

ABSTRACT

ANALYZING, SIMULATING, AND VISUALIZING COMPLEX SOCIAL SYSTEMS

M. Baniukiewicz, M.S.
Department of Computer Science
Northern Illinois University, 2017
Dr. Philippe J. Giabbanelli, Director

Policy formulation and implementation is a multi-dimensional process, which requires a common platform to build communication between all sides involved. The growing availability of data along with the development of information and communication technology solutions (ICTs) supports this process by providing virtual platforms to design and evaluate policies. This thesis seeks to develop systems for policy-making with an emphasis on exploring and identifying the interacting causes that shape health. Our computational methods are primarily applied to the cause of obesity. In particular, we identify the relationships between fast-food outlets and schools at a national level, whereas it was previously done at a city-level. This thesis goes beyond the development of virtual platforms, by also contributing to newer approaches to analyze their output. Specifically, we develop interactive visualizations to help decision-makers in finding key patterns from large simulations of complex systems. Overall, this work has a few limitations. Despite the wealth and scale of data used in our study, it neither captures every single aspect that drives population health, nor does it track them with high temporal and spatial accuracy. Future work should explore the application of our model as a test platform for possible interventions, for instance through usability studies with policy-makers and an extended cost-benefit analysis of simulation results.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

DECEMBER 2017

**ANALYZING, SIMULATING, AND
VISUALIZING COMPLEX SOCIAL SYSTEMS**

BY

M. BANIUKIEWICZ
© 2017 M. Baniukiewicz

A THESIS SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

Thesis Director:
Dr. Philippe J. Giabbanelli

ACKNOWLEDGEMENTS

This work has been created thanks to the support and motivation of many individuals. I would like to express my gratitude to all of the people involved in the creation of this thesis. Most of all, I am highly indebted to my advisor, Prof. Philippe J. Giabbanelli, who guided me throughout my research and studies. I thank him for being a source of endless support, inspiration, and motivation. His advices and boundless knowledge allowed me to transform from a student to a professional researcher. I could not have imagined my education and exploration of carrier possibilities without him as my mentor.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Nicholas Karonis and Prof. Michael Papka for their interest in my research and precious suggestions regarding the topic of my research. Special thanks to Prof. Jose J. Padilla, whose insight helped my to put my research in the right perspective.

I would also like to thank my friends and colleagues at the Northern Illinois University, Eric Lavin, Venkata Sai Sriram Pillutla, and Marcus Nguyen, for many inspiring conversations, which made my stay and study more enjoyable.

This thesis was supported by the College of Liberal Arts & Sciences and the Department of Computer Science at Northern Illinois University. Research reported in this thesis was also supported by the Global Obesity Prevention Center (GOPC) at Johns Hopkins, and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) and the Office of The Director, National Institutes of Health (OD) under award number U54HD070725. The content is solely the responsibility of authors and does not necessarily represent the official views of the National Institute of Health. This work used resources of the Center for Research Computing and Data at Northern Illinois University.

DEDICATION

To my mother, my greatest teacher of compassion, love and fearlessness.

To my father, without whom none of my success would be possible.

To my sister for being a guiding light to follow.

*“Our greatest weakness lies in giving up.
The most certain way to succeed is always to
try just one more time.”*

— THOMAS A. EDISON

TABLE OF CONTENTS

	Page
List of Tables	x
List of Figures	xii
Chapter	
1 Introduction	1
1.1 Introduction	1
1.2 Objectives	3
1.2.1 Computer science and simulation research goals	3
1.2.2 Public health and obesity research goals	4
1.3 Methods	5
1.3.1 Overview	5
1.3.2 Aim 1: Analyzing the relationship between the fast-food environment and schools at the national scale	9
1.3.3 Aim 2: Anticipating the consequences of public health interventions on exposure to fast-food outlet	11
1.3.4 Aim 3: Visualizing the output of discrete simulations with replica- tions and multiple time steps	12
2 Background	15
2.1 The societal importance of obesity	15
2.2 Determinants of obesity and conceptual frameworks	17
2.2.1 Role of environmental factors in determining people’s health	20
2.2.2 Importance of socioeconomic status	23

Chapter	Page
2.2.3	Regulatory environment and new interventions 28
2.3	Measuring the access to fast-food outlets. 30
2.3.1	Overview 30
2.3.2	Principles of graph theory and spatial separation indices 33
2.3.3	Application of graph theory to clustering patterns 35
2.3.4	Beyond graph theory: performing nation-wide analyses 39
3	Understanding a nation: using large-scale network analysis to capture the fast-food landscape in England. 40
3.1	Introduction 41
3.2	Assembling a dataset. 44
3.2.1	Overview 44
3.2.2	Step 1: Dividing England into Local Authority Districts (LADs) . . . 47
3.2.3	Step 2: Finding the road segments within each LAD 49
3.2.4	Step 3: Assigning schools and fast-food outlets to road segment 51
3.2.5	Step 4: Identifying the Lower layer Super Output Area (LSOA) for each road segment 55
3.2.6	Step 5: Adding the deprivation level of each road segment via its LSOA 56
3.3	Analytical methods 56
3.3.1	Overview 56
3.3.2	Computing shortest-path distances. 57
3.3.3	Relating the presence of fast-food outlets to centralities 59
3.4	Results. 62
3.5	Discussion 66

Chapter	Page
4 Simulating a nation: a data science approach to developing a model of the fast-food environment in England	71
4.1 Introduction	72
4.2 Background	74
4.2.1 Theoretical models supporting clustering	76
4.3 Design of the Model and Experiments.	78
4.3.1 Model boundaries	78
4.3.2 Simulation rules	80
4.3.3 Strategy for model validation.	82
4.3.3.1 Running the simulation model	82
4.3.3.2 Strategy for model validation	84
4.4 Results.	85
4.5 Discussion	90
4.5.1 Principal Findings	90
4.5.2 Limitations	91
4.5.3 Future work	94
5 A novel visualization environment to support modelers in analyzing data generated by cellular automata	97
5.1 Introduction	98
5.2 Designing the visualization environment	101
5.3 Application to classical cellular automata models.	103
5.3.1 Epidemics	103
5.3.2 Sandpile.	105
5.3.3 Fire spread.	106
5.4 Feedback from modelers.	106

Chapter	Page
5.5 Discussion	108
6 Analyzing Spatio-Temporal and Multi-Run Data Produced by Simulations from Cellular Automata	111
6.1 Introduction	112
6.2 Background	114
6.2.1 Visualizing time-series of simulation outputs	115
6.2.2 Visualizing multi-run simulation data	118
6.3 Design of the visualization environment.	119
6.4 Design of the empirical evaluation	121
6.4.1 Objectives and overview	121
6.4.2 Detailed Study Design.	124
6.4.2.1 Pre-study questionnaire and overview video	125
6.4.2.2 Observational study and short interview	126
6.4.2.3 Post-study questionnaire	127
6.4.3 Participants and Apparatus	128
6.5 Results.	129
6.5.1 Overview	129
6.5.2 Quantitative analysis of the observational study	130
6.5.2.1 What affects the participants' experience?	130
6.5.2.2 How are participants affected by the visualization?	133
6.5.3 Qualitative analysis of the interview	136
6.5.4 Mixed methods analysis of the post-survey	138
6.6 Discussion	139
6.7 Conclusion	142

Chapter	Page
7 Navigating complex systems for policymaking using simple software tools	144
7.1 Introduction	145
7.2 Functionalities to navigate networks with a policy focus	150
7.2.1 Functionalities required	150
7.2.2 Functionalities supported by existing software	154
7.3 Proposed software: ActionableSystems	157
7.3.1 Overview	157
7.3.2 Design Principles	158
7.3.3 Implementation of the five functionalities	159
7.4 Demonstrations.	162
7.5 Discussion and conclusion	165
8 Conclusions and Perspectives.	166
8.1 Introduction	166
8.2 Achievements	167
8.2.1 Contributions to computer science and simulation research.	167
8.2.2 Contributions to public health and obesity research	168
8.3 Limitations.	170
8.4 Future work	173
8.4.1 A new visualization approach to multiple run and multiple replicated data	174
8.4.2 Tracking peoples' walking paths.	175
8.4.3 Modelling public policies	176
References.	178

LIST OF TABLES

Table	Page
2.1 Diseases related to obesity	16
2.2 Evidence linking the built environment with obesity.	24
2.3 Summary of SES and obesity papers.	27
2.4 Categories of policies and regulations	30
2.5 Accessibility indices' groups	33
2.6 Small scale studies with network analysis	36
2.7 Customer-side forces encouraging clustering of economic activities.	38
3.1 Network science studies of road networks.	43
3.2 Minimum and maximum values for each subdivision type in the UK.	47
3.3 Key features of previous studies of fast-food outlets in the UK.	48
3.4 Datasets combined for our study.	49
3.5 Hypothetical example of data produced during network analysis	53
3.6 Equations and fit (R^2) of the distributions of distances d for fast-food outlets	66
4.1 Correlation ranges.	88
4.2 Non-network national studies	93
6.1 Study components, process, and recent studies employing the same process .	122
6.2 Results of the NASA Task Load Index (NASA-TLX)	139
7.1 Current ICT software and the five key functionalities	155

8.1 Prioritization of public policies 172

LIST OF FIGURES

Figure	Page
1.1 Modeling & Simulation approaches	6
1.2 Model relationship of physical space and internal state representation	8
1.3 Approximation of a street network using a grid for cellular automata.	9
1.4 Key steps in the methodology of the first goal	10
1.5 Clock glyph to aggregate data within each simulation component	13
2.1 Prevalence of obesity and poverty in main English cities.	17
2.2 Ecological model of obesity	20
2.3 Rising popularity and clustering of fast-food outlets.	22
2.4 Conceptual framework of the local food environment	22
2.5 The Index of Multiple Deprivation	26
2.6 Hierarchy of methods to measure food access.	32
2.7 The three most common types of centralities.	35
3.1 Pre-processing overview	46
3.2 Example of roads encoding.	50
3.3 Three cases regarding the relationship between a road and an area.	51
3.4 Visualizations of largely disconnected road networks.	52
3.5 Distribution of the number of fast-food outlets and schools across LADs.	54
3.6 Approximation of k parameter used in centrality computation.	62
3.7 Examples of nodes with zero value of betweenness centrality	63

