weather forecast. In Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826), Vol. 7. IEEE, 4216–4221.
[16] Pablo Rozas Larraondo, Inaki Inza, and Jose A Lozano. 2017. Automating weather forecasts based on convolutional networks. In Proceedings of the ICML Workshop on Deep Structured Prediction, PMLR, Vol. 70. JMLR.org.
[17] LKSoft. [n.d.]. JSDAI. https://www.jsdai.net/
[18] Scott Lundberg and Su-In Lee. 2016. An unexpected unity among methods for interpreting model predictions. CoRRabs/1611.07478 (2016). arXiv:1611.07478 http://arxiv.org/abs/1611.07478
[19] Meteomanz.com. [n.d.]. Datos meteorológicos de SYNOPS/BUFR - Predicciones GFS/ECMWF - Meteomanz.com. http://www.meteomanz.com/
[20] Christoph Molnar. 2020. Interpretable Machine Learning. Lulu.com.
[21] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2018. iml: An R package for interpretable machine learning. Journal of Open Source Software 3, 26 (2018), 786.
[22] Y. Radhika and M. Shashi. 2009. Atmospheric Temperature Prediction using Support Vector Machines. International Journal of Computer Theory and Engineering (2009), 55–58.
[23] Claude E Shannon. 1948. A mathematical theory of communication. The Bell system technical journal 27, 3 (1948), 379–423.

[24] Alexander Shapiro. 2003. Monte Carlo sampling methods. Handbooks in operations research and management science 10 (2003), 353–425.
[25] Lloyd S Shapley. 1953. A value for n-person games. Contributions to the Theory of Games 2, 28 (1953), 307–317.
[26] Lloyd S Shapley, Alvin E Roth, *et al.* 1988. The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press.
[27] Erik Strumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems 41, 3 (2014), 647–665.
[28] Wayne Thorpe. [n.d.]. A guide to the WMO code form FM 94 BUFR. Technical Report. Office of the Federal Coordinator for Meteorological Services and Supporting Research, NOAA, Silver Springs. MD,122pp.
[29] Alfredo Vellido,JoseĎavid Martiń-Guerrero,and Paulo JG Lisboa.2012.Making Machine Learning Models Interpretable. In EAANN, Vol. 12. Citeseer, 163–172.

5.3 Epidemic Spreading Simulation on Distributed Process Systems

Epidemic Spreading Simulation on Distributed Process Systems

Ngoan Thanh Trieu
LabSTICC, UMR CNRS 6285, University of Brest, Brest, France
CICT, Can Tho University, Can Tho 902070, Vietnam
ttngoan@cit.ctu.edu.vn

Hiep Xuan Huynh
College of Information and Communication Technology, Can Tho University
Can Tho 902070, Vietnam
hxhiep@ctu.edu.vn

Bernard Pottier
LabSTICC, UMR CNRS 6285 University of Brest
Brest 29200, France
bernard.pottier@univ-brest.fr

Vincent Rodin*
LabSTICC, UMR CNRS 6285 University of Brest
Brest 29200, France
vincent.rodin@univ-brest.fr
*Corresponding author

***Abstract*—Epidemic spreading is still an attractive topic for public attention because of the regular occurrence of pandemics throughout history. Authorities collect health statistics based on geographic divisions and make it open for everyone to create potentially impacted tools. In this study, we give an approach of using distributed process systems for epidemic spreading simulation with the segmentation of geographic divisions. A cellular automata model is defined on this irregular cell space with the initial conditions acquired from Open Data repositories. The spreading process is performed by local exchanges in an adjacent neighborhood. Open data repositories offer multiple parameters that can be used to approximate local behaviors inside automata specifications. Experiments are conducted on Covid-19 spreading simulation and data in a one-year period is analyzed to deduce the transition function. Experiment results show that the epidemic propagation trend is caught although the simulated incidence rates are generally lower than the real incidence rates collected. The practical interest is to understand epidemic spreading in time and space. Principles can be reproduced in a number of situations provided that accurate geographic segmentation and related data are available.**

Keywords—Epidemic spreading simulation, Epidemic control, Irregular cell space, Distributed simulation systems.

I. INTRODUCTION

Control of contagion is a problem that regularly occurs throughout history. Cholera pandemics [1] occurred in the 19th and 20th centuries with a hypothesis that the epidemic transmission was drinking water. In the London outbreak, a pump was identified to supply contaminated water and the cholera was spread from this source to the wider surrounding area. The Great Plague of Marseille arrived in France in 1720 [2]. From the original source of the infection, the epidemic spread to the surrounding localities. Attempts to stop the spread of plague tried to separate Marseille and the rest of Provence with a plague wall and the remains of the wall can still be seen today. Covid-19 is a respiratory syndrome caused by coronaviruses [3] that can spread between people in close contact due to small liquid particles. The epidemic spread from one source to other communities and places far from the source due to the flights of many inhabitants. Before the availability of vaccines, some non-pharmaceutical interventions such as quarantines, school closures, and banning public gatherings were used to delay and flatten pandemic peaks.

There have been some studies that model the epidemic spreading, especially the effects of spatial dimension for propagation. The Covid-19 spreading in Italy has been studied in [4]. The epidemic model was an extended version of the Susceptible-Exposed-Infected-Recovered (SEIR) model including mobility data from 107 provinces in Italy. The spatial nature of the model was helpful in terms of making different mobility restriction policies for the authorities. The combination of the Susceptible-Infected-Recovered (SIR) model and Cellular Automata (CA) was used to examine the spatial and temporal propagation of epidemics [5]. Doran and Laffan [6] used the SIR-CA model to simulate the spatial dynamics of the foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia. The cell system was a square lattice system implemented using the map algebra system. The probability of interactions between the grid cells depended on the density of susceptible herds. The research showed that depending on the season the outbreak is initiated (wet or dry), the evolution results were completely different in the two regions assessed.

This work provides an approach using geographic divisions as an irregular cell space for epidemic spreading simulation in distributed process systems. Open Data collected from geographic divisions are analyzed with machine learning techniques to approximate local behaviors inside automata specifications. As many as 15 parameters are referenced by related databases, it is necessary to be correlated inside territories such as regions or districts. Permutation feature importance is used to select parameters to define local transition rules. The study region is in Brittany, France but the method can be applied in other territories taking into account the local data.

The rest of the paper is organized as follows. Section 2 shows the irregular spatial segmentation with geographic divisions and the definitions of distributed cell systems. Section 3 provides a data analysis process to form the transition function. The experiments and simulation results are presented in Section 4. The perspective and conclusion are presented in Section 5.

II. EPIDEMIC SPREADING WITH IRREGULAR SPATIAL SEGMENTATION

A. Epidemic Counting Policy in Administrative Levels

Health authorities have established a counting policy based on administrative data collection with an antigen test or a real-time polymerase chain reaction test. Counting is done in

This work is funded by a Brest Métropole, France and Can Tho University, Vietnam.

medical entities, following criteria such as age, vaccines, and medical history of patients. A country is usually divided into smaller geographic entities (such as departments, districts, and communes) for management and data collection related to the population. In France, large communes are divided into several IRIS units. IRIS[1] is defined by the National Institute of Statistics and Economic Studies (INSEE) to prepare for the dissemination of the population census. The study of pandemics is based on real data collected from the fragmentation of territories for spatial and temporal simulation. These divisions are bound to a set of data including demographic data, social measurement, meteorological data, and spatial dimension.

Epidemic evolution is measured as the number of infected cases, and the influence is its propagation, according to the transmission rate and the infection rate. The key point is to represent physical facts as program data, and then to process these data to represent what nature will do. Data models follow the CA approach with distributed algorithms showing the interactions between all segments. This allows the reproducing of physical evolution and dependency between cells according to a transition function.

B. *Irregular Cell Systems with Geographic Divisions*

In this work, geographic divisions are used to generate distributed process systems for simulation with models obtained from historical data analysis. For the designer, process systems are templates embedding geographic definitions as polygon shapes associated with local states defined by a set of variables. Thus, a whole process system is a discrete spatial and temporal dynamic that has four main elements: a set of cell processes, a neighborhood, a set of initial conditions, and a transition function.

1) Cell processes

The space is segmented by geographic divisions represented as polygons on a map. A segment is a discrete interpretation of reality, which allows separating physical concerns and easing the description of behaviors. Each segment is a process in a distribute system. As an example, in Fig. 1a, the space is divided into four irregular cells with different sizes and shapes. The set of cell processes can be represented as a matrix $C_{m,n}$, where m is the number of cells and n is the number of local parameters bounding to a cell (population, temperature, social measures, etc.).

$$C_{m,n} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix}$$

2) Neighborhood

A **neighborhood** of a segment is defined as the adjacent polygons containing all segments sharing a point or a line with it (likened to the traditional Moore neighborhood) [7]. Given a set of polygons (Fig. 1a), the adjacent neighborhood represents a set of closest surrounding polygons. The interactions between neighbors follow this natural neighborhood structure. An irregular process system (Fig. 1b) illustrates the combination of distributed systems and CA approach. The cell processes in distributed systems interact with others within a local neighborhood by exchanging messages that represent physical influences. Let $G=(V,E)$ is a bi-directional graph representing a distributed system. Each node $i \in V$ is associated with a process and each edge $(i,j) \in E$ is associated with a channel connection between two processes i and j.

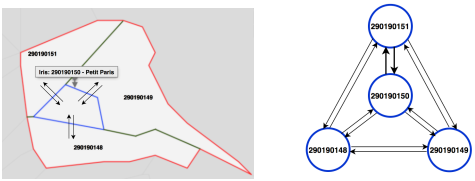


Fig. 1. a) A polygon with its neighbors; each polygon has an identification code. b) Process communication via channels between the polygons.

3) Initial conditions and synchronization

A set of variables S specifies the **initial state** of a process and all processes change their states **synchronously** at discrete time steps based on influences from its neighbors. Message passing involves processes sending and receiving messages and a barrier is synchronization point that requires all processes to reach before proceeding. The evolution time is segmented according to a given clock or predictable events. In the case of an epidemic, the state is the density of infected cases and the evolution time is one day for capturing the changes of impacted cases.

4) Transition function

The interaction and local evolution are modeled by a **transition function** (Fig. 2). A function f: $S^n \rightarrow S$ where n is the size of the neighborhood is the global transition function of the system. This function receives the states of n neighbors one time step before to determine the current state of a process. Although the behavior of each process is very simple, the interactions between all processes lead to intricate global behavior. This illustrates the complex behaviors of a system can arise from simple rules as in the Game of Life [8]. The evolution of a system is determined by an initial state and each cell changes states simultaneously over time according to the rules.

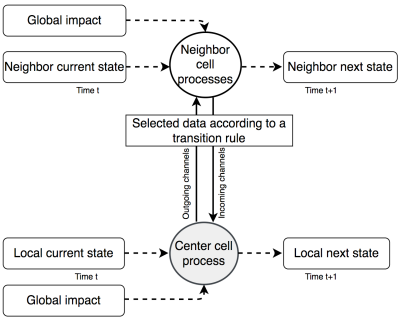


Fig. 2. Process state changes. Each process has its current state at time step t. At time step $t+1$, each process changes state according to a transition function with influences from its neighbors.