Figure	Page
3.8	Approximation of epsilon term 64
3.9	Results for distances distribution for both outlets and schools 65
3.10	Fit between the output and logarithmic curves, across levels of deprivation . . 66
3.11	Distribution of correlations between the density of outlets and centralities . . 67
4.1	Examples of centralities measures 73
4.2	Correlation cities average 86
4.3	Statistics of the greatest and lowest correlations 87
4.4	Comparison of correlation in bins of values 89
4.5	Comparison of results in respect to cities characteristics 96
5.1	The EXP V2 environment 99
5.2	Comparison of temporal glyphs for a dataset with continuous values 99
5.3	Modeling the spread of HIV within the body 100
5.4	Data aggregation in 4, 8, or 16 segments 101
5.5	Visualization of an epidemic in a CA 104
5.6	Visualization of a sandpile in a CA 105
5.7	Visualization of a forest fire in a CA 107
5.8	Varying the segment length to encode information 108
6.1	Example of the cellular automaton technique 114
6.2	Visualization of clock glyph combined with cellular automaton grid 116
6.3	Prototype visualization along with implemented tools 117
6.4	Aggregation method for clock glyph 120
6.5	Proposed environment based on compact visualization 121
6.6	Simple visualization environment using a slide 125

Figure	Page
6.7 Results for age, gender, and prior experience of participants	129
6.8 Contribution of independent factors and pairs of factors to the total variance .	132
6.9 Distribution of participant's confidence across tasks for both visualizations .	134
6.10 Distribution of participant's errors across tasks for both visualizations.	135
7.1 The Foresight Obesity Map	148
7.2 Extending a model to capture relationships between inputs.	149
7.3 Examples of loops and disjoint paths in obesity	151
7.4 Rendering of a System Dynamics model as hierarchical network	160
7.5 Using ActionableSystems on a System Dynamics model	161
7.6 Visualizations of the PHSA obesity and well-being map	164
7.7 Comparing two maps for obesity.	164
8.1 Overview of policy cycle and stakeholders. Adapted from [121]	167

CHAPTER 1

INTRODUCTION

1.1 Introduction

Many countries are faced with a relatively high prevalence of overweight and obesity among children. In the United Kingdom, based on the National Child Measurement Programme (NCMP) 2013-2014, one third of children aged 10-11 and over a fifth of those aged 4-5 were overweight or obese [185]. This has an array of possible long-term consequences which include high blood pressure, glucose intolerance, and adult obesity [51, 37]. The current policy landscape in the United Kingdom emphasizes the role of eating patterns in achieving a healthy weight. For example, the UK Department of Health stated that

“increasing physical activity is important but [...] eating and drinking less is key to weight loss.” [167]

Among the different types of food outlets, fast-foods have received particular attention since their products tend to be calorie dense, high in sugar, salt and fat, and low in fruit and vegetables. This project focuses on fast-foods around schools, and their impact on children. Studies in parts of the UK found that, over a 20 years period, fast-foods increased by 45% [153]. Given that food outlets tend to cluster around schools [11, 59, 66], longitudinal evidence from London has confirmed that schools have been exposed to an increasing number of fast-foods [215]. These outlets play a significant role in childrens’ nutrition: British secondary school children get more food from ‘fringe’ shops than from the school canteen [210]. In addition, even when there is a stay-on-site policy for lunch, the most popu-

lar time to buy food is after school [211]. This situation has led policymakers to increasingly advocate for the regulation of fast-foods as part of an overall strategy of obesity prevention in school neighborhoods. The 2011 National Institute for Health and Clinical Excellence (NICE) guidance recommended that local authorities regulate the number of fast-foods in specific areas, such as within walking distance of school [75]. The 2013 Academy of Medical Royal Colleges' report advocated to

“reduce the proximity of fast-food outlets to schools, colleges, leisure centers and other places where children gather.” [168]

The attention devoted to fast-foods around schools in the British policy landscape increased at the end of 2014. Two reports were published in October 2014 recommending a restriction of fast-foods around schools [119, 44]. Combined with growing local concerns, this has increased the pressure on local authorities to act through planning. In this context, meetings between researchers and planners were organized, such as the November 2014 Research and Policy Meeting where school environments and fast-foods were central items on the agenda [74]. Local authorities have stated planning as the tool of choice, owing to their views that improving nutritional quality

“is not an issue that will be satisfactorily resolved by voluntary improvement, education, advice or any other “easy” intervention. Without political will and a determination to limit the proliferation of takeaway food businesses we are unlikely to make any meaningful impact on the impact of poor diet on significant parts of the population.” (Peter Wright, Gateshead Council) [74]

As we enter a period where local authorities increasingly prepare and implement zoning restrictions for fast-foods, it is particularly time sensitive to inform the design of these policies. Indeed, these policies currently face two obstacles. First, they rest on limited

evidence on how fast-foods around schools' impact obesity. A recent systematic review found evidence for this association in 14 studies, while 2 studies showed the opposition association and 4 showed no association [244]. Second, these policies have historically differed widely in design as they can restrict takeaways in terms of clustering and concentration (e.g., maximum percentage of takeaways or minimum distance between them) or by setting buffers of varying sizes around schools [232]. There is thus a pressing need for a systems approach to find synergistic combinations of policy options to tackle obesity in an environment with high uncertainty. This thesis seeks to support policymakers in designing effective interventions by creating a simulation environment, serving as a virtual laboratory, in which policymakers can safely test different policies. The focus is on schools' exposure to fast-food restaurants, which has a significant impact on children's eating patterns. Moreover, we propose a new approach to visualize time-varying data, to allow a deeper understanding of simulated phenomenon. A more detailed description of these work goals is presented in next paragraph.

1.2 Objectives

1.2.1 Computer science and simulation research goals

In the context of policy making, analyzing data is crucial to understand the underlying structure of complex social phenomena, and leverage that understanding to design and/or evaluate interventions. Several parts of that process require technical innovations, thus making fundamental contributions to computer science. In particular, this thesis seeks to further the use of Modeling & Simulation (M&S) methods by making the following two contributions:

1. Increasing the scale of network-based simulations of the fast-food environment.

2. Visualizing the output of discrete simulations with replications and multiple time steps.

In the first contribution, we focus on developing the very first network-model for fast-food environments at a national scale. This is a novelty, as previous attempts were limited to single cities or regions. This objective is not technical straightforward: creating a large model cannot be achieved by simply telling previous models to ‘expand’, or letting run for longer. Indeed, new models have to be designed, supporting datasets must be identified and combined, and approximation algorithms may have to be calibrated in order to run simulations at a large scale (even when using resources such as high performance computing cluster).

In the second contribution, we are interested in visualizing the output of discrete simulations produced by Cellular Automata (CA). Since CA are run for multiple time steps (e.g. until a desired number of time steps or stabilization is achieved), and with many repeats (e.g. when the model is stochastic then repeats allow to assess the distribution of the output), a difficult is to visualize CA with multiple time steps and replications. While solutions have been identified previously for multivariate data, none has so far been used for CA. Our contribution for M&S research is thus on better navigating the output of simulation models using novel visualizations.

1.2.2 Public health and obesity research goals

The overarching contribution of this thesis to public health and obesity research is to support the identification and evaluation of new policies for obesity. Specifically, this contribution relies on two specific aims:

1. Anticipating the consequences of public health interventions (e.g., zoning) on exposure to fast-food outlets.

2. Providing researchers and policymakers with tools to design and evaluate interventions taking place in complex social systems [124, 71].

The prevalence of overweight and obesity among children in England has remained high for the past decade. Therefore, communities and the government seeks effective interventions which would limit the extent to which children are exposed to unhealthy foods. This thesis will investigate the spatial patterns of fast-food outlets, and explore factors (such as the presence of schools) that may be involved in determining the location of new outlets. The national scale of our analysis will support policymakers in designing national interventions, whereas the current policy landscapes consists of fragmented interventions decided by local policymakers with a paucity of evidence. This thesis will also provide practical tools that policymakers can use to take a systems science approach, both for the study of fast-food outlets and for policies in general.

1.3 Methods

1.3.1 Overview

As aforementioned, the key problem is that policymakers need better evidence to develop public health interventions. It is important to note that this problem belongs to the broader field of complex systems. Indeed, a growing literature has been devoted to viewing the design and evaluation of public policies as a complex system because of the many actors (e.g., communities, local governments, firms) and factors involved through non-linear interactions [124, 71]. In addition to being multi-actor or multi-sectoral, complex systems such as policy planning are also at the intersection of several fields of science [81]. Modeling will be the key approach in this thesis to support policymakers in designing interventions

in the complex system that drives, and results from, fast-food outlets and health. Modeling such systems faces several challenges. Complex systems often have an ability to change and adapt. Due to nonlinearity, the result of that adaptation can be difficult to predict. Studies of M&S applied to social problems use a broad range of methods, and each approach is able to capture only specific features of a problem [4, 188, 77]. For this reason, a variety of techniques will be used for our three main goals. Network simulation is a popular method in Modeling & Simulation (M&S) for simulating complex problems (Figure 1.1). It has been used in a wide variety of fields ranging from the analysis of population dynamics [234, 129, 15] to the representation of individual health behaviors [85, 88] or the management of complex ecological issues [182]. We note that there are also many other M&S methods (e.g., cellular automata, genetic algorithms) [231].

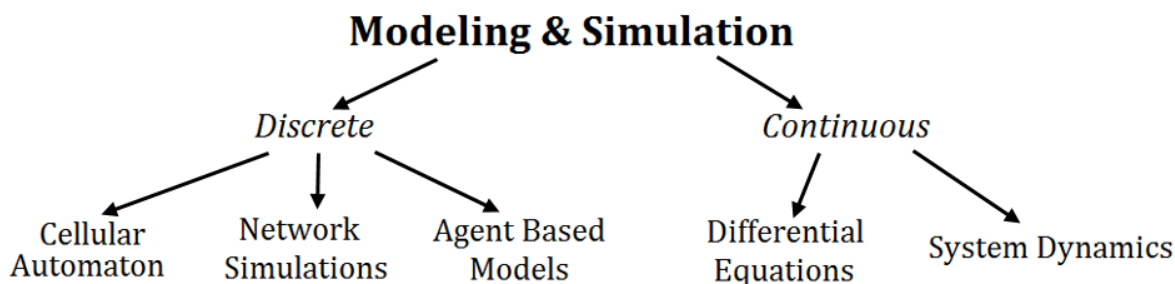


Figure 1.1: Hierarchy of commonly used Modeling & Simulation approaches.

The first and second goals will be operationalized by analyzing network data and developing a network model centered on the relationship between fast-food premises and schools in England. This will support policymakers in understanding how combined planning measures around schools affect the English food landscape across different levels of deprivation. The main steps will consist of:

- Assembling, pre-processing, and mining large heterogeneous datasets to understand the relationship between fast-food outlets and schools using graph measures (i.e. street networks) [39, 23].

- Implementing, calibrating, and validating the first national-level network model of the dynamics of regulating fast-food outlets, while monitoring effects on food exposure to children across levels of deprivation [96].

Building the network model is only half the journey: its results still have to be analyzed. Simulation models generate outputs, which may be used to control other systems or given to human experts. Using a simulation model is thus not only a matter of developing it and running simulations: it also requires the ability to analyze and visualize its output data. Generally speaking, data can be analyzed through a variety of techniques including machine learning and visualizations. While such techniques have been used for the data provided as input to the model, there has been relatively less research on how to apply them to the output [94]. That is: given a specific simulation model, how can its characteristics be leveraged when analyzing the data that it produces? Several interactive visualizations have been developed, which can be adapted to discrete simulations such as network simulations (Figure 1.1). This has been a particularly active area of research for simulations in engineering [134]. An example of a discrete modeling approach that commonly employs visualization is Cellular Automaton modeling, where the user is often presented with a grid of cells and a slider to navigate through time [152]. The third aim thus examines how to create interactive visualizations for discrete simulation models. That is, models consisting of a collection of elements that are updated over a period of discrete, fixed time steps, based on surrounding elements and system-wide rules. To avoid challenges related to the network visualizations (i.e. arbitrary arrangement of elements in space), a simpler approach will be adopted, the Cellular Automata (CA) [242] (i.e. fixed layout of square cells over a grid), as presented in Figure 1.2. In particular, we will consider a CA, which can be used to present complex and multidimensional output (e.g., from network simulation) in more accessible form. Network model can be approximated as a Cellular Automaton (Figure 1.3) via a process known as

tessellation. Although, as conversions exist between some types of model, visualizing cellular automata is only the beginning to visualize more complex models. This thesis will improve on existing methods for visualizing two-dimensional CA with square cells, via an innovative use of the 'temporal clock glyph' (known in data visualization) [242]. In addition, our proposed approach will be evaluated empirically, in terms of usefulness for modelers and performance. Modelers will include Dr Piper J. Jackson (Simon Fraser University) and Dr Vijay K. Mago (Lakehead University), with whom the senior thesis advisor has long-standing collaborations.

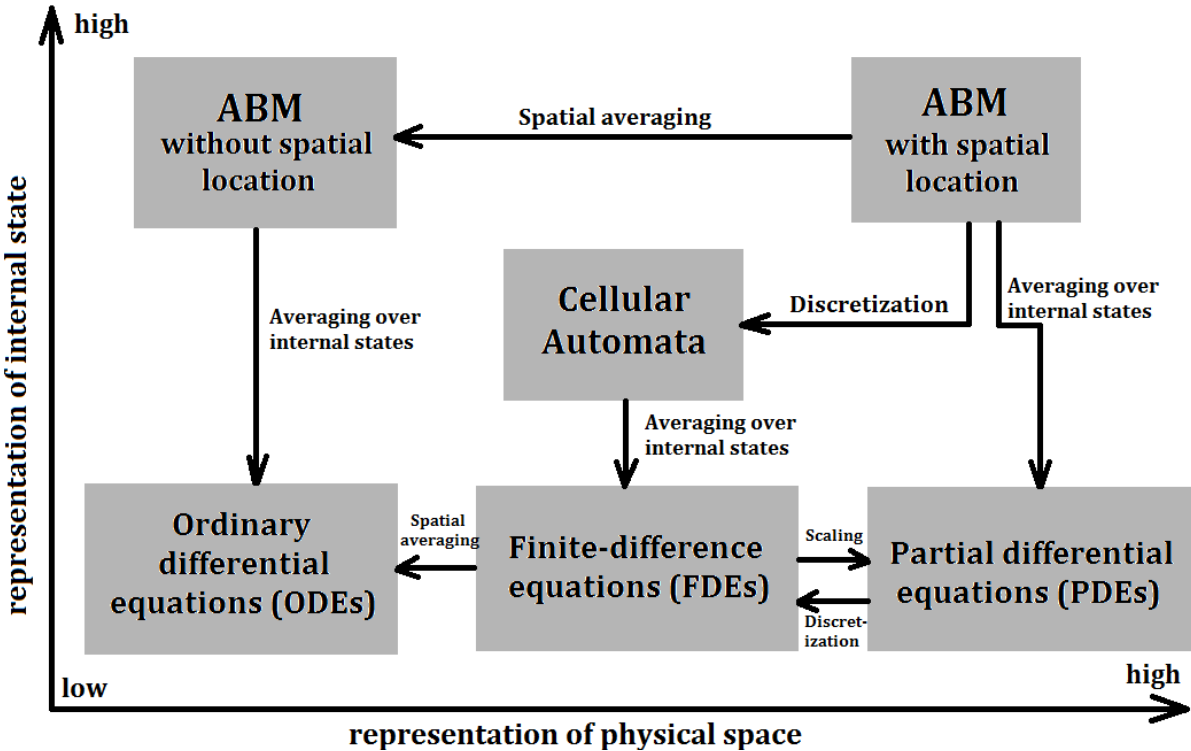


Figure 1.2: Model relationship with respect to physical space and internal state representation. Adapted from Deutsch & Dormann [62].

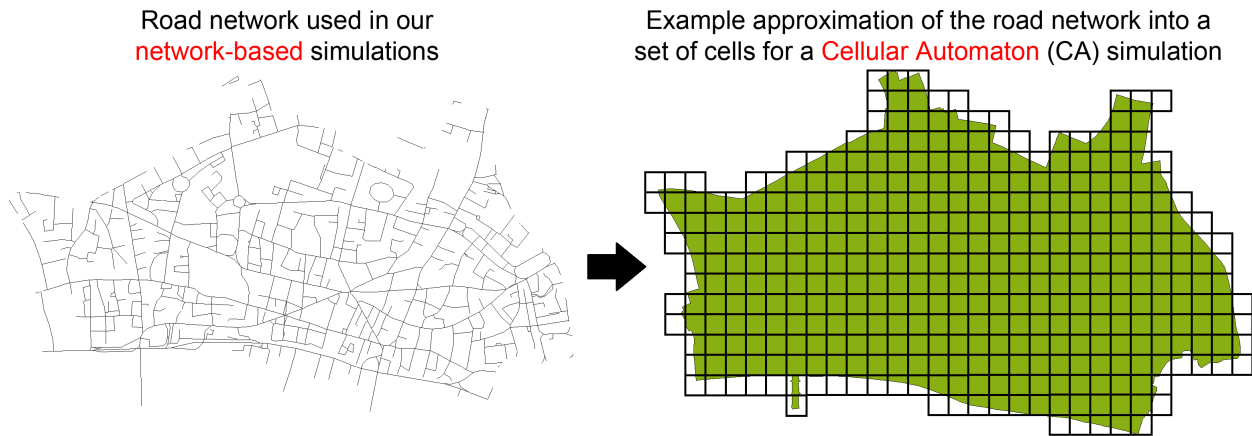


Figure 1.3: Approximation of a street network using a grid for cellular automata, as done in geography when applying kernel-based methods.

1.3.2 Aim 1: Analyzing the relationship between the fast-food environment and schools at the national scale

The aim is to create a system, which would contribute to better understanding of how combined zoning restrictions and policies affect distribution of fast-food premises at the national level. There are three main reasons supporting this study:

1. There is a high (and historically growing) prevalence of obesity among British children [185].
2. Schools have been exposed to an increasing number of fast-food restaurants. Children are highly exposed to unhealthy food on their way to school and back home [11, 59].
3. Current policies are attempting to regulate fast-food premises as part of an overall strategy of obesity prevention in school neighborhoods [197].

This project will start with collecting, pre-processing, and analyzing data. The spatial resolution will be the same as typically used when performing research on food geography

at the level of city, but scale to the whole of England. Consequently, handling large datasets through scientific computing will be a critical component to this project. For each area, we will collect: (i) the road network, (ii) the location of every fast-food outlets and schools, (iii) the level of deprivation as measured by the English Index of Deprivation [73], and (iv) the population predictions for the next 26 years. All of these datasets already exist, thus they do not require new data collection endeavors. The next step will focus on understanding the current geography of the fast-food landscape, using a very detailed geographical resolution over an unprecedented scale: the whole of England. Specifically, we will conduct network analysis. A street network is represented as a set of nodes (road intersections) connected by edges (roads). We will assume edges to be undirected, as we do not capture the directions of roads. Multiple edges will also be removed, as we do not capture the number of lanes but only whether there is a way to get from point A to point B. This undirected network will be analyzed structurally with respect to (i) the distance between fast-food outlets and schools, measured using the shortest path; and (ii) the correlation between the centrality of nodes and the number of fast-food outlets. We note that similar analysis (at the scale of a city) have previously been conducted [39]. This thesis’s approach is based on the widely used combination of Graph Theory and Spatial Separation Indices [39].