[1] https://www.insee.fr/en/metadonnees/definition/c1523

In epidemic modeling, the transition function is defined as in equation 1, where D_t is the density representing the infected cases at time step t, inf is the infection rate in a segment, s_{send} is the density sending to its neighbors, $s_{receive}$ is the density receiving from its neighbors, and r is the reduction rate.

$$D_{t+1} = D_t + D_t * inf + s_{receive} - s_{send} - D_t * r \quad (1)$$

The infection rate shows the increase of the density in the cell segment. This variable varies depending on the actual situation of each cell as temperature, humidity, and population density. A high infection rate will cause the epidemic to quickly spread out to a broader area. The reduction rate shows the decrease of the density in a cell. This variable is defined by the government measures against the epidemic spreading including social distancing and vaccination. The density D_t of the segment also changes depending on the interactions with its neighbors. Assuming that the epidemic is transmitted to its close neighbors, the infected cases of a cell moving around and transmitting the epidemic to its neighbors but they still live in their place. Thus, in this case, the s_{send} is zero. The influence ($s_{receive}$) of the neighborhood is defined in equation 2, where n is the number of neighbors, D_{t_i} is the density of neighbor i at time step t, and trs_i is the transmission rate of neighbor i. The transmission rate may vary depending on the actual situation of the epidemic in the neighbors.

$$s_{receive} = \sum_{i=1}^{n} D_{t\_i} * trs_i \quad (2)$$

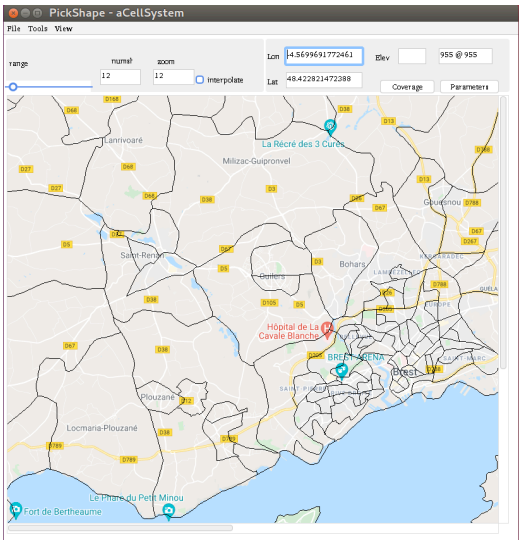


Fig. 3. A process system organization over geography divisions: Divide the surface by population; each segment is bound with a set of input data according to its identification.

C. System Generation

The process systems generation is supported by a set of tools developed in the University of Brest, France (Fig. 3). The tools allow reading geographical data and visualizing a graphical window for zone selection. Space segmentation is obtained on each geographic division and channels between segments are computed based on their adjacent neighbors. Precise information related to each division is bound to each segment using its identification. The data are stored in shapefiles or local databases and then queried in the system generation process. After loading the shapefile and zooming in to an area, the geo-coordinates of the bounding box of that area are used to query into the local database to take all divisions in the area for processes and channels generation. The generated system consists of all divisions partially or fully contained in the selected zone.

Let's consider binding population density into geographic divisions in South France. The population density per km^2 by IRIS is presented in Fig. 4. By 2019, Metropolitan France is divided into 48,590 IRIS, in which each IRIS has an average area of 11.3 km^2 and an average population of 1,323 people. Each IRIS has a clear geographical boundary that can be bound with a set of local data. The population of each IRIS [9] is bound to each IRIS process through its identification. The IRIS cell space represents the correct local data managed at government administrative levels.

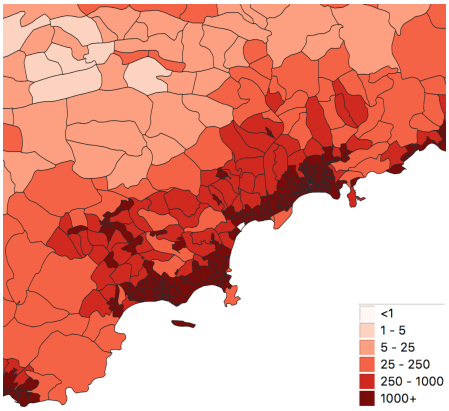


Fig. 4. South France population density in IRISs provided by INSEE.

Occam/CUDA code is generated for parallel computation and high-performance simulation on Graphics Processing Units. An illustration of the process system in Fig. 1 is presented in Occam code as below. Channel declarations describe all the channels in the system that can carry *diam.proto* data. This is a data type declared by users. Each process has an array of outgoing channels and an array of incoming channels. The *Node* describes how each process sends/receives data to/from its neighbors and updates its status. This is the user's responsibility to modify the transition functions for specific simulation problems. The *Mux* declares the data to be presented in the standard output.

```
PROC aCellSystem(CHAN OF BYTE stdin, stdout,
stderr)
  -- Channel declarations
  CHAN OF diam.proto P01.P02, P01.P03, P01.P04 :
  -- Channel table declaration for nodes
  P01.out IS [ P01.P02, P01.P03, P01.P04 ] :
  P01.in IS [ P02.P01, P03.P01, P04.P01 ] :
  [4]CHAN OF BYTE toMux:
  -- Program Body
  PAR
    Node( P01.in, P01.out,0, toMux[0])
  Mux(toMux,stdout)
:
```

III. From Parameters to Transition Rules

In this section, we analyze data related to Covid-19.

A. Selecting Parameters

The parameters related to the epidemic spreading are social distancing [10], demographic [11], mobility [4], and weather condition [6]. A data analysis process is done to detect the dependencies between the parameters and the incidence rate. The important parameters will be selected to model local evolution.

A random forest model is fit [12] for interpreting the relations between parameters in the pandemic incidence rate predictions. It is a combination of a large number of classifiers $\{h(x,\Theta_k), k=1,...\}$ where the Θ_k are independent identically distributed random vectors that cast a unit vote at input x. The common prediction outcome of these decision trees will be used as the final output. The decision trees are built in an approach to produce the purest tree nodes with impurity degree measuring methods, such as the Gini index used in classification and regression trees. Gini index [13] is calculated by the sum squared probabilities of each class from 1 (equation 3). The tools used for analyzing important parameters are mainly packages in R.

$$Gini(S) = 1 - \Sigma^{n}_{i=1} p_i^2 \quad (3)$$

Permutation feature importance [14] is a global interpretation method that will shuffle the values of the features and measure the drop in performances. An important feature is a feature that will increase the model error if permuting its values. The interpretation methods have been explored in our previous work on weather data analysis and prediction [15].

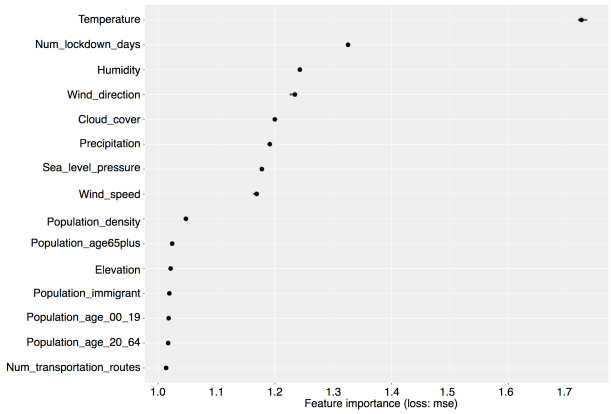


Fig. 5. Permutation feature importance. Temperature, humidity, and the number of days applied lockdown measure are the three most important features in predicting the epidemic incidence rate.

Permutation feature importance (Fig. 5) is used in this case to analyze data of the study region. The analysis shows that the important features affecting the predictions are temperature, humidity, and the number of days applied lockdown measure. The mean squared error is chosen to measure the loss in performance. Features associated with a model error increase by a factor of 1 (means no change) were considered not important for predicting the pandemic incidence rate. In Fig. 5 the most important feature is temperature associated with an error increase of 1.72 after permutation.

B. Transition Rules

Classification of parameters allow to select the most important ones for modeling the pandemic spreading after the data analysis process. The temperature, humidity, previous day incidence rate, and the neighbors' incidence rate are the factors that will decide the infection and transmission rate of the pandemic. The lockdown measure and vaccination will affect the reduction rate.

Poisson distribution [16] is a discrete probability distribution used to test the relevance of realistic stage-period distribution on the dynamics of epidemic outbreaks [17-19]. Suppose some events occur λ times with an interval, the probability P of k times occurrences of the same event in the same interval is given by equation 4.

$$P(k|\lambda) = e^{-\lambda}(\lambda^k/k!) \quad (4)$$

Poisson regression models the dependency between the response and covariates by assuming that the response y has a Poisson distribution. The dependency $\hat{y}$ is calculated as equation 5.

$$\hat{y} = e^{\lambda} \quad with \quad \lambda = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \quad (5)$$

The remaining is to find the parameters $\beta_0, \beta_1, ..., \beta_n$ which maximize the possibility P by the maximum likelihood estimation.

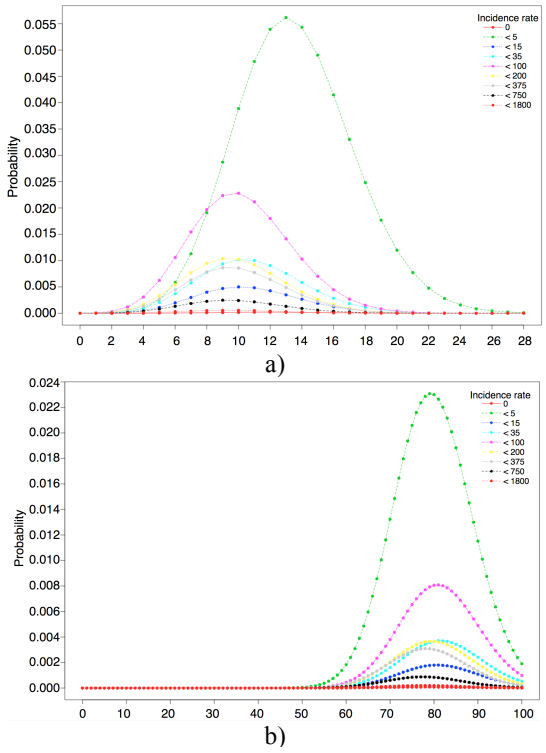


Fig. 6. Poisson distribution of incidence rate based on (a) temperature and (b) humidity.

A Generalized Linear Model with Poisson distributions is fit to model the dependency between the pandemic incidence rate and the important parameters. This will provide the β values showing changes in the parameters affecting the incidence rate as in equation 6, where t is a time step, ir_t is the incidence rate at time step t, and δ_i is the dependency of parameter n on the incidence rate.

$$ir_{t+1} = ir_t + ir_t * e^{\lambda} \quad and \quad \lambda = \sum_{i=1} \delta_i \tag{6}$$

The general value of λ is calculated as equation 6. We analyze each parameter to deduce the value of δ_i. Fig. 6 shows the Poisson distribution of incidence rate based on temperature and humidity. It is observed that the probability mass function is expected to be highest when the temperature is around 10 degrees and humidity is around 80 percent. The rules are calculated with the variables as presented in Table I.

TABLE I. VARIABLES AND EXPLANATION

Variable : Value	Variable - Explanation
β_1 : 0.005151	$temp_t$: temperature at time t
β_2 : 0.11403	$humid_t$: humidity at time t
β_3 : 0.00002795	$popu$: population density
β_4 : -0.0512048	$lock_t$: number of days applied lockdown measure
β_5 : -0.0455449	$vaccin_t$: percent of vaccination/population at time t
β_6 : 0.0008964	nir_{jt} : incidence rate of neighbor j at time t
	m : number of neighbors

The transition rules for epidemic spreading simulation are presented as follows.