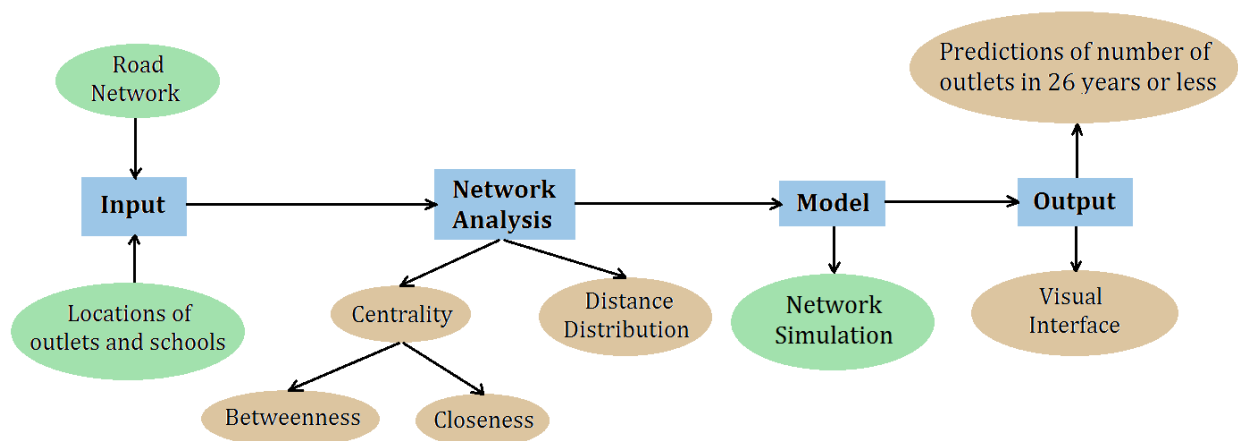


Figure 1.4: Key steps in the methodology of the first goal.

1.3.3 Aim 2: Anticipating the consequences of public health interventions on exposure to fast-food outlet

Once the analysis is completed, a large-scale network simulation will be developed. It will provide a virtual laboratory for testing policies and can thus be used to investigate what effects to expect without having to first conduct natural experiments [96]. Our model will focus on how new zoning policies may re-shape the fast-food landscape over years, thus impacting the exposure of children to unhealthy food options. That is, we will assume that school locations and roads are fixed, and that the normal rate at which fast-food outlets close is driven by extrinsic economical dynamics. For simulating changes in the fast-food landscape, results of network analysis (Aim 1) will be necessary to determine process of locating new restaurants. Thus, we will influence where new fast-food locations can be open, through policy. This will, in turn, modify the exposure that children have to fast-food outlets, with the understanding that a lower exposure would translate to a lower utilization, which in turns will impact health benefits. A health impact assessment study would be necessary to quantify these benefits exactly, and is beyond the scope of this thesis. The essential parts of the simulation are:

- **Input data:** Road network for given area, location of fast-food outlets and schools, deprivation level, and population predictions.
- **Output data:** For each step of simulation, the number of opened and closed outlets with their localization will be collected, as a proxy to schools' exposure to fast-food restaurants.

- **Timesteps:** One iteration of a simulation will correspond to one year. This scale allows to capture noticeable changes in the fast-food environment, while the process itself is not too slow.
- **Model boundaries:** The drivers of children exposure to fast-foods are highly complex and includes factors such as walking paths to and from schools, free meals plan available for children at schools, healthy options offered by restaurants, eating patterns of families etc. The scope of this project is limited to changes in population density in given area, impact of social-economic status on, and effectiveness of implemented zoning restrictions.

This model may be employed by policymakers to evaluate the effects of a new policy. Our time frame will be 26 years. A visual interface will be provided so policymakers can run a simulation with selected parameters in any area of England. The actual integration of our network simulation into the current workflow of policymakers, as well as improvements in usability and the evaluation of our method once policies are constructed are all important aspects. However, these steps can take several years given the complexity of the policy landscapes and the many actors involved. Consequently, they are beyond the scope of the proposed thesis, which will stop aim 2 after having developed a network simulation and before its integration in the policy workflow.

1.3.4 Aim 3: Visualizing the output of discrete simulations with replications and multiple time steps

Simulation such as network simulations generate outputs which can have many time steps, as well as many repeats (used to estimate the distribution of the output measure when there

is randomness in the model). Slider-based visualizations are commonly employed to navigate datasets, by displaying the state of all elements at each time step. However, this approach has several issues (e.g., change blindness) and may not scale well to many replications or many time steps. The aim is to support modelers by creating a new visualization framework for easy and effective exploration of the output of discrete simulation models. The goal will be reached by improving on the current slider-based approach by employing a clock glyph to aggregate data within each simulation component (Figure 1.5).

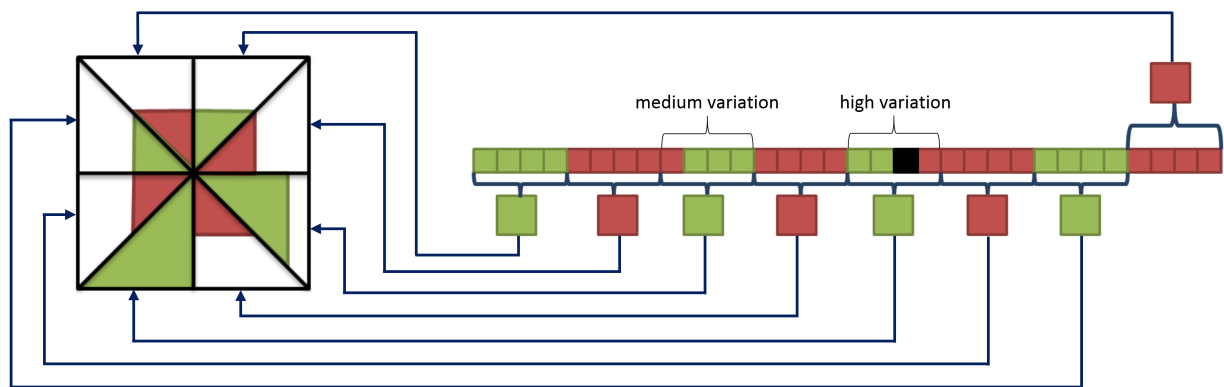


Figure 1.5: The 32 successive states of one simulation item can be aggregated by dividing them into chunks, and taking the majority state in each chunk (right). All chunks can be displayed simultaneously by wrapping them using a clock glyph, starting from the first chunk (at 'noon') and proceeding clock-wise. Research will explore the appropriate number of chunks, and how to encode additional simulation features such as variability across simulation runs (left).

This study will be conducted in following order:

1. **Prototyping** creating a prototype model, using the clock glyph design;
2. **Gathering feedback from modelers** - feedback on prototype will serve as the foundation for the design of the full visualization environment. As mentioned on p.4, there will be a minimum of two modelers (Dr Piper J. Jackson, Dr Vijay K. Mago). Associates in their respective research laboratories may also provide feedback;

3. **Updating model** model will be changed according to suggestions gathered from modelers in earlier step;
4. **Observational study** empirical evaluation of the performance and usefulness of our proposed final environment;

The new environment will have several components: three small ones, and one occupying most of the space (referred to as ‘the main’). The main visualization will represent the whole CA at once, where each cell is divided into equal segments whose color represents the main state during the corresponding part of the simulation. The three smaller components will include a flow diagram, which is automatically inferred from the simulation data and shows the different states (circles) and transitions (directed arrows). This new design proposed in this thesis will be examined in an empirical evaluation to:

- Assess whether important tasks for modelers can be performed more efficiently using the new environment than the one based on a slider; and
- Examine how performances are influenced by key simulation factors (length of the simulation, the number of replicas, and phenomenon being simulated [200, 10]).

CHAPTER 2

BACKGROUND

2.1 The societal importance of obesity

Food is an inseparable part of people's lives. It provides the energy and nutrients needed by the organism, thus playing a key role in weight dynamics and health. An excess in calories (i.e. a positive energy balance) is stored by the body as fat (either by adding to existing fat cells or recruiting new ones). Consuming more calories than needed for a person is an important contributor to overweight and obesity [232]. This condition is not limited to adults, but is also found in children and young people. This raises a question: are there risks associated with this condition? Studies on the co-morbidities of overweight and obesity have established links to many conditions, presented in Table 2.1 [63]. Given these detrimental consequences as well as the fact that individuals living with obesity often do so for an extended period of time, obesity is recognized as a chronic condition.

The prevalence of overweight and obesity has become high in many countries. Predictions show that this tendency will remain high. For example, linear models extrapolating from 1990-2008 data suggested that 51% of American adults will be obese by 2030 [72]. England is no exception to this 'obesity epidemic'. It has one of the highest rates of obesity in Europe and the developed world [232], with two thirds of adults, a quarter of 2-10 year old and one third of 11-15 year old already obese or overweight [187]. Figure 2.1 provides concrete examples of the prevalence of obesity on children in 9 English cities, and also illustrates the double burden of poverty and obesity [232, 61, 184].

Table 2.1: Diseases related to obesity [63, 187]

<i>Category</i>	<i>Diseases and conditions</i>
Metabolic Disease	Type 2 diabetes mellitus
Cardiovascular Disease	Coronary artery disease Congestive heart failure Ischemic heart disease Hypertension
Respiratory Disease	Asthma Obstructive sleep apnea Chronic obstructive pulmonary disease Obesity hypoventilation syndrome
Musculoskeletal Disease	Osteoarthritis Chronic back pain
Cancer	Colorectal cancer Kidney cancer Breast cancer Ovarian cancer Endometrial cancer
Cancer	Polycystic ovary syndrome Pulmonary embolism Stroke Gallbladder disease

In addition to the burden it creates for individuals' health and well-being, the high prevalence of obesity also has economic impacts on society. The annual cost of obesity in England was estimated to 27 billion pounds, with 352 million pounds of this amount due to social care and 13.3 million pounds spent on obesity medication only [187]. Other examples of costs include direct patient treatment for type 2 diabetes, which was estimated at 8.7 billion a year in 2010/2011, that is, approximately one tenth of NHS expenditure [111].