Rule 1: δ_1

$$\delta_1 = \begin{cases} (temp_{t+1} - temp_t) * \beta_1 & if \quad temp_{t+1} < 10 \ and \ temp_t < 10 \\ (temp_t - temp_{t+1}) * \beta_1 & if \quad temp_{t+1} \geq 10 \ and \ temp_t \geq 10 \\ (10 - temp_t + 10 - temp_{t+1}) * \beta_1 & otherwise \end{cases}$$

Rule 2: δ_2

$$\delta_2 = \begin{cases} (humid_{t+1} - humid_t) * \beta_2 & if \quad humid_{t+1} < 80 \ and \ humid_t < 80 \\ (humid_t - humid_{t+1}) * \beta_2 & if \quad humid_{t+1} \geq 80 \ and \ humid_t \geq 80 \\ (80 - humid_t + 80 - humid_{t+1}) * \beta_2 & otherwise \end{cases}$$

Assuming that when applied lockdown measure, the transmission between people in the population is trivial, and without any measurement, the transmission probability is around 10 percent. The transmission is calculated as rule 3. There is a moderate correlation between the incidence rate and the number of days applied lockdown measures. The incidence rate tends to decrease when there is an increase in the number of lockdown days. The Pearson correlation coefficient is a measure of the strength of a linear association between two variables. The measured number is -0.36691, meaning that these two variables tend to lie on opposite sides of their respective means. The effect of lockdown measures, vaccination, and neighborhood on incidence rate is calculated as rule 4.

Rule 3: δ_3

$$\delta_3 = \begin{cases} 0 & if \quad lock_t > 0 \\ 0.1 * popu * \beta_3 & otherwise \end{cases}$$

Rule 4: δ_4, δ_5, δ_6

$$\delta_4 = lock_t * \beta_4$$
$$\delta_5 = vaccin_t * \beta_5$$
$$\delta_6 = \sum_{j=1}^{m} nir_j * \beta_6$$

IV. EXPERIMENTS

A. Data and Tools

The data used in this study is related to Covid-19 in a one-year period from May 2020 to May 2021. The study region is Brittany, France. The Covid-19 incidence rate daily and weekly is provided by the French public health agency[2].

QuickMap/PickShape is a set of tools developed at the University of Brest, France. The tools allow reading geographical shapefile and visualizing on a graphical window for zone selection and generation of irregular cell systems.

B. Simulation Results

Fig. 7a shows the average incidence rate in time series (historical data). After the ease of restrictions, during August 2020 incidence rate began to rise again. The average incidence rate continued to rise and the government decided to enter a second nationwide lockdown from 30 October 2020. This measure ends on 15 December 2020. The third national lockdown was proposed by the government starting on 5 April 2021 and lasted until 3 May 2021.

Fig. 7b shows the average incidence rate after simulation. The simulation experiments show that we can follow the trend of the average incidence rate in the study region. People send and receive viruses depending on population density, vaccination, lockdown measure, and weather condition. Each cell in the system has a state defining the number of infected cases and synchronously updates its state in a discrete time by the transition rules considering the neighborhood states. The simulation software computes epidemic spreading over irregular cell systems, allowing us to examine its evolution with transition rules synthesized from historical data. The data used is in a specific region but the method can be applied in other territories taking into account local available data.

Although the global propagation trend is caught, it has to be noted that we cannot obtain a high accuracy in the simulation because of the chaotic behaviors of the pandemic. It is observed in Fig. 7b that the simulated incidence rates are generally less than the real data collected. Another limitation of the study is that we only consider the infected cases while ignoring the hospitalized, recovered, or dead cases. This is one of the reasons for decreasing simulation accuracy.

The important parameters pointed out and used in our model are similar with the results provided by several previous studies. Bastos and Cajueiro [20] analyzed epidemic data in Brazil with the SIR model to consider the effect of the social distance policy. It was shown that the policies of social distance can flatten the contamination. However, a short-term distancing policy can only shift the peak of infection into the future. This conclusion is close with our simulation results. When there is a lack of vaccination, social distancing can only

[2] https://geodes.santepubliquefrance.fr

move the epidemic peak to the future as shown in Fig. 7b (the epidemic peak ships from November 2020 to April 2021). The emergence and replication of the virus have been shown to have a connection with weather factors. Adly Anis [21] demonstrates that the temperature plays an important role in the epidemic spreading. He showed that cool weather is the most appropriate for virus activity and transmission. Aly Kassem [22] also showed the relationship between temperature and transmission speed of the Covid-19 virus. Most respiratory viruses are known to show seasonal infection [23]. Warm places and regions have a lower risk of respiratory viruses. Cold and dry conditions can be the factors affecting the spread of the virus. This is also one of the results shown in our study that two important factors affecting the spread of the virus are temperature and humidity.

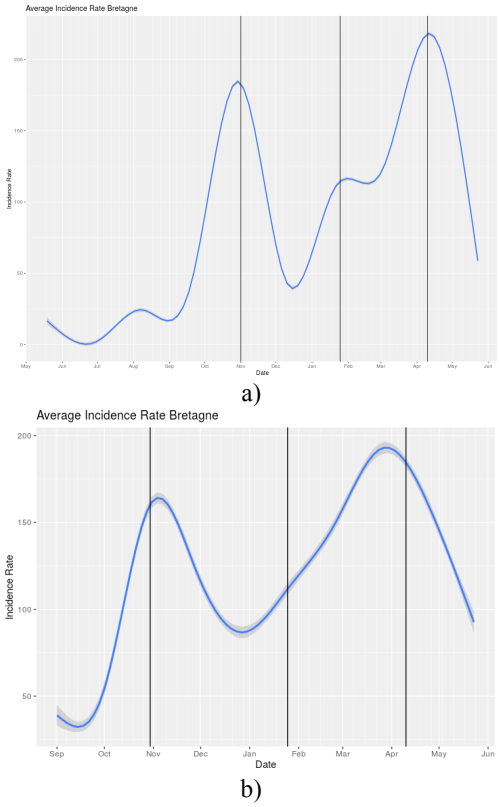


Fig. 7. The average incidence rate in time series. a) Historical data. b) Simulation results. The simulation results follow the trend of the average incidence rate in time series.

In addition, we base our simulation on transmission between neighboring geographical areas assuming that people in neighboring areas will spread the disease to others. This is a result shown in Schimit Pedro's study [24] describing the transmission of Covid-19 with a square lattice of n x n cells. The local interactions were based on an extended Moore neighborhood with a radius affecting the probability of interacting between one cell and the others. Their results show that Covid-19 is likely to spread in densely populated regions and between geographically adjacent regions since people in these regions are more likely to interact with each other.

C. *Variation of Behaviors will Assist Control*

In addition, different regions will have different orders of important parameters. Fig. 8 shows the first three important parameters in 3 different zones in Brittany, France including Brest, Carnac, and Landivisiau. We consider data in Brest (administrative and industry zone) with 27 IRIS having an average population density of 8,130 people/km^2. In Carnac (tourism zone), we take 9 IRIS having an average population density of 158 people/km^2. In Landivisiau (agriculture zone), we take 9 IRIS having an average population density of 593 people/km^2.

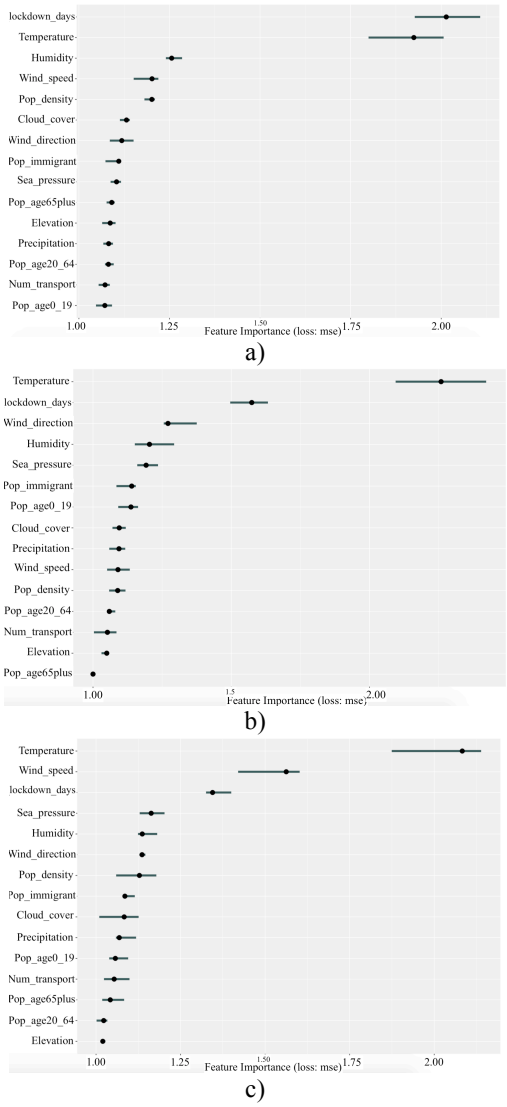


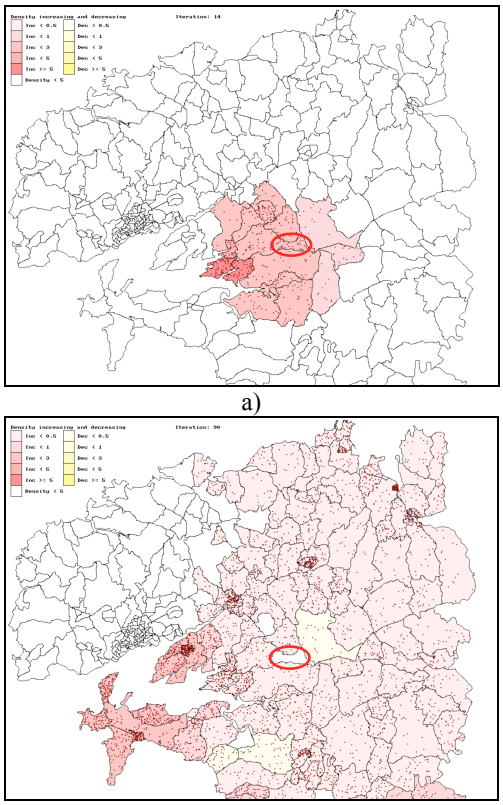
Fig. 8. Permutation feature importance in 3 different zones in Brittany. a) Brest; b) Carnac; c) Landivisiau.

The number of days applied lockdown measurement is the most important feature in places with high population density (Brest). In places with lower population density (Carnac, Landivisiau), temperature is the most important feature in predicting the epidemic incidence rate. The humidity and wind are also important features in these places. Thus, different

behaviors in zones will be taken into consideration to improve control measures. As an example, places where lockdown measures are less important are candidates to reduce the number of lockdown days after checking the simulation.

D. Case Study

We explore different transition functions with a process system generated. In modeling a epidemic spreading, it is possible to start from a pure hypothesis of a virus appearing in a restricted place. Fig. 9 shows an illustration of an epidemic spreading from the first place highlighted with a red circle. In Fig. 9a, we will discover the geometry of the epidemic spreading around the initial spot. Segments are presented on a map with color intensity representing the density changes and the points inside each segment represent the density (impacted cases). The increasing density is shown in red polygons, the decreasing density is shown in yellow polygons, and the white polygons are less affected places with density less than 5.



a)

b)

Fig. 9. Epidemic simulation with points represents the density and colored polygons represent the density changes. a) A source of contamination (in red circle) spreads to its surrounding neighbors after 14 days. b) The spreading after 90 days without control.

The transition function in this case is a variation of equation 1. We define a transition function as in equation 7, where *recv* is the recovery cases. The idea is that the infected cases will recover after a number of time steps and once people have been infected, they will be protected. The average time for recovery is 7 days in this case. The influence to its neighborhood is defined in equation 8, where *trs* is the transmission rate of the segment.

$$D_{t+1} = D_t + D_t * inf + s_{receive} - s_{send} - recv \qquad (7)$$

$$s_{send} = \Sigma^{n}_{i=1}\ D_t * trs \qquad (8)$$

Fig. 9b shows the epidemic spreading to the broader area after 90 rounds without any counteractions. It is noticed that, in some segments, the epidemic disappeared because herd immunity was achieved through previous infections. Not shown in the figures that the epidemic can oscillate between increasing and decreasing. The density increases with the infection rate and the density received from its neighbors. It decreases by the recovery cases and the density sending to its neighbors. As an example, when the neighborhood is not affected, the epidemic of a segment decreases because of sending to its neighbors and later increases when the surrounding neighbors are affected. At some points, it decreases again because of the recovery.

Spreading could be managed given government open data and measures to counteract the epidemic. As an example, a barrier with vaccination or social distancing can be erected in an effort to slow down the epidemic spreading. In Fig. 10, a *virtual wall* (in red line) was built to control the spreading to the north of the region by tuning the infection rate and the transmission rate of the polygons in the red line. After 90 rounds, the north part of the region is less affected than the case without counteractions shown in Fig. 9a.

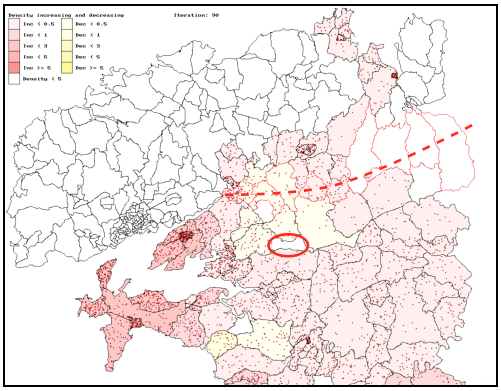


Fig. 10. The spreading after 90 days is controlled with vaccination and social distancing (polygons in the red line).

The global status of the epidemic changes with different transition functions. Another example is a variation of the transition function in equation 7. We define a transition function as in equation 9. The idea is that the density is receiving from its neighbors but not sending to its neighbors.

$$D_{t+1} = D_t + D_t * inf + s_{receive} - recv \qquad (9)$$

In Fig. 11, the epidemic starts from an initial place and explodes to a broader region. The epidemic quickly increases and achieves herd immunity. Thus, the places that were affected first will return to an unaffected state after a period of time.

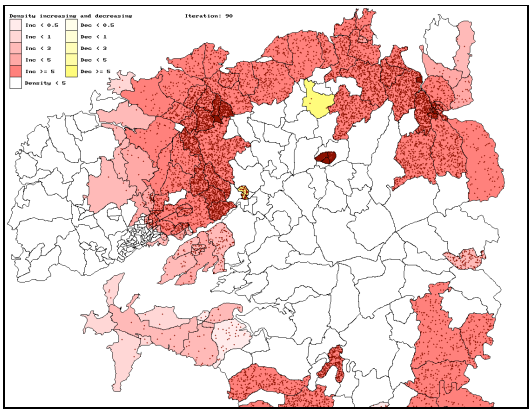


Fig. 11. Epidemic simulation starting from an initial source and explodes to a broader region.

V. Perspective and Conclusion

The Covid-19 pandemic quickly spreads worldwide and was declared a global pandemic in March 2020 by the World Health Organization. Authorities collect Covid-19 related health statistics based on geographic divisions and make it open for everyone. This study uses distributed process systems with the segmentation of geographic divisions and Open Data to model epidemic spreading. The epidemic evolution in each process depends on its local data and influences from the neighborhood. Open Data in geographic divisions are analyzed to select parameters and define transition rules for epidemic evolution. The practical interest is to understand the spreading in time and space so that control measures can be defined. Simulations allow us to warn about the risks. Given a state set, it is possible to explore capabilities for control. Simulating these measures is useful to reduce dangers and estimate the measure effectiveness. Especially, simulation allows checking counteractions in different regions with their characteristics. As an example, the lockdown measurement is more effective in one place and but not effective in another place.

References

[1] S.W. Lacey, "Cholera: calamitous past, ominous future." Clinical Infectious Diseases, vol. 20, pp. 1409–1419, 1995. https://doi.org/10.1093/clinids/20.5.1409

[2] N. Varlık, "Rethinking the history of plague in the time of covid-19." Centaurus, vol. 62, pp. 285–293, 2020. https://doi.org/10.1111/1600-0498.12302

[3] W.j. Guan, Z.y. Ni, Y.Hu, W.h. Liang, C.q. Ou, J.x. He, L. Liu, H. Shan, C.l. Lei, and D.S. Hui, "Clinical characteristics of coronavirus disease 2019 in china." New England Journal of Medicine, vol. 382, pp. 1708–1720, 2020. https://doi.org/10.1056/NEJMoa2002032

[4] M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, and A. Rinaldo, "Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures." Proceedings of the National Academy of Sciences, vol. *117*, pp.10484-10491, 2020. https://doi.org/10.1073/pnas.2004978117

[5] G.C. Sirakoulis, I. Karafyllidis, and A. Thanailakis, "A cellular automaton model for the effects of population movement and vaccination on epidemic propagation." Ecological Modelling, vol. *133*, pp. 209-223, 2000. https://doi.org/10.1016/S0304-3800(00)00294-5

[6] R.J. Doran, and S.W. Laffan, "Simulating the spatial dynamics of foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia, using a susceptible-infected-recovered cellular automata model." Preventive Veterinary Medicine, vol. *70*, pp.133-152, 2005. https://doi.org/10.1016/j.prevetmed.2005.03.002

[7] D. Stevens, and S. Dragićević, "A GIS-based irregular cellular automata model of land-use change." Environment and Planning B: Planning and design, 34(4), 708-724, 2007. https://doi.org/10.1068/b32098

[8] C. Bays, "Introduction to cellular automata and Conway's Game of Life." Game of Life Cellular Automata, pp. 1-7, 2010. https://doi.org/10.1007/978-1-84996-217-9_1

[9] INSEE: 2016 population census: sub-municipal databases - iris (2016), http://www.progedo-adisp.fr/enquetes/XML/lil.php?lil=lil-1369, distributed by ADISP

[10] M. Qian, J. Jiang, "COVID-19 and social distancing." Journal Public Health (Berl.) 30, pp. 259–261, 2022. https://doi.org/10.1007/s10389-020-01321-z

[11] V. Colizza, R. Pastor-Satorras, and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks." Nature Physics, vol. 3, pp. 276–282, 2007. https://doi.org/10.1038/nphys560

[12] L. Breiman, "Random forests." Machine Learning, vol. 45, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[13] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and regression trees. Routledge, 2017.

[14] C. Molnar, Interpretable machine learning. Lulu.com, 2020.

[15] N.T. Trieu, B. Pottier, V. Rodin, and H.X. Huynh, "Interpretable machine learning for meteorological data." In The 5th International Conference on Machine Learning and Soft Computing, pp. 11–17, 2021. https://doi.org/10.1145/3453800.3453803

[16] M. Sankaran, "275. note: The discrete poisson-lindley distribution." Biometrics, pp. 145–149, 1970.

[17] N. Hernandez-Ceron, Z. Feng, and C. Castillo-Chavez, "Discrete epidemic models with arbitrary stage distributions and applications to disease control." Bulletin of Mathematical Biology, vol. 75, pp. 1716–1746, 2013. https://doi.org/10.1007/s11538-013-9866-x

[18] C. Kremer, A. Torneri, S. Boesmans, H. Meuwissen, S. Verdonschot, K.V. Driessche, C.L. Althaus, C. Faes, and N. Hens, "Quantifying superspreading for covid-19 using poisson mixture distributions." Scientific Reports, vol. 11, pp. 1–11, 2021. https://doi.org/10.1038/s41598-021-93578-x

[19] P. Yan, "Distribution theory, stochastic processes and infectious disease modelling." Mathematical Epidemiology, pp. 229–293, 2008. Springer. https://doi.org/10.1007/978-3-540-78911-6_10

[20] S. B. Bastos and D. O. Cajueiro, "Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil." Scientific Reports, vol. 10, pp. 1–10, 2020. https://doi.org/10.1038/s41598-020-76257-1

[21] A. Anis, "The Effect of Temperature Upon Transmission of COVID-19: Australia And Egypt Case Study." Available at SSRN 3567639, 2020. http://dx.doi.org/10.2139/ssrn.3567639

[22] A. Z. E. Kassem, "Does temperature affect COVID-19 transmission?" Frontiers in public health, vol. 8, 554964, 2020. https://doi.org/10.3389/fpubh.2020.554964

[23] P. Mecenas, R. T. d. R. M. Bastos, A. C. R. Vallinoto and . Normando, "Effects of temperature and humidity on the spread of COVID-19: A systematic review", PLoS one, vol. 15, e0238339, 2020. https://doi.org/10.1371/journal.pone.0238339

[24] P. H. Schimit, "A model based on cellular automata to estimate the social isola- tion impact on COVID-19 spreading in Brazil." Computer methods and programs in biomedicine, vol. 200, 105832, 2021. https://doi.org/10.1016/j.cmpb.2020.105832

5.4 Shore Pollution Simulation Based on Tidal Currents and Ground Effects

Shore Pollution Simulation Based on Tidal Currents and Ground Effects

Ngoan Thanh Trieu[1,2], Hiep Xuan Huynh[2], Vincent Rodin[1,*], and Bernard Pottier[1]

[1]Lab-STICC, UMR CNRS 6285, Computer Science Department, University of Brest, Brest, France
[2]College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam
Email: ttngoan@cit.ctu.edu.vn (N.T.T.); hxhiep@ctu.edu.vn (H.X.H.); vincent.rodin@univ-brest.fr (V.R.); bernard.pottier@univ-brest.fr (B.P.)
*Corresponding author

Manuscript received December 5, 2023; revised January 8, 2024; accepted January 11, 2024; revised May 27, 2024

***Abstract*—Marine pollution comes from different sources including agricultural, industrial, and domestic wastewater discharge from human activities in coastal areas. Environmental simulation can represent ground and sea characteristics, modeling spreading occurring in both spaces. These characteristics are variable, due to soil capability and reaction, and sea behavior, in particular currents and tides. This work presents a heterogeneous tiling approach modeling sea behaviors in coastal areas based on tidal currents and ground effects. The ground is segmented into irregular cells following administrative divisions for collecting observations while the sea area is segmented into regular geographical tiles. The impact of the interactions is represented by messages carrying qualities and quantities of physical pollution. Channels link cells following cellular automata or distributed system paradigms. This system architecture allows to produce a synchronous message passing program suitable for massive parallel execution. The status of cells and messages are produced step by step and can be interpreted graphically. Green tides caused by eutrophication appear when nutrients circulate in high concentrations in coastal waters. These nutrients come from land use, accumulate, and propagate to the shores mainly through rivers end up joining the sea or the oceans. Our simulations show when and where tides are able to increase concentration levels, producing space and time characteristics.**

***Keywords*—environment simulation, monitoring shores, distributed systems, hybrid systems, tidal currents**

I. Introduction

Coastal interactions appear to be mainly from ground to sea, especially in the case of intensive agriculture, urban development, and industry. The nature of these exchanges is chemical, biological, or sediments, with sometimes huge quantities of pollutants sent to the sea (nitrates, nitrites, phosphates, or biological). These exchanges are in fact bi-directional, ocean activity producing, in turn, sediments and biological effluents that are spread back to the coast. A first classification inside this system can distinguish a set of behaviors: ground circulation (1) depending on the nature of ground pollutants (2), ocean circulation (3), and the nature of spreading (4). The introduction will review this system, and explicit quantities published for an intense agriculture region, with a strong influence of marine currents. Then, we will focus on sea spreading, to demonstrate basic circularity due to tide currents. A predicted schedule of tide directions and strength will allow for simulation spreading revealing places with a harmful sea-to-ground accumulation, after periodic effects of tides.