In section 2.2 we describe the complexity of obesity and introduce popular frameworks to conceptualize its determinants (i.e., factors that contribute to obesity). Specifically, we begin with the problem of capturing all determinants of obesity and their impacts (which are important for this section of the thesis). Then, Section 2.2.3 introduces the policy landscape for obesity in England. As this section of the thesis is primarily interested in zoning

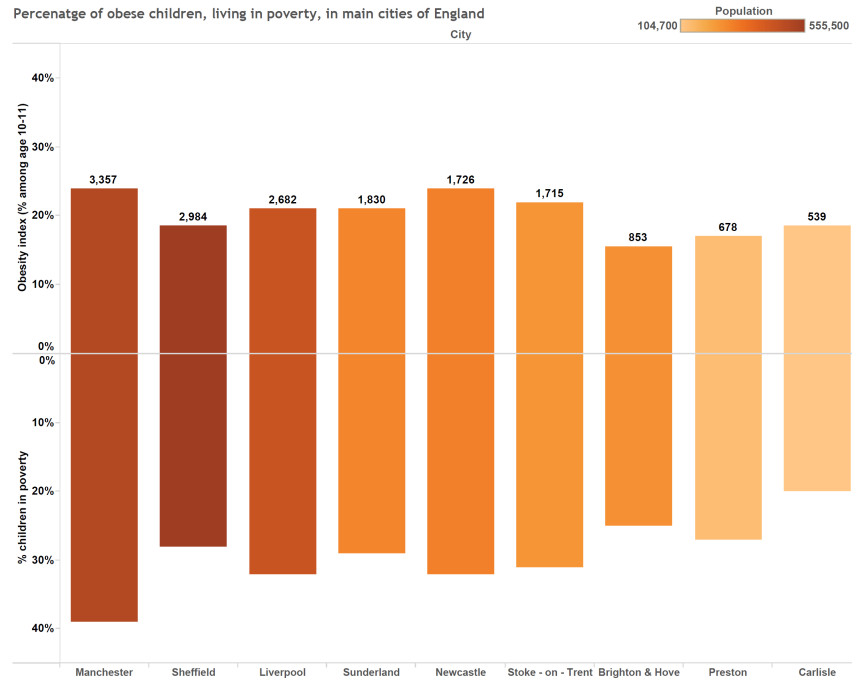


Figure 2.1: Prevalence of obesity (top bar) and poverty (bottom bar) for 9 English cities, with the population displayed as the bar’s hue. The top number shows how many children are living with obesity [232, 61, 184].

regulations, we briefly explain spatial distribution analysis in Section 2.3.2. In particular, this section covers both (i) the use of graph theory to perform small-scale analyses of street networks (e.g., at the city level), and (ii) nation-wide analysis without using the underlying street network.

2.2 Determinants of obesity and conceptual frameworks

Obesity research has undergone at least two significant conceptual shifts over the last decades, with a third shift currently underway.

First, there is now a broad understanding and agreement that obesity is a complex issue, deeply rooted within modern societies, which requires a multi-sectoral, multidisciplinary

perspective [187]. This view has not always been prevalent, as there was historically an emphasis on individual responsibility for health [158, 35], that is, on *proximal* (or ‘down-stream’) determinants of obesity rather than *distal* (or ‘up-stream’) ones. The historical reliance on proximal factors may be explained by a lack of studies showing clear correlations between distal factors and obesity, as well as a moral stance on blaming individuals rather than regulating. There is now conclusive evidence about the role of distal factors, and particularly socio-environmental factors [158, 219].

Second, obesity research has also shifted from seeing influences as *linear* to studying non-linear, system-wide dynamics. This gives rise to the concept of uncertainty and unpredictability [204]. Spontaneous decisions and events make models less predictive and much more sensitive to initial conditions, thus prompting some researchers to even view some behaviors as chaotic [191]. Reports have started to emerge that suggest using systems tools to deal with problems such as obesity [171].

Third, a shift was advocated from obesity to well-being [183]. This posits that the goal of interventions isn’t to change ‘silhouettes’ but rather to improve people’s lives in a comprehensive manner, which includes physical *and* mental well-being. This shift also emphasizes weight stigma and the negative consequences of attitudes that blame individuals, which continue to prevail despite scientific evidence on distal determinants.

Once a conceptual lens has been chosen to study the problem, specific public health interventions will be chosen. Those targeting food behaviors can take a broad range of forms, including price promotions/subsidies on healthier foods (e.g., fruits and vegetables), labeling less healthy options (e.g., those high in fat, sugar or salt), and limiting the proliferation of fast-food outlets through zoning [226, 187]. While there is also a range of tools to act on obesity through physical activity (e.g., improving cycling or walking), this section of the thesis is primarily concerned with food behaviors and focuses accordingly. Childhood obesity receives particular attention, as children are exposed to factors contributing to obesity early

on. Moreover, studies suggest that the likelihood of persistence of overweight into adulthood is moderate and effective interventions are essential to stop this process [212]. While some interventions may target the population as a whole (e.g., zoning), others may be designed specifically for them (e.g., limiting TV advertisement to children or school meal policies) or tuned. As an example of tuning, zoning interventions may take into account educational institutions, thus limiting the proliferation of fast-food outlets specifically in the vicinity to such institutions.

Interventions (and their evaluations) cannot be conceived in a vacuum. Social, environmental and political factors overlap and together shape humans' reaction to different stimuli [201]. In terms of food behavior, this translates to influences such as the environment, negative marketing, social impacts, or government actions [139, 141]. Each one can be divided into many, more specific factors, carrying positive and/or negative influences. The social-ecological framework allows to find a balance between proximal and distal contributors to individual health. This framework combines multiple dimensions (e.g. intrapersonal, interpersonal/network, community) and describes how they interact [157]. Figure 2.2 presents an example of social-ecological model for the problem of obesity [201]. While using such conceptual model can provide guidance to the design of interventions (Section 2.2.3), the very use of a model already has benefits as it allows individuals to situate their role in a complex system of interrelated factors. As a comprehensive examination of all determinants of human health is beyond the scope of this thesis, the next three sections address three categories of influence: environmental (2.2.1), socio-economical (2.2.2), and political (2.2.3).

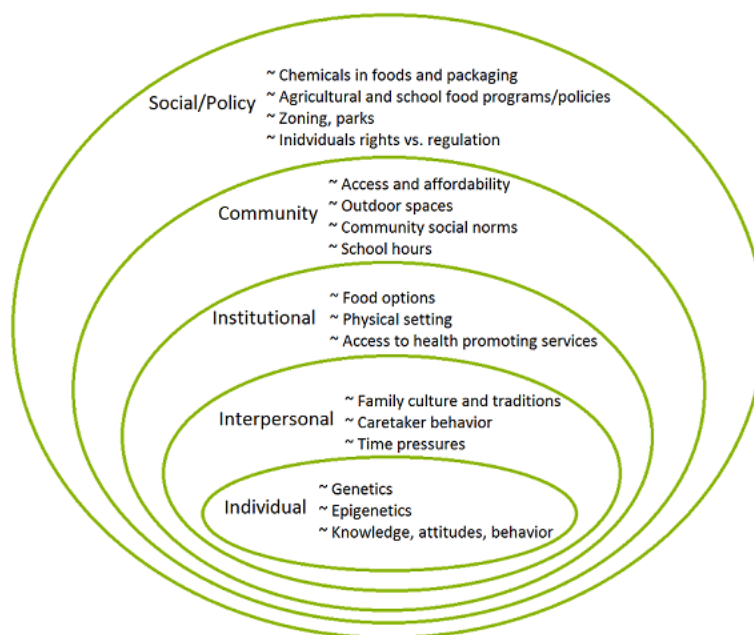


Figure 2.2: Ecological model of obesity from The Woman’s and Children’s Health Policy Center, adapted from [201]

2.2.1 Role of environmental factors in determining people’s health

The importance that environmental influences have on shaping peoples’ health might not be easy recognizable, since these sources of influence are both distal and multiform. Many studies have explored, and demonstrated, that this multitude of influences does act collectively on obesity. This was made possible in part due to methodological improvements in gathering reliable data on individual and contextual determinants of obesity [205]. That is, such demonstrations require precise operationalization of ‘environment’ and ‘obesity’, for example by focusing on characteristics of people’s neighborhood and measuring obesity through the Body Mass Index (BMI) [32]. Beyond providing proofs of influences, such

studies also contribute to identifying potential approaches to obesity prevention that target determinants at multiple levels [205].

More generally, the literature points at three conduits through which environmental influences are exerted [141]:

“The environment can be related to health through: (1) its physical design (the built environment); (2) the socio-cultural rules that govern these environments; and (3) the socio-economic status of these environments.”

More specific examples will be offered in the next section through the impact of deprivation on diet quality. This section of the thesis is primarily concerned with one environmental feature¹ which received increased attention in recent years: fast-food outlets. Mounting evidence points to the association of fast-foods with obesity [76, 122, 33] as well as broader detrimental effects on health. Reasons may include (i) their composition [166, 105] and (ii) the underestimation by customers of their energy consumption when eating in fast-food outlets [25]. As shown in Figure 2.3, fast-food outlets are a mounting concern in England, particularly as they are disproportionately found in proximity to schools [209, 59, 11]. Policymakers thus see school neighborhoods as a key place for health promotion [220]. A conceptual framework was recently proposed for urban planning policy and practice [161]. It shows the hypothesized pathways through which the built environment may influence immediate behavioral outcomes and even impacts long-term health effects. The framework accounts for food outlets availability and accessibility to neighborhood-level attributes (Figure 2.4), and it supports the view that there is a positive correlation between density of fast-food restaurants and obesity. This view is however debated, given the conflicting evidence.

¹The literature offers various interpretations of the term ‘environment’. According to Papas *et al.*, people perceive their surrounding as everything that is “external to individual” [172]. Others have referred to this term as a collection of factors such as availability of unhealthy food and “the walkability” of the environment – the term describing the level of environment’s adaptation to make individual’s surroundings more physical activity friendly and encourage people to walking [149].

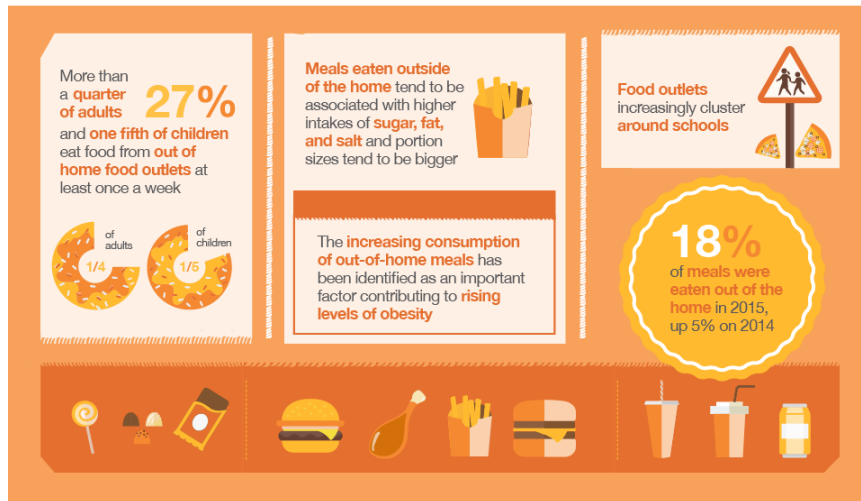


Figure 2.3: Importance of fast-food outlets for child nutrition based on a March 2017 report from Public Health England [186]. Additional data on the rising popularity of fast-foods can be found in [34]. Findings are similar in other countries such as the US, with respect to both rising popularity and clustering patterns. Studies in four large US cities found an increased mean count of neighborhood fast-food outlets over time [203], while spatial analyses found a clustering of fast-food outlets around schools [11].

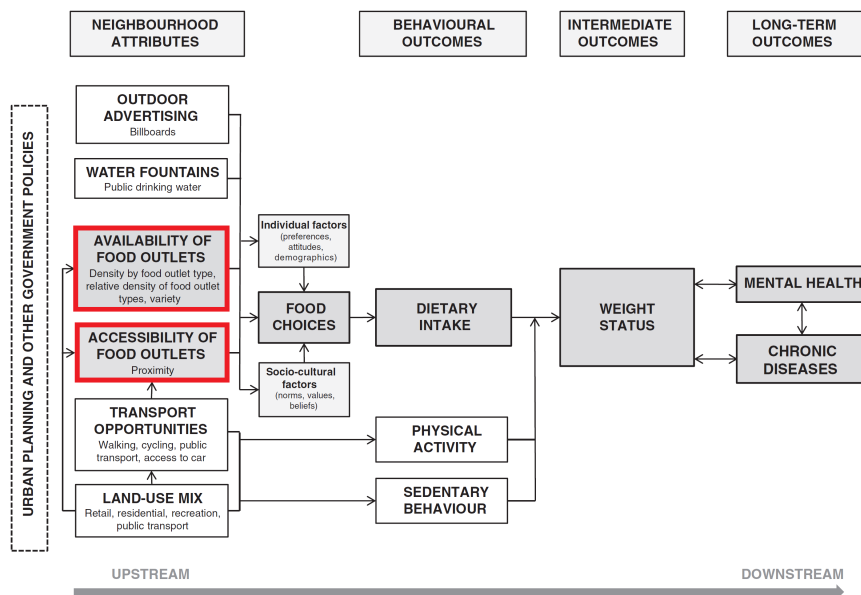


Figure 2.4: Conceptual framework of the local food environment and health to inform urban planning policy and practice. Adapted from Murphy et al. [161]

The evidence in North America indicates the food environment contributes to problems of energy imbalance and the ensuing weight gain in society. Most of the studies investigating the potential relationship between environmental features and likelihood of being obese are congruent with each other. The common suggestion is that easy access to unhealthy food options has a significant impact on peoples' weight statuses, however, not all research seems to support that hypothesis (Table 2.2).

While evidence does link the built environment with obesity, methods varied across studies from indirect methods (e.g. combination of survey data to estimate socio-economic status) to intermediate (e.g. use of telephone book, yellow pages or marketing databases) and direct ones (e.g. face to face interviews by trained investigators) [27]. Consequently, the current evidence base is not sufficient to clearly guide governmental environmental interventions into the modification of supposedly 'obesogenic' neighborhoods. However, these blurred findings and variety of convincing hypothesis do not discourage planning authorities from considering the built environment as one of the main contributor to the obesity epidemic [246, 47, 46, 48].

2.2.2 Importance of socioeconomic status

The ecological model of obesity (Figure 2.2) emphasized the multiple spheres of influences on health behaviors. The previous section discussed one such sphere (the built environment) as part of different components of environmental influences. Another component is the socio-economic status, which is a concept rather than an actual metric (as it can be operationalize in different ways e.g. depending on the stage of life) [141]. In England, it is measured using the Index of Multiple Deprivation (IMD) which weights different domains. Half of its value comes from income and employment deprivation (in equal shares) and a fourth from education/skills and health deprivation (in equal shares) [73].

Table 2.2: Review of studies investigating relation between BMI and environmental factors.

Study	Region	Findings
Positive correlation		
Obesity and the built environment [27]	Portland (Oregon)	This study showed that among 120 neighborhoods with higher fast-food restaurant density, there was a strong association among older adults visiting fast-food or buffet style restaurants and their likelihood of being obese. This study suggested that the availability of fast-food outlets in neighborhoods where participants live may increase their likelihood of being obese.
Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in Cambridgeshire, UK: population based, cross sectional study [33]	Cambridge, UK	Exposure to takeaway food outlets was positively and significantly associated with consumption of takeaway food. Moreover, associations between exposure to takeaway food outlets and body mass index were equally consistent.
Associations between Food Outlets around Schools and BMI among Primary Students in England: A Cross-Classified Multi-Level Analysis [245]	England	Cross-classified study of primary students BMI showed some evidence that fast-food outlet densities in a child's home neighborhood have an effect on BMI, particularly among girls.
Local food outlets, weight status, and dietary intake: associations in children aged 9–10 years [122]	Norfolk, England	Cross-sectional study investigated the correlation between weight status and dietary intake of over 1600 children with neighborhood food outlets. Results showed association between features of the built environment relating to food purchasing opportunities and weight status in children.
Fast food and obesity: a spatial analysis in a large United Kingdom population of children aged 13–15 [76]	United Kingdom	A spatial analysis made based on a data from a large United Kingdom population of children aged 13-15 revealed a 2% increase in body fat percentage associated with eating fast-foods.
No association		
Assessing the obesogenic environment of North East England [32]	North England	No significant relationship of environment with BMI, contradicting the suggestion that higher food outlets density correlate with higher BMIs.

An overview of the deprivation level across England reveals an unequal distribution across space² (Figure 2.5). This difference may in turn impact overweight and obesity through multiple pathways, including the exposure to fast-food outlets. Note that patterns of deprivation not only vary spatially but also across time, as can be seen through the different study waves.

Several studies have examined the distribution of fast-food outlets in relation to deprivation. In particular, a review of 12 studies showed a significant positive association between deprivation and the availability of fast-food outlets. None of them presented significant negative association and two studies indicated no significant association. However, only two of the studies were conducted in United Kingdom, while the bulk was performed in the US [77]. In addition, the studies were highly heterogeneous in terms of methodology³.

We assembled a sample of studies in Table 2.3, which shows that findings from the United Kingdom tend to be in agreement. Despite the studies different in geographical scale, they conclude that a greater level of neighborhood deprivation comes with an increased likelihood to (i) be exposed to fast-food outlets, and (ii) gain additional outlets over time. The one study that does not support this view is from Glasgow (Scotland). In short, in England specifically, the few studies conducted so far provide conclusive evidence that deprivation level is an important factor shaping the fast-food outlet environment⁴.

²The 2015 report by the Department for Communities and Local Government, based on 32,844 small areas or neighborhoods, detailed the distribution of deprivation across the country [73]. The summary showed that 61% of local authorities districts contain at least one of the most deprived neighborhoods in England. The report also listed the local authorities with the highest proportions of neighborhoods among most deprived in England (Middlesbrough, Knowsley, Kingston upon Hull, Liverpool, and Manchester). According to the executive summary, the 20 most deprived local authorities are largely the same as found for the 2010 Index, with few exceptions such as Tower Hamlets and Newham becoming relatively less deprived.

³Regression was used in few studies [153, 76, 32], while others employed repeated measures analysis of variance (RMANOVA) [153], one-way analysis of variance [55], cross-classified multi-level model [245], and correlation, chi-square and ANOVA [32].

⁴Several putative pathways have been put forth. One suggested that more affordable land-use costs attract restaurant owners more to deprived areas [55, 174], whereas another suggested that retailers may be catering to a greater customer demand or fast-foods in more deprived places [174]. The last hypothesis, which is of particular importance here, is that least deprived areas put higher constraints on small businesses; conversely, opening up a new fast-food outlet in more deprived places may be easier. [174]