Ground circulation: Pollution spreading on land can be approximated by elevation analysis or observed by sensors. During rainfall events, water runoff picks up pollutants from the land and carries them into water bodies. This runoff contains sediments, nutrients, pesticides, and other chemicals. The river system is the main connection for sending ground impacts to the oceans. Another connection is drainage basins propagating pollutants within the soil. Heavy rainwater propagation can be simulated by geographic segmentation according to a grid, augmented by elevation. Truong *et al*. [1] presented the real case of a heavy rain event with a flooding effect on the ground, computed over regular tiles. Monitoring ground pollution can be achieved by dedicated sensors and data collection. The observation results are produced and published following geographic boundaries. The core segmentation is geographic divisions also used to collect or observe human activities associated with pollution sources. Fig. 1 displays a map of infected beaches in Brittany, France.

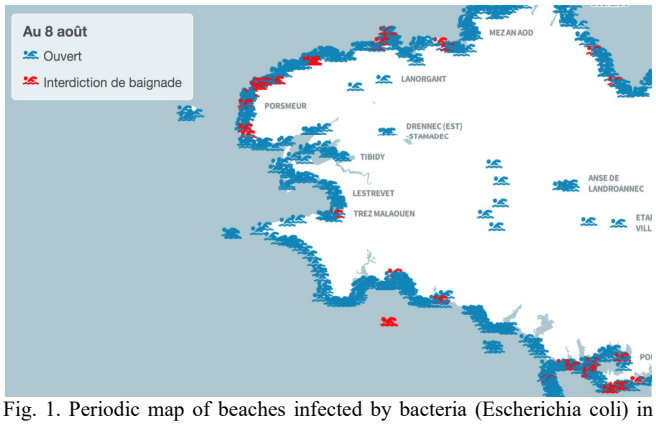


Fig. 1. Periodic map of beaches infected by bacteria (Escherichia coli) in Brittany, France. It is changing with rain, temperature, and tourism activities (Le Telegramme: 08/08/2023).

Nature and quantity of ground pollutants: Pesticides, chemical waste, cleaning products, petroleum products, mining waste, rubbish, and sewage are examples of marine pollutants of land-based origin [2]. Nitrogen-based fertilizers in agricultural activities flow to rivers and estuaries before ending up in the sea, where they encourage the growth of destructive algae, deplete the ocean's oxygen supply, and create several dead zones in which marine life cannot survive. Furthermore, in 2000, the population in coastal provinces in China reached 529 million (whole nation population: 1.29 billion), uncounted seasonal tourism, and temporary inhabitants engaged in construction work and other short-term employment. Sewage discharges into the ocean of more than 10 billion tones with dominating pollutants of

inorganic nitrogen, phosphate, and heavy metals [3]. The contamination from land runoff caused red tide events to occur 28 times involving an area of 10,000 km^2 in 2000 and 77 times in 2001 affecting an area of 15,000 km^2.

Ocean circulation: The oceanic model is made of currents carrying objects. SHOM is the French's Naval Hydrographic and Oceanographic Service [4] that operates for reference maritime and coastal geographic information. One of the products provided by SHOM is the tidal current prediction [5] including the components of surface tidal currents hour by hour and for two characteristic tidal coefficients (45 and 95). The data come from calculations of finite element models as also developed using TELEMAC-2D. It is a computational system that calculates free surface flows in two dimensions of horizontal space.

TELEMAC-2D is an open-source system, which is helpful in developing a model suitable for different situations. The system is of general application to hydraulic problems as it is able to take into account various phenomena. Awad and Darwich [6] used the TELEMAC-2D model to simulate the quality of seawater affected by wind and local currents. The aim was to determine how much the pollutants outlet into the sea with the sea current intensity and direction. Three scenarios were taken into consideration with the average daily spill of sewage produced by the inhabitants living in the river basin. Li *et al*. [7] studied the impact of building representations on urban flooding using the TELEMAC-2D model. By physical experiments and numerical modeling, the results showed that it is largely applicable to the simulation of urban flooding although there are some differences exist in the simulation results. The model was used in [8] to simulate the dynamics of the rivers during a flood period in Brazil. The results were validated using the city flood map provided by the government.

Nature of spreading: The nature of spreading in sea areas is different with different kinds of pollutants. As an example, oils tend to spread out on top of the water for hundreds of nautical miles. Plastic trash can be transported around the world with ocean currents. Waves produce sprays carrying back marine pollution to the ground. Aerosolized toxins from harmful algae blooms are transported on land with wind effects and subsequent exposure and inhalation of the generated aerosols can induce adverse human health effects [9].

The rest of the paper is as follows. Section II presents a brief discussion of marine current estimation methods. Section III shows the tidal currents modeling based on 2D vectors. We present illustrations of water circulation and pollution monitoring with tidal currents in Section IV and a conclusion was drawn in Section V.

II. Marine Currents

A. Physical Causes

Marine currents are complex motions of water produced locally or globally by physical forces: wind, water density differences, and tides resulting from sun and moon attraction and Coriolis effect. The wind pushes against the sea surface and drives currents in the down-wind direction, with strength decreasing following depth. Winds have a tendency to drive localized currents along coastlines, which can lead to phenomena like coastal upwelling. Differences in temperature (thermal) and salinity (haline) drive a vertical circulation that transports heat from the tropics toward the poles. Thermohaline circulation-driven currents are significantly slower moving than tidal or surface currents and can be found in both deep and shallow ocean depths. Currents driven by wind and thermohaline circulation are non-periodic currents to be predictable since they are associated with changing weather. The gravitational pull of the moon and sun causes tides. Water moves up and down over a long period of time during tides. The currents produced by tides are strongest near the shore, in bays, and in estuaries along the coast. They are referred to as "tidal currents". The currents are generally measured in knots (1 knot = 1.85 kilometers per hour) and the direction of a current is the direction it is headed for or where the current is flowing towards counted from 0^o to 360^o (0^o being the geographic north) clockwise. Strong tidal currents can travel at speeds of eight knots or more.

B. Currents Estimation and Measurement

Current modeling is critical for people working in marine activities.

1) Measurement

Sensing observation: Current measurement is typically done with sensors, grouped as single-point current meters and current profilers. A current meter is an oceanographic device for flow measurement by mechanical, tilt, acoustical, or electrical means. It is an instrument for measuring the velocity of the flow of a fluid in a stream. A magnetic compass can be incorporated into the current meters to determine the orientation of the instrument with respect to the magnetic north. Single-point current meters will only measure the current in the exact depth where they are installed [10]. As sea currents vary significantly with depth, it's preferred to measure the current profile for the entire water column by using a dedicated tool. The Acoustic Doppler Current Profiler is extensively used to measure ocean currents [11]. The American National Oceanic and Atmospheric Administration scientists (NOAA) typically deploy Acoustic Doppler Current Profilers to measure currents throughout the water column at various locations for a period of one to four months. The device operates using reflections of the sound wave from drifting particles for the measurement. The profiler sends an acoustic signal into the water column and that sound bounces off particles in the water. The instrument calculates the speed and direction of the current by knowing the frequency of the return signal, the distance it traveled, and the time it took for the signal to travel.

Radar observation: Radio antennas and high-frequency Radio Detecting and Ranging systems are used to map surface ocean current patterns over a large area in coastal areas [12]. These shore-based instruments use the Doppler effect to determine when currents are moving toward or away from the shore or to measure the velocity of a current. The utilization of high-frequency radar systems in coastal areas has rapidly increased alongside the use of moored current meters [13]. It has been demonstrated that coastal

high-frequency radar can resolve quick changes. However, their coverage remains limited although the number of high-frequency radars has been augmented.

Satellite observation: A combination of surface current measurements by satellite and high-frequency coastal radar is a promising approach to cope with both the resolution and fast dynamics characteristic of coastal areas and the medium size and slower evolution of surface currents in the open ocean regions. Satellites provide information about the ocean bathymetry, sea surface temperature, ocean color, coral reefs, and sea ice. The sea surface temperature shows patterns of water circulation characterized by cold water and warm water currents.

2) Prediction

Observation data are collected by different types of sensors (currents, salinity, temperature, oxygen, and pressure) at selected locations to track the movement of various water masses. These data are used to generate current predictions in the locations. To create accurate prediction models, it is necessary to periodically resurvey various coastal and estuarine locations [14]. Extensive numerical models have been used to study the characteristics of tidal currents based on physical and geographical parameters temperature, salinity, water depths, vertical turbulent viscosity, and diffusion [15]. These numerical models need to be assessed and calibrated using observational data, such as observation with current sensors, to produce reliable current models.

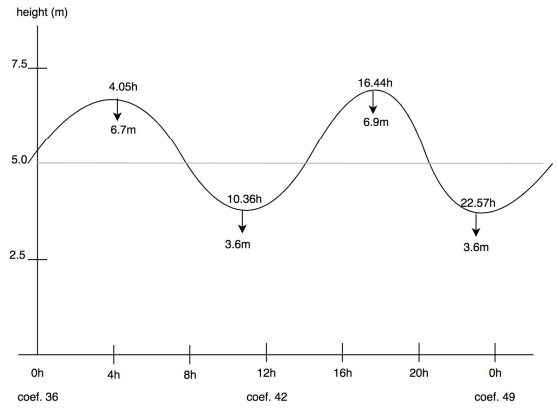
III. Tidal Current Model

In this section, we will introduce the principles of tides, tidal estimation from an oceanographic service, and a formal model of tidal currents' velocity and direction used in this study.

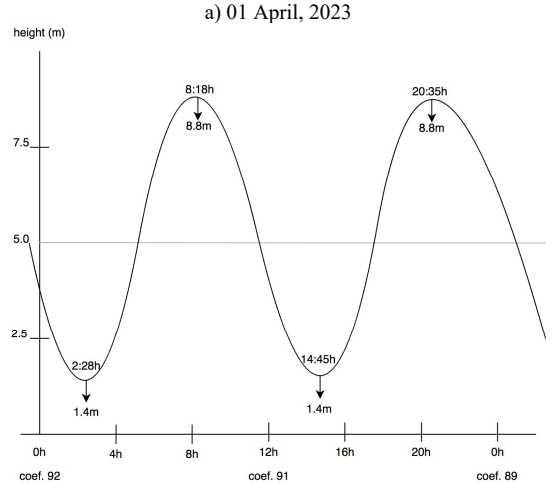
A. Tide Principles

Tidal currents are periodic currents making prediction for future dates possible. They embed height range and speed that affect coastal activities such as fishing, shipping, and tourism. Tidal currents can break up some pollutants or carry others back to shore and contribute to shoreline changes due to erosion.