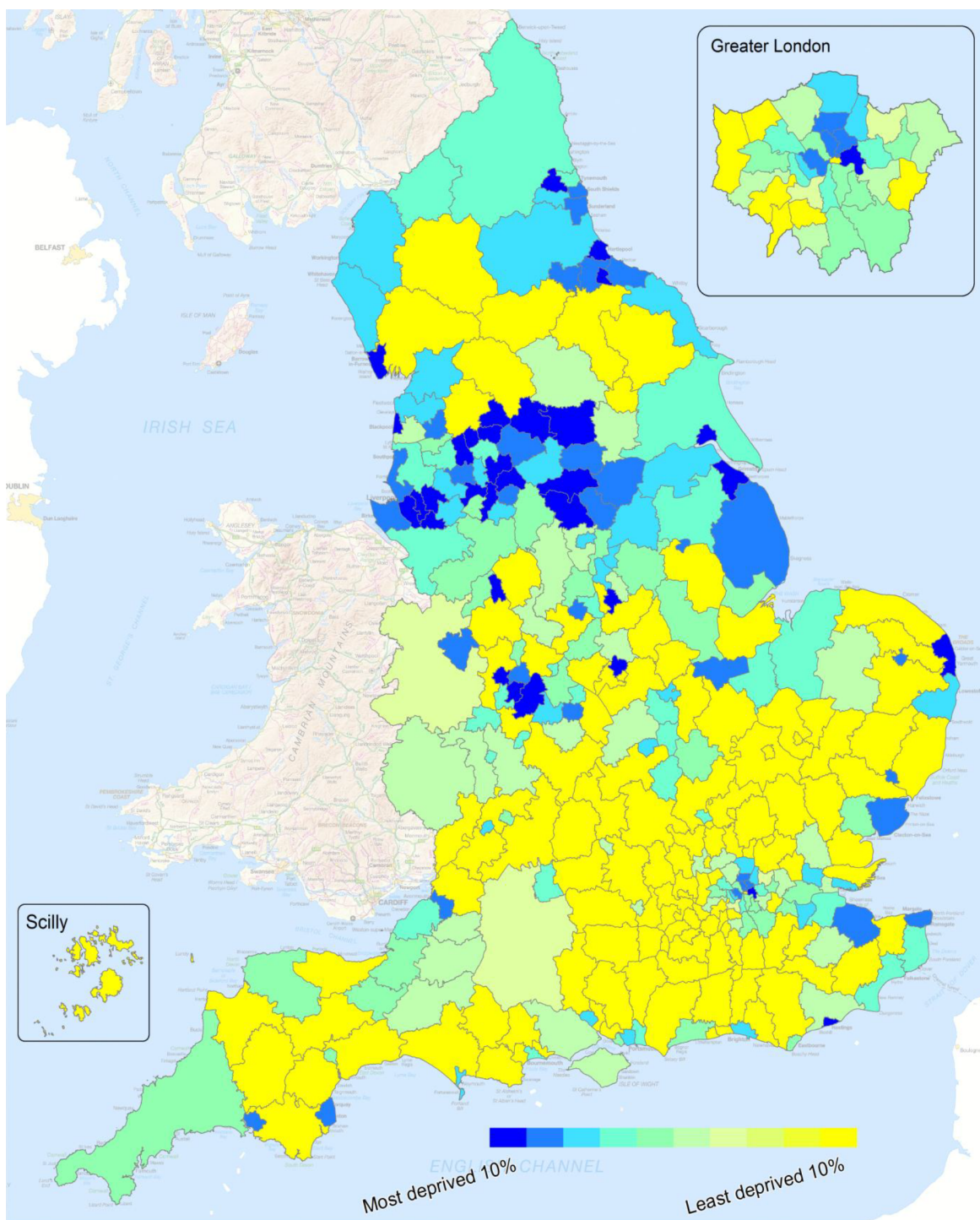


Figure 2.5: The Index of Multiple Deprivation 2015 among local authority districts based on proportion of their neighborhoods in the most deprived decile nationally. Adopted from The English Indices of Deprivation 2015 [73]

Table 2.3: Summary of papers that have been written on the distribution of fast-food outlets with respect to the socio-economic groups.

Ref.	Area covered	Year	Main results
[55]	Scotland and England	2005	For English SOAs, there is a statistically significant, positive linear association between quintile of area deprivation and mean number of outlets per 1000 residents as areas become more deprived, the mean number of outlets per 1000 people increases.
[174]	New Zealand	2007	Median travel distance to the nearest fast-food outlet varied by neighborhood deprivation, with travel distance being as least twice as far in the least-deprived compared to most-deprived areas. When distance is calculated to fast-food outlets from each school across the country, access similarly tends to be better around more-deprived schools.
[153]	County of Norfolk, UK	2015	The most deprived wards had the highest mean density of takeaway food outlets at every time point. In the most deprived tertile density in takeaway food outlet density increased by 2 outlets per 10,000 population over the time period from a mean of 4.6 in 1990 to 6.5 in 2008; a 43% increase.
[151]	The city of Glasgow, UK	2005	35% of out-of-home outlets, and nearly 50% of the fast-food chain restaurants, were located in Q2, the second least deprived quintile. The highest density per thousand population of each type of outlet, and for all combined, was in Q2 and the lowest density was in Q4.
[190]	City of Melbourne, Australia	2002	In the lowest income category (SES 4), there is approximately one outlet per 5641 head of population, compared to one outlet per 14,256 in the highest income category (SES 1). The second lowest income category (SES 3) has a relative density of outlets of 0.65, the next wealthiest income category (SES 2) has a relative density of 0.55 and this drops to 0.4 in the wealthiest income category (SES 1).
[26]	City of New Orleans, USA	2004	The regression equation for shopping areas with 1-mile buffer demonstrates that for every 10% increase in fast-food restaurant density, neighborhood median household income decreased by 4.8% and the percentage of black residents increased by 3.7%.
[154]	United Kingdom	2014	Differentials were observed in the consumption of the three food groups (fruit and vegetables, red and processed meat, oil fish) examined and non-milk extrinsic sugars by all the three socio-economic indicators. The socio-economic gradients identified for red and processed meat intake may be more pronounced between the consumption of lean, fresh meat compared with processed meat.
[160]	King County, WA, USA	2012	Mean diet cost was higher in persons with higher educational attainment and higher household income. Persons with the highest educational attainment reported diets that were 11.5% points higher in nutrient density than diets reported by the lowest group (mean of 95.7 versus 84.2 %).

2.2.3 Regulatory environment and new interventions

The complexity and prevalence of obesity has prompted policymakers to design many interventions. These can broadly be classified as prevention (e.g., avoiding weight gain in the first place) and treatment⁵. Techniques range from emphasizing healthy social norms (to cause behavior change) to create environments where the healthy choice is the easy choice [147].

Interventions differ in their ease of realization. Organizations including the UK Health Forum and the World Obesity Federation applied a Food Environment Policy Index (Food-EPI) to England. The aim was to influence government policy to create healthier food environments [225]. The policy areas rated as less well implemented (in comparison with best practices from other countries) included planning regulations and zoning to encourage healthy food outlets, and government-led systems based approach to improving the food environment. This suggests room for improvements in these two interventions, which will be at the core of the next chapter. Specifically, the experts prepared an action list to fill these policy gaps, in part by developing supplementary planning guidance and supporting local authorities. More specifically, a recommendation for policies aiming at improving the food environment in England is to monitor the food environment and strengthen planning laws to discourage less healthy food offers [225].

The role of local authorities for health in England has increased. It is well-established that they have a range of legislative and policy levers at their disposal [40]. On April 2013, that range widened as a reorganization of health services transferred responsibilities from

⁵The distinction between prevention and treatment is important to understand the policies investigated and simulated in this thesis. As policymaking is a professional field, there are many other specific terms which may occasionally appear. In the context of British policies, those include ‘hugs’ (direct incentives such as vouchers in return for healthy behavior), ‘shoves’ (measures restricting choices such as restricting takeaways for schools), and ‘smacks’ (bans e.g. on smoking in public places) [147].

the National Health Service (NHS) to local authorities [146]. Locally tailored strategies for obesity are thus at the forefront of the policy agenda. In recent years, local authorities have started to use the legal and planning systems to regulate the growth of fast-food outlets, including those near schools. One of the biggest initiatives is called “UK Healthy Cities Network” and refers to one of 20 networks accredited by World Health Organization (WHO). Several large cities in the United Kingdom are included in the network, such as Liverpool, Manchester, Newcastle or Sheffield [232]. Nonetheless, obstacles have emerged. First, authorities face issues in evaluating the possible effectiveness of these restrictions. This is particularly important when authorities are legally challenged. Second, the classification of food premises can create barriers. Most of the limitations developed by local authorities refers to new fast-food restaurants near schools. Before 2005, all hot food takeaways were classified as Class A3, which is much less restrictive than current Class A5. This means that, historically, hot food premises may have given planning permission under Class A3 if they have been in existence since before 2005, so the new rules do not apply to them [40]. Finally, other political challenges can make the situation even more complex. For example, Brexit (i.e. leaving the European Union) can change the price of food products importing from Europe, thus impacting the dynamics of the national market [142].

The documents controlling planning permissions for hot food takeaways are called Supplementary Planning Documents (SPDs). Through SPDs, local authorities are able to enact restrictions, within the limits of the National Planning Policy Framework. In their SPDs, fifteen councils have cited concerns relating obesity to hot food takeaways. Table 2.4 summarizes the main restrictions that authorities can create, through their SPDs [246, 47, 46, 48]. Note that while SPDs are a typical vehicle, authorities can also use other planning documents: five authorities used local plans and two relied on Development Plan Documents (DPDs) to control hot food takeaways [197].

Table 2.4: Categories of most popular common regulations implemented in English neighborhoods [246, 47, 46, 48].

<i>Type of regulation</i>	Description	Councils where it was adopted
<i>Exclusion zones to restrict hot food take-aways</i>	Restrict access to hot food take-aways, generally a 400-meter zone with variations applied to: schools, leisure centers, playing fields, youth facilities	Warrington Borough Council, City of Bradford MDC, Barking and Dagenham, Gateshead Council, Islington Council, Solihull Council, St Helen Council, London Borough of Newham
<i>Overconcentration/proliferation</i>	Limitations on the number of hot food takeaways in shopping centers and along high streets (in locations outside of exclusion zones)	Solihull Council, Gateshead Council, Barking and Dagenham
<i>Hours of operation</i>	Regulating the opening times, to limit the sale in late hours and lunch time, especially for restaurants close to schools	City of Bradford MDC, Warrington Borough Council
<i>Locations with high levels of obesity</i>	Planning permission will not be granted for hot food takeaways in wards where there is more than 10% of the year 6 pupils classified as obese	Gateshead Council

2.3 Measuring the access to fast-food outlets

2.3.1 Overview

Section 2.2.1 introduced the association of fast-foods (and thus outlets serving them) with obesity, while section 2.2.3 summarized the key tools that local planners can use to control the proliferation of fast-food outlets. In other words, ‘control’ seeks to make fast-food outlets less accessible. It is thus essential to understand (i) the extent to which fast-food outlets are *currently* accessible, and (ii) how this accessibility would be impacted by new control

measures. The next two chapters will use network analysis and agent-based modeling to address these two aspects in turn. To situate the contribution of these chapters, this section provides an overview of how accessibility has been computed so far. The next section is devoted to one specific approach to computing accessibility, that is, network analysis.

Studies exploring relationship between the food environment and diet have used a wide variety of methodologies to measure the degree of food access for study participants [39]:

1. **Geographic Information Systems (GIS)**, which is a general term for a *set* of measures, commonly operationalizes food access via store density (buffer distances) or proximity [43]. One challenge is to find appropriate and consistent criteria for defining geographic boundaries, which will be discussed in details in the next chapter [43];
2. **Store audits**, in which researchers visit stores and perform a manual assessment. While this approach requires intensive manpower (and thus has limited scale), it has been used on a few occasions. As exemplified by the Nutrition Environment Measures Study (NEMS) [98], such approach is most relevant when researchers are concerned with more than just physical accessibility. Their assessment may include price (affordability) and other study-specific metrics (e.g., product variety);
3. **Respondent-based perceived measures** ask the participants for the perceived availability and accessibility of food or food stores. As it is a perceived measure, it is less reliable than direct measures;
4. **Mixed** methods combine perceived and objective measures. Only a few studies have used them, for example by combining GIS methods with store audits. This allowed to assess the number of stores selling certain foods (audit) near the participants' homes (GIS).