The *tidal coefficient* represents the height of the tide in relation to its mean tide. Typically, it fluctuates between 20 for low tides and 120 for high tides. The strong tides, called spring tides, occur near the times of a new moon or full moon when the Earth, moon, and sun attraction are all lined up so that the tidal bulges caused by the moon and by the sun add to each other (mean spring tides: coefficient 95). The weak tides, called neap tides, occur near times of first quarter and third quarter the tides when the moon and the sun work against each other (mean neap tides: coefficient 45). The tidal range, or the variation in water height between high and low tide, increases with increasing tidal coefficient. In the example shown in Fig. 2, the tidal range is 3.1 m corresponding to lower coefficients (around 40) and the tidal range is 7.4 corresponding to higher coefficients (around 90). This indicates that the water level rises and retreats considerably. The mean water height is different between places. In our example of Roscoff, it is 5 m, whereas the Bay of Mont Saint-Michel has a mean water height of 7.5 m.



a) 01 April, 2023



b) 08 April, 2023

Fig. 2. Water height in relation with tidal coefficients in Roscoff, France. Larger coefficients: water level will increase and decrease in a large range indicating a strong flow of water. Smaller coefficients: water level will increase and decrease in a small range indicating a weak flow of water. Increasing water level means the water move towards the English Channel (a) and decreasing water level means the water move away from the English Channel (b).

B. Tidal Estimation Following SHOM Model

While predicted tide coefficients are used by many users, prediction systems are only used by scientific departments, thanks to astronomic considerations known in advance. In France, prediction data are produced by SHOM. SHOM also provides algorithms allowing to compute strength according to discrete time and location.

Two basic measurements are needed for the estimation of the tidal current velocity [16]. The first data parameter is the spring water level H that will be needed for the evaluation of the tidal coefficient. It represents the difference between high and low tides. The second data parameter groups the spring and neap water current velocities, v_{sw} and v_{nw}, respectively. The data parameters are produced by SHOM depending on geographic locations. These data are formulated as tables communicated in CSV format. An example of the current data is shown in Fig. 3.

```
Roscoff
4843.628 -358.975
0 0 4 12 11 8 3 -6-11-11 -7 -1 0 * 0 0 0 -1 -1 -1 -1 0 0 1 -1 0 0
2 5 7 8 7 5 1 -5 -8 -7 -6 -4 -2 * 0 0 -1 -1 0 0 0 0 0 0 0 0 0 0
```

Fig. 3. Tidal current in one data point in Roscoff from SHOM.

Each data file includes a header for the reference shore name, followed by geographical coordinates then the surface

vertical *v* and horizontal *u* components hour by hour. The first line is for spring water currents, the second line is for neap water currents. The data points are presented as colored dots in Fig. 4 grouping four overlapping systems with different grid paths. In the data, spring tides have a coefficient of 95 and neap tides have a coefficient of 45. Measures are taken hour by hour, from −6 hours to +6 hours in reference to the time in which occurs the tides in 0.1 knots. The *u* and *v* components are separated by an asterisk.

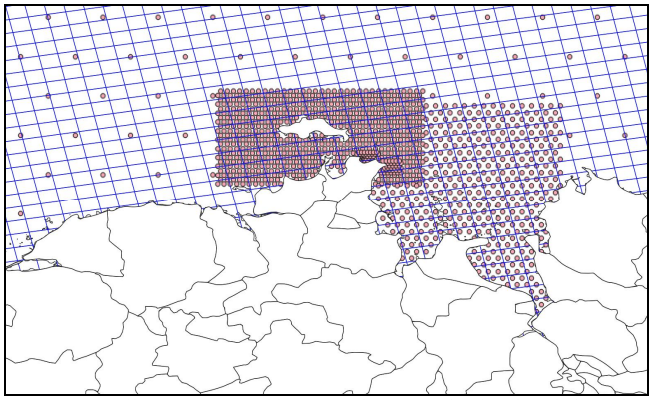


Fig. 4. Ocean tile versus ground layout: polygons on land and grid aligned on 1 km^2 in the ocean. Tidal currents are presented as data points depending on SHOM scales choice.

C. Formal Model as 2 Dimensional Vectors

The current velocity is calculated by equation 1 [4]. In which, a_0 is the tidal coefficient for neap tides (mean 45) and b_0 is the tidal coefficient for the spring tides (mean 95). The current direction is calculated by equation 2 also depending on the mean tidal coefficient of spring and neap tides.

$$v = v_{nw} + \frac{C - a_0}{b_0 - a_0}(v_{sw} - v_{nw}) \quad (1)$$

$$d = d_{nw} + \frac{C - a_0}{b_0 - a_0}(d_{sw} - d_{nw}) \quad (2)$$

where:

v: velocity in a place

$v_{nw} = \sqrt{u_{45}^2 + v_{45}^2}$: velocity in neap tides

$v_{sw} = \sqrt{u_{95}^2 + v_{95}^2}$: velocity in spring tides

C: coefficient in a time

d: current direction in a place

$d_{nw} = a\tan 2(u_{45}, v_{45}) * 180/\pi + 180$: velocity in neap tides

$d_{sw} = a\tan 2(u_{95}, v_{95}) * 180/\pi + 180$: velocity in spring tides

u_{45}, v_{45}: the horizontal and vertical components in coefficient 45 provided by SHOM

u_{95}, v_{95}: the horizontal and vertical components in coefficient 95 provided by SHOM

To model the impact of ground activities on marine pollution, we need to represent the two spaces in simulation systems. So, the sea area is segmented into regular tiles and the ground area into irregular cell spaces. Doing this way, the impact of the ground to sea can be represented by interactions between neighboring cells. The simulation systems are generated and used in experiments as described below.

IV. Tidal Current Simulation

A. System Generation

On land, geography divisions are used to query local values monitored in the territories (Fig. 5). In the ocean, the segmentation on the ocean is regular tiles with each cell having a size of 1 km^2 provided by the European Environment Agency [17]. For each country in the European Union, three types of vector polygon grid shape files for 1, 10, and 100 km^2, are available covering at least the country borders plus a 15 km buffer (not reflecting the extent of the territorial waters). As can be seen in Fig. 4, the 1 km^2 grid is different from the SHOM grid of currents. Thus, we use the nearest point interpolation method to estimate currents in the 1 km^2 grid. The ground and ocean segments are connected with channels carrying messages at an abstract level. Messages are the only way to exchange inside the neighborhood.

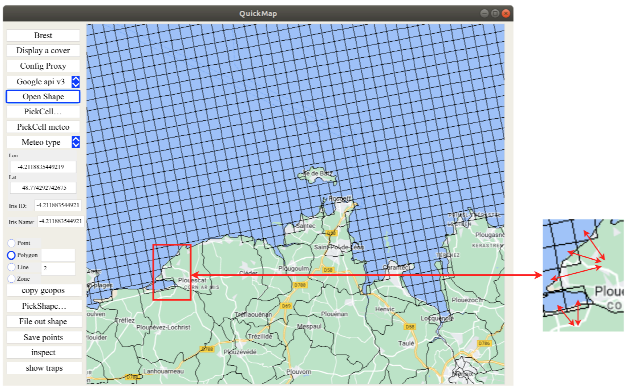


Fig. 5. Segmentation on ground and sea area. Irregular segments on ground as the administrative divisions by government and regular tiles on the ocean. UBO tools allow zooming and panning for zone selection and generating process systems in the selected zone.

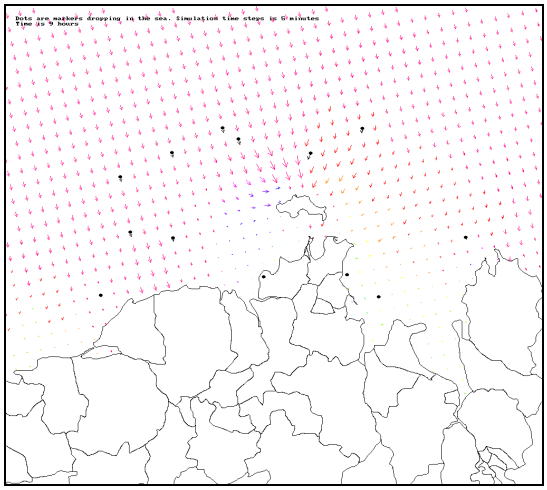
Process systems are generated as the cooperation between irregular polygons in land and ocean tiles. This cooperation provides a new approach to environmental simulation, especially pollution on shores related to chemicals released from land. It produces a process system similar to Cellular Automata including polygons and a grid of 1 km^2. Each process interacts with other processes within the adjacent neighborhood. The simulation in coastal areas strongly depends on tidal currents. The initial conditions are tidal currents in tiles and local data collected on land. UBO tools allow to zoom and pan to select specific locations and generate cell systems with adjacent neighborhoods in the location depending on input shapefiles.

B. Water Circulation with Currents

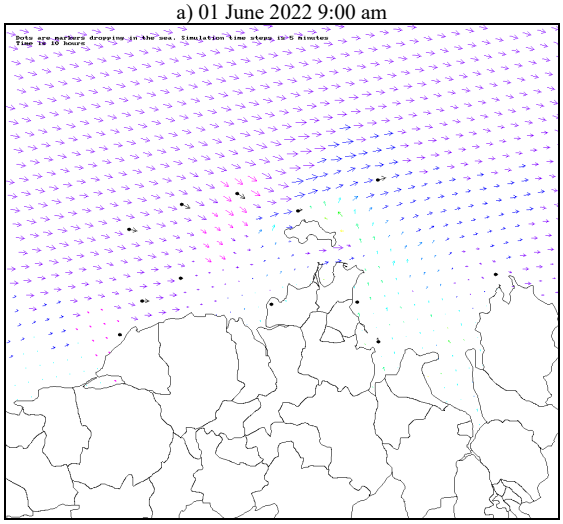
Experiments are presented with a system consisting of 891 cells (with 854 grid cells). In Fig. 6, current directions are represented by arrows and current strengths are the length of arrows. The current direction and strength will be changed hourly with different tidal coefficients. In this case, we use tidal coefficients provided by SHOM in Roscoff harbor [18] in a period of 2 months, June and July 2022.

Water circulation is described by currents in a sea area. This is useful in monitoring marine pollution since water flow transports and spreads pollution in aquatic environments. In areas with limited water circulation, pollutants tend to accumulate and persist, leading to serious environmental and

ecological consequences. Lack of water circulation promotes the accumulation of pollution, such as the growth of harmful algal blooms.



a) 01 June 2022 9:00 am



b) 01 June 2022 10:00 am

Fig. 6. Marker movement by currents. Current direction is represented by arrows and current strengths is the length of arrows. Notice the decreased current strength and the direction, reflecting orientation of the sea entering the English channel. Also notice the influence of Ils-de-batz on current direction.

As shown in Fig. 6(a), the global direction of the currents is southeast with flood currents (water moves toward the shore). After an hour, in Fig. 6(b), the global direction is changed along with the current strength. The figures reveal three retention zones with a lack of water circulation that can concentrate and retain pollution without being advected offshore by currents. These places show the need for regular monitoring to identify the potential sources of contamination. This is crucial to prevent the development of high pollutant concentrations in marine coastal waters.

C. Monitoring Virtual Markers

We monitor virtual markers moving in sea areas with currents. Following the cellular automata approach, the simulation is based on synchronous processes communicating over channels. These processes execute a single program operating on local data and messages representing neighbor interactions. Data are tidal currents considering geographical locations. A simulator fires all the cells in parallel, taking care of transports, cell to cell, whatever the nature of the transport. Simulation steps follow the physical application rate, and the necessity to cover large time periods.

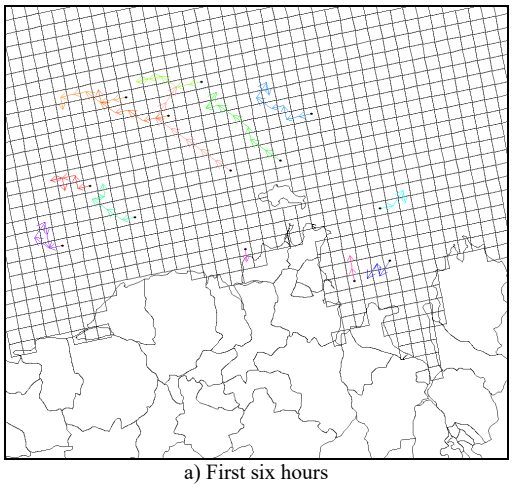
The maximum current strength reaches 7 knots (12.95 km/h) on Brittany shores, France. Thus, to catch the movement between cells in 1 km^2 (move 1 cell at most by each iteration), the simulation time step is 5 min. We drop virtual markers in the ocean area and monitor the movement of markers as shown in Fig. 6, where 13 markers are represented by black dots. The marker movement in grid cells is calculated by current speed and direction as shown in Algorithm 1.

Algorithm 1: Marker movement in grid cells

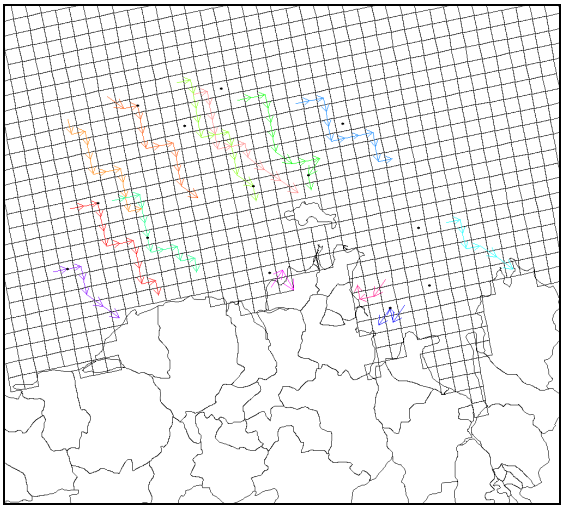
```
Initialize n markers in n random grid cells
if This cell has a marker then
    Calculate current speed;
    Calculate the marker position in this cell;
    if marker position > 1 km then
        Move marker to my neighbor;
    foreach n: neighbors do
        if n move a marker to its neighbor then
            Calculate n current direction;
            Calculate n current speed;
            //check moving by current direction
            if n move a marker to this cell then
                Update marker position in this cell; //by current speed
                Update marker identity for tracking;
end
```

It is possible to keep a history of visited places of each marker to track biological or physical alterations (Fig. 7). Fig. 7(c) shows the positions of markers after 4 days. It can be seen that there are loops in the direction of the markers. Some markers close to the coastal line cannot escape to the sea affected by flood and ebb currents. A clearer view is provided in Fig. 7(a) and 7(b). These figures show the marker movement in different places in the first 6 hours and the second 6 hours.

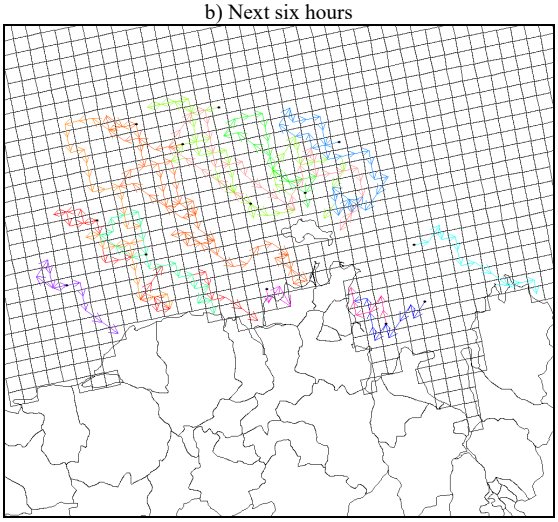
A subsystem in the simulation is that cells will have pipelines to slow down transfer objects internally. Transfers are achieved upon the agreement of the neighbors. Management of virtual markers in studied zones is similar to pollution accumulation on shore: with counting bags storing objects in excess. The local count represents the density of objects (pollution). This gives a potential way of monitoring the spatial and temporal distribution of marine buoyant plastics by providing the dynamics and pathways of how plastic waste moves in the sea area and enters the shores.



a) First six hours



b) Next six hours



c) Four days

Fig. 7. Marker positions after simulation with tidal currents showing loops in direction.

D. Monitoring Pollution Coverage Area

In the same space, green algae development can be monitored with the effect of currents. Ground information is coming from administrative data collection, with known timed samples of activities producing pollution. Colored polygons represented the area with nitrate concentration [19] on land as shown in Fig. 8. These polygons contain sources of pollution that propagate nutrients to coastal areas.

Nitrate is monitored in Brittany, France since 1995 as a parameter of water condition. In 2020, 718 stations were taken to evaluate nitrate concentrations (c) based on five classes: bad ($c > 50$ mg/L), poor ($25 < c \leq 50$ mg/L), medium ($0 < c \leq 25$ mg/L), good ($2 < c \leq 10$ mg/L), and very good ($c \geq 2$ mg/L). The limitation value is 50 mg/l used as an indication of bad water condition. The Brittany region is classified as a "vulnerable zone" with high nitrate concentration reported. Action programs have been initiated in these territories concerning balanced fertilization and good agricultural practices that must be respected. In agriculture, nitrates are essential plant nutrients, but in excess amounts, they have a negative impact on the water's quality. Together with phosphorus, nitrates in excess amounts can accelerate eutrophication causing a dramatic rise in aquatic plant growth as well as changes in the types of plants and animals that inhabit the stream.

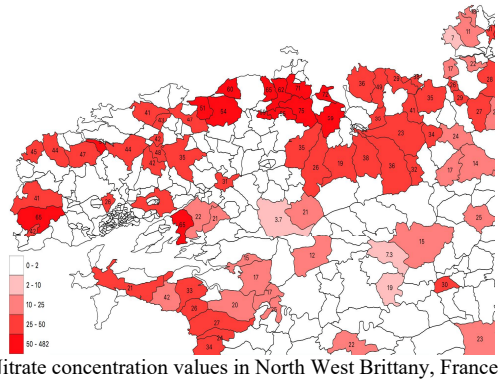
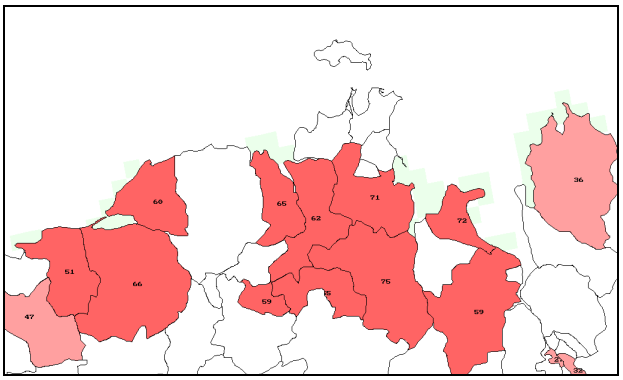
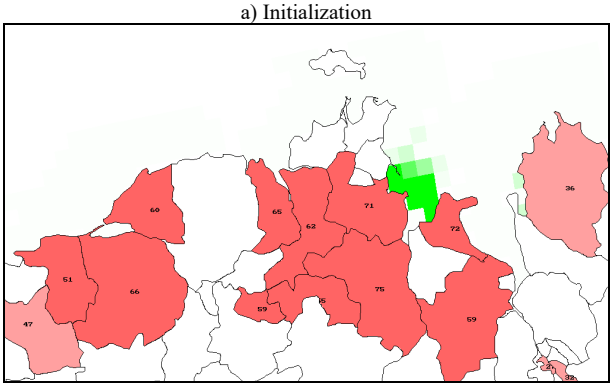


Fig. 8. Nitrate concentration values in North West Brittany, France in 2020. The values are presented inside polygons in mg/L units.

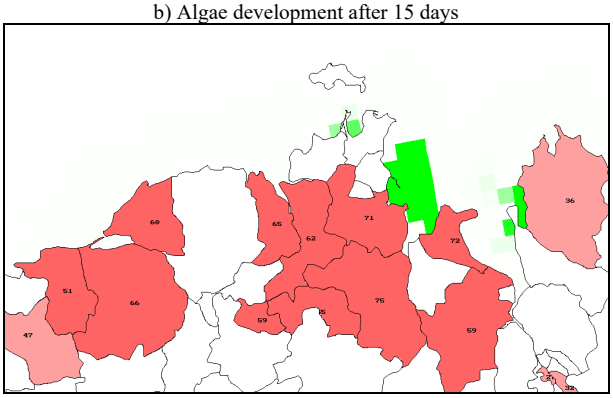
Fig. 9 shows the simulation of green algae development affected by currents. We initialize grid cells in the sea area with algae density (green color) in places close to colored polygons (Fig. 9 (a)). The interaction between grid cells is followed by a Moore neighborhood with a simple transition rule defined as in equation 3. The algae density is the growth rate and the density sending/receiving to/from its neighbors. Density transmission between cells is affected by current strength and current direction.



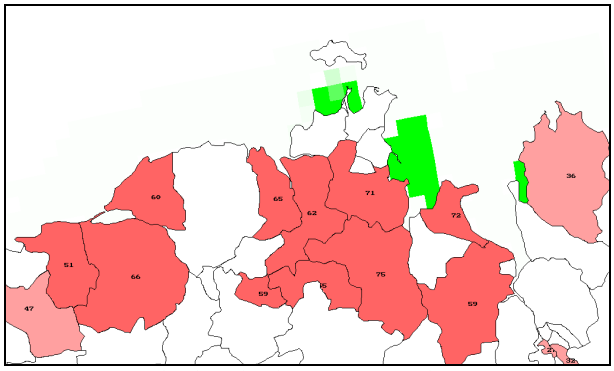
a) Initialization



b) Algae development after 15 days



c) Algae development after 40 days



d) Algae development after 55 days

Fig. 9. Green algae development affected by currents. Algae density is represented in green color. Initialize grid cells in sea area with algae density in places close to colored polygons, which are sources of pollution that propagate nutrients to coastal areas.

$$D_{t+1} = D_t + D_t * r - sum_{send} + sum_{receive} \quad (3)$$

where:

D_t: algae density in a cell at time step t

r: growth rate per day, in this simulation, it is 75% meaning 0.0026% per iteration of 5 minutes

sum_{send}: algae density sending to its neighbors calculated by Eq. 4

$sum_{receive}$: algae density receiving from its neighbors calculated by Eq. 5

$$sum_{send} = \frac{D_t * \sqrt{u_C^2 + v_C^2}}{100} \quad (4)$$

where:

u_C and v_C: tidal current of coefficient C represented by west-east and south-north components

Sending a part of its density depending on its current strength

$$sum_{receive} += D_i * \frac{v_i}{100} * \alpha \quad (5)$$

where:

D_i: algae density in cell neighbor i at time step t

v_i: current strength of cell neighbor i

α: a fraction receiving from neighbor i depending on its current direction with west-east and south-north components

As shown in Fig. 9(b), after 15 days, the algae density develops in the area near Saint-Pol-de-Leon whereas other beaches show less chance of growing green tides due to water circulation. The sea currents spread progressively algae to neighbor cells, especially in retention zones. A cell will send its algae density to its neighbors with a percentage depending on its current strength. It can send to at most three neighbors depending on its current direction with west-east and south-north components. After 40 days (Fig. 9(c)) and after 55 days (Fig. 9(d)), the algae density develops in the area near Roscoff.

The eutrophication phenomena are generally confined to large enclosed water systems. The nutrients are entrapped by tidal currents causing eutrophication phenomena. In open areas with suitable dilution factors and widespread dispersal, they would not give rise to eutrophication phenomena. The simulation results show three places with the possibility of algae development problems. Especially, the polygons in Roscoff do not contain sources of pollution that propagate nutrients to the sea area. Thus, it is important to simulate the pollution propagation with water flow to examine the pollution sources to have the correct and appropriate solutions.

V. Conclusion

Based on the cooperation between regular and irregular cell spaces, we show how to model a variety of pollution on shores. The green tides are discussed as the major reason for eutrophication accelerating by excess amounts of nitrates from agricultural runoff. Modeling ground activities can be done in two ways, simulation propagation of pollution sources on a 2D ground with elevation differences or querying observation data within administrative boundaries. Modeling ocean behaviors is done based on tidal currents in a regular grid of 1 km^2. An illustration of tidal currents is provided with marker movements in the coastal area near Roscoff, France. We also provide a simple model of green algae development in the effect of tidal currents with nitrate input sources in the same place. The simulation mechanism reveals risk from ground activities, with spatial and time characteristics. It can be used for a wide variety of accidental or systematic activities. This provides a general approach to model pollution on shores with currents showing water circulation in places suspected of pollution, which can be employed as a quick assessment tool for studying marine pollution issues. It is possible to test different scenarios by modifying the initial conditions and the transition function. One could track the pollution propagation over time given a source of pollution. Additionally, making the simulation accessible to policymakers will help support them in making decisions related to marine pollution prevention. Specifically, the marine areas vulnerable to pollution will need to be monitored more closely. Different coastal localities may have different safety indicators for pollutants due to differences in water circulation. Overall, it is important to manage ground activities in a way that reduces the inputs of nutrients and other pollutants into coastal waters and promotes the health and resilience of coastal ecosystems. The limitation side is that tidal currents need to be re-computed in different places with respect to the geographical locations and bathymetry. In addition, tidal currents showing water circulation are important in monitoring the shores however it is also important to take into account the effects of other parameters such as wind, waves, and water surface temperature. The random behavior of spreading pollutants on the surface of the sea is highly affected by wind and wave.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization and methodology were formulated by Bernard Pottier and Vincent Rodin. Trieu Thanh Ngoan conducted the research, analyzed the data, and wrote the drafted manuscript. Bernard Pottier designed two compatible systems for sensor networks (NetGen), then physical

simulation (PickCell). These developments have supported international research initiatives funded by French foreign office. In the present work, he also shared a marine knowledge coming from his favorite recreative activity cruising and racing sailboats. Vincent Rodin, Bernard Pottier, and Hiep Xuan Huynh supervised the research, revised, and finalized the manuscript. All authors had approved the final version.

ACKNOWLEDGMENT

This work is funded by a Brest Metropole, France. It is initiative related to Open Data for environment simulation. It is rooted in SAMES project (Ministère des Affaires Étrangères 2016 - 2018) grouping mainly researchers from University of Brest, Can Tho University, and now from places in Indonesia, West Indies, and Africa around software developments for environment modeling and simulation.

REFERENCES

[1] T. P. Truong, B. Pottier, and H. X. Huynh, "Cellular simulation for distributed sensing over complex terrains," *Sensors*, vol. 18, no. 7, 2018, 2323. https://doi.org/10.3390/s18072323

[2] H. L. Windom, "Contamination of the marine environment from land-based sources," *Marine Pollution Bulletin*, vol. 25, 1992, pp. 32–36.

[3] D. Li and D. Daler, "Ocean pollution from land-based sources: East China Sea, China," *Ambio*, vol. 33, no. 1, 2004, pp. 107–113.

[4] SHOM. (2005). Courants de Marée des Côtes de France (Manche/Atlantique). [Online]. Available: https://diffusion.shom.fr/media/wysiwyg/pdf/courants_2d_notice.pdf accessed on 01 April 2023.

[5] SHOM. (2005). Courants de Marée 2D. [Online]. Available: https://diffusion.shom.fr/marees/courants-de-maree/courants-2d.html accessed on 06 February 2024.

[6] M. M. Awad and T. Darwich, "Evaluating sea water quality in the coastal zone of north Lebanon using Telemac-2D TM," *Lebanese Science Journal*, vol. 10, no. 1, 2009, p. 35.

[7] Z. Li, J. Liu, C. Mei, W. Shao, H. Wang, and D. Yan, "Comparative analysis of building representations in TELEMAC-2D for flood inundation in idealized urban districts," *Water*, vol. 11, no. 9, 2019, 1840.

[8] A. Forster, J. Costi, W. C. Marques, A. G. Wormsbecher, and A. R. R. Bendo, "Application of the TELEMAC-2D model in the fluvial hydrodinamics simulation and reproduction of flood patterns," *In Defect and Diffusion Forum*, vol. 396, pp. 187–196, 2019.

[9] C. C. Lim, J. Yoon, K. Reynolds, L. B. Gerald, A. P. Ault, S. Heo, and M. L. Bell, "Harmful algal bloom aerosols and human health," *EBioMedicine*, vol. 93, 104604, 2023, pp. 1–23.

[10] P. Collar and G. Griffiths, "Single point current meters," in *Encyclopedia of Ocean Sciences*, John H. Steele (Ed.). Academic Press, Oxford, pp. 2796–2803, 2001. https://doi.org/10.1006/rwos.2001.0326

[11] B. H. Brumley, R. G. Cabrera, K. L. Deines, and E. A. Terray, "Performance of a broad-band acoustic Doppler current profiler," *IEEE Journal of Oceanic Engineering,* vol. 16, no. 4, 1991, pp. 402–407. https://doi.org/10.1109/48.90905

[12] K. Hickey, R. H. Khan, and J. Walsh, "Parametric estimation of ocean surface currents with HF radar," *IEEE Journal of Oceanic Engineering*, vol. 20, no. 2, 1995, pp. 139–144. https://doi.org/10.1109/48.376678

[13] I.-F. Jordi *et al*., "Remote sensing of ocean surface currents: A review of what is being observed and what is being assimilated," *Nonlinear Processes in Geophysics*, vol. 24, no. 4, 2017, pp. 613–643. https://doi.org/10.5194/npg-24-613-2017

[14] NOAA. What is current survey? [Online]. Available: https://oceanservice.noaa.gov/facts/current-survey.html accessed on 06 February 2024.

[15] A. F. Blumberg and G. L. Mellor, "A description of a three-dimensional coastal ocean circulation model," *Three-Dimensional Coastal Ocean Models*, vol. 4, 1987, pp. 1–16. https://doi.org/10.1029/CO004p0001

[16] T. Tawil, J. F. Charpentier, and M. Benbouzid, "Tidal energy site characterization for marine turbine optimal installation: Case of the Ouessant Island in France," *International Journal of Marine Energy*, vol. 18, 2017, pp. 57–64. https://doi.org/10.1016/j.ijome.2017.03.004

[17] European Environment Agency. (2013). EEA Reference Grid for France (1km). [Online]. Available: https://sdi.eea.europa.eu/catalogue/srv/api/records/ada072ce-a203-4e36- 87f4-cbd021ab6435f accessed on 01 April 2023.

[18] SHOM. (2022). Tides coefficient Roscoff (France). [Online]. Available: https://maree.shom.fr/habor/ROSCOFF/coeff?date=2022-07-01, accessed on 06 February 2024.

[19] Brittany Environment (OEB). Nitrates cours eau Bretons. [Online]. Available: https://brittany-environnement.fr/nitrates-cours-eau-bretons-datavisualisation accessed on 06 February 2024.

COLLEGE MATHSTIC
DOCTORAL BRETAGNE
BRETAGNE OCEANE

UBO
Université de Bretagne Occidentale

Titre : Données ouvertes et Simulation environnementale - Simulation environnementale et sociale sur des systémes de processus distribués basés sur un espace cellulaire irrégulier

Mot clés : Données Ouvertes, Espace Cellulaire Irrégulier, Divisions de Géographie, Simulation de Processus Distribués, Propagation Épidémique, Surveillance de la Pollution sur Les Côtes

Résumé : La combinaison d'automates cellulaires (CA) et de systèmes distribués offre un moyen simple de modéliser les problèmes environnementaux et sociaux en divisant les zones d'intérêt en segments spatiales discrètes pour le calcul parallèle. L'évolution de l'état de chaque segment est divisée en étapes temporelles discrètes. Les divisions géographiques en tant qu'espace cellulaire irrégulier permettent de tirer parti des données ouvertes pour alimenter les systèmes de simulation. Les données sont analysées pour en déduire les règles de transition apportant des influences distribuées dans un quartier. Une étude de cas de modélisation de la propagation épidémique basée sur les divisions administratives est présentée. Étant donné l'hypothèse que l'épidémie se propage aux personnes vivant dans le quartier, un système de simulation est généré en fonction des voisins adjacents avec des conditions initiales de collecte à partir du portail de données ouvertes du gouvernement. Une approche hybride est introduite avec la coopération entre les tuiles régulières et les espaces cellulaires irréguliers dans la modélisation des activités côtières. Une simulation environnementale est nécessaire pour représenter les caractéristiques du sol et de la mer qui se propagent dans les deux espaces. Ces caractéristiques sont très différentes en raison de la capacité et de la réaction du sol, et du comportement de la mer, en particulier les courants et les marées. Le problème des marées vertes est modélisé lorsque les nutriments sont présents en concentrations élevées et piégés par les courants de marée.

Title: Open Data and Environment Simulation - Environmental and social simulation on distributed process systems based on irregular cell space

Keywords: Open data, Irregular cell space, Geographic divisions, Distributed process simulation, Epidemic propagation, Monitor pollution on shores

Abstract: The combination of Cellular automata (CA) and distributed systems provide a simple way to model environmental and social issues by dividing the relevant areas into discrete spatial segments for parallel computation. The state evolution of each segment is divided into discrete time steps. Geographic divisions as irregular cell space give a chance to take advantage of Open Data in feeding the simulation systems. Data are analyzed to deduce the transition rules bringing distributed influences in a neighborhood. A case study of epidemic propagation modeling based on geographic divisions is presented. Given an assumption that the epidemic is spreading to people living in the neighborhood, a simulation system is generated based on adjacent neighbors with initial conditions collected from the government open data portal. A hybrid approach is introduced with the cooperation between regular tiles and irregular cell spaces in modeling shore activities. Environmental simulation is needed to represent ground and sea characteristics modeling spreading occurring on both spaces. These characteristics are very different due to soil capability and reaction, and sea behavior, in particular currents and tides. The problem of green tides is modeled when nutrients are presented in high concentrations and entrapped by tidal currents.