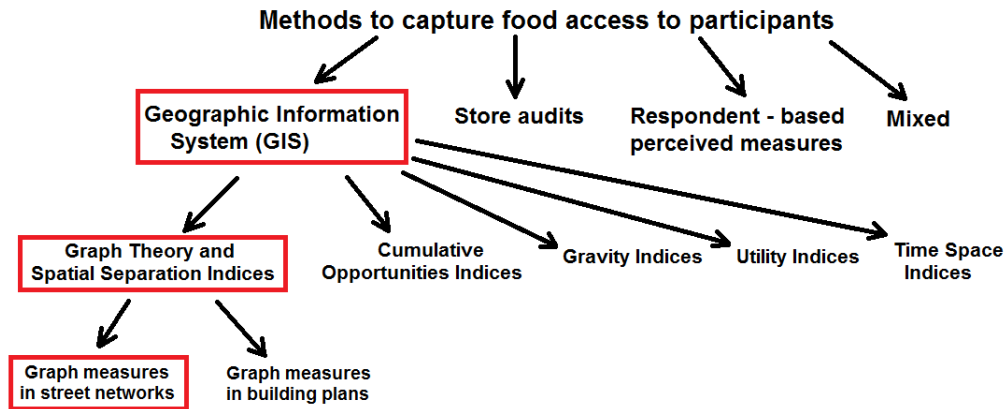


Figure 2.6: Hierarchy of methods used to measure the degree of food access, based on studies exploring the relationship between the food environment and diet. Adapted from Caspi *et al.* [39]

The majority of studies exploring the spatial distribution of food premises used a GIS approach [39, 156]. Trends suggest the use of GIS will even increase [43]. As the next chapter of this thesis belongs to this line of research, we now summarize its sub-divisions (Figure 2.6), culminating with the graph approach to street networks (detailed in the next section).

In our context, GIS are used to study how *urban form* (the geometry of streets, blocks, parcels, and buildings) relate to *land-use patterns* (spatial distribution of institutions and activities). *Activity measures* are used to relate these two notions⁶. Researchers have divided the various existing accessibility indices into five groups [23, 206], summarized in Table 2.5. The approach used in this thesis belongs to the first group, ‘Graph Theory and Spatial Separation Indices’.

⁶The notation of accessibility is somewhat similar to the notion of density, but what distinguishes them is the scale used to summarize features of built environment. Density focuses on number of features per unit of area, while accessibility analyzes the environment as seen from a specific location.

Table 2.5: Five groups of existing accessibility indices. Adopted from Bhat, Handy et al. [23]

<i>Accessibility group</i>	<i>Methodology</i>
<i>Graph Theory and Spatial Separation Indices</i>	Describes the spatial impedance factors that separate locations, without considering the nature of the activities separated. It typically measures accessibility from a particular location to either all other locations in the study area or to all other locations that fall within a certain distance threshold from the location of interest.
<i>Cumulative Opportunities Indices</i>	The Cumulative Opportunities Index differs from a graph theory measure for its inclusion of a destination type parameter. The index defines a travel time or distance threshold around a location and counts the number of destinations located within a distance threshold as the accessibility measure for the location.
<i>Gravity Indices</i>	A gravity type measure combines the attractiveness of the opportunities and the travel times required to reach them. The accessibility of a location can be quantified by calculating the time-distance relationship between the location and all possible destinations around that location.
<i>Utility Indices</i>	It is based on random utility theory, which assumes that the probability that an individual will patronize a particular destination depends on the relative utility of that choice compared to the utilities of all other possible choices.
<i>Time Space Indices</i>	The motivation behind this approach to accessibility is that individuals have only limited time periods to undertake actions. As travel times increase, the size of their prisms shrinks. This method allows for better evaluation of trip changing.

2.3.2 Principles of graph theory and spatial separation indices

Graph theory (also referred to as *network theory*) studies *graphs*, which are collections of *nodes* linked by *edges*. This approach will be formally developed in the next chapter. To investigate the relations between locations and the geometry of the environment [19], a *street network* is created. This is a specific type of graph, in which edges represent street segments, and nodes stand for street intersections. Nodes are usually the unit of analysis. An important strength of street network analysis is that it can measure access from each element of urban form (e.g. street segment) to every other element. That is, graph-based indices can

estimate a location's accessibility to all surrounding people or activities, regardless of their type. Specifically, a graph-based index can illustrate how some locations in the graph are closer, more 'between', or otherwise better accessible than others. To capture these varying aspects of nodes' accessibility within street network, Sergio Porta and colleagues created a methodology called Multiple Centrality Assessment (MCA). In short, different centrality measures from graph theory are used to perform a spatial analysis. Three common metrics employed in MCA are [178] (illustrated in Figure 2.7):

1. **Betweenness centrality.** It is the fraction of shortest paths between pairs of nodes in a network that pass through a given node [78]. A simple way to calculate it is to create a matrix of shortest paths between all node pairs. As will be discussed in the next chapter, this computation can be prohibitive for large networks and approximation algorithms are then used.
2. **Closeness centrality.** It is the inverse of the distance required, from one node, to reach all other nodes through shortest paths. The value is usually normalized by the number of nodes in the graph to make the measure comparable in different networks. The main aim of this approach is to describe how far each location is from all other locations.
3. **Straightness centrality** illustrates the extent to which the shortest paths from a node to all others resemble straight Euclidean paths. The straightness measure increases as the distance become larger. Higher values indicate a larger deviation from the shortest Euclidian paths.

While street network analysis provides a very detailed picture, it comes at the expense of computational requirements. For example, computing the shortest path between all pairs of origin-destination nodes at a country level would be both prohibitive (in resources) and

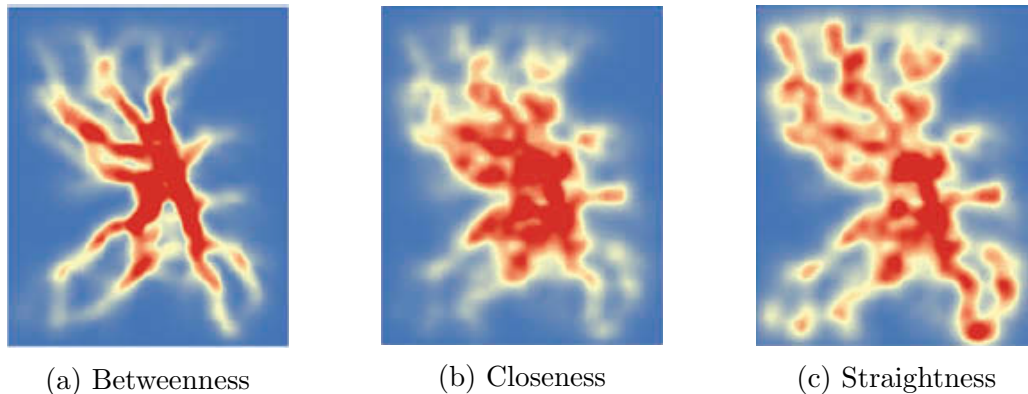


Figure 2.7: The three most common types of centralities.

uninformative (as accessibility is about what people can reasonably access rather than anything existing in the country). Consequently, it has so far been limited to the (relatively small) scale of one or a few cities⁷, as shown in selected studies (Table 2.6). In these studies, the three indices previously mentioned are the most popular. We also observe that the focus is usually on retail activities or land use.

2.3.3 Application of graph theory to clustering patterns

Owners and businessmen wish to ensure that customers can easily access their fast-food outlets. To maximize sales, it is thus important to find the right location. The literature on restaurant site selection shows that planning takes into account several aspects, including the physical network of roads that may draw in potential consumers, neighborhood demographics, and the presence of other businesses [206]. This last aspect is particularly important.

⁷A distinction was also made between a ‘global’ and a ‘local’ case [202]. In a global case, the centrality of a node depends on the entire network. In a local case, the network around a node is limited using a buffer area of radius d (e.g., 10km to match the average radius suburbs of Stockholm in [202]), and the node’s centrality is computed in this limited network. This distinction is most useful in a study using multiple spatial resolutions, but it can be misleading to compare studies. For example, a study may perform a ‘local’ case and still have a much bigger network (depending on d) than a study doing a ‘global’ case, whose network may be smaller to start with.

Table 2.6: Review of papers using the underlying street network at a much smaller scale (e.g., city).

Ref.	Cities	Network metrics	Phenomenon
[202]	Stockholm (Sweden)	Centrality metrics: closeness, betweenness, straightness	Land use, distinguished on types: built-up areas and green areas
[247]	Cardiff (Wales)	Centrality metrics: closeness, betweenness	Property prices
[170]	8 Israeli cities: Kfar Saba, Raanana, Bat-Yam, Beer Sheva, Ashdod, Modiin, Lod and Ramle	Degree of nodes, centrality metrics: closeness, betweenness	Spatial patterns of retail activities
[179]	Five neighborhoods in London (England): Telegraph Hill, Battersea, Brixton, Barnsbury, Dalston	Centrality metrics of betweenness	Dynamics of economic and social uplift (gentrification)
[241]	East Baton Rouge Parish of Louisiana (USA)	Centrality metrics: closeness, betweenness, straightness	Land use
[177]	Barcelona (Spain)	Centrality metrics: closeness, betweenness, straightness	Economic activities density
[178]	Bologna (Italy)	Centrality metrics: closeness, betweenness, straightness	Retail and services activities
[54]	Zhengzhou (China)	Centrality metrics: closeness, betweenness, straightness	Spatial patterns of POIs in city to understand urban land use and urban planning
[221]	Edinburgh (Scotland), Leicester (England), Sheffield (England), Oxford (England), Worcester (England), Lancaster (England), Catania (Italy), Barcelona (Italy), Bologna (Italy), Geneva (Switzerland)	Distribution of street length, angles formed between street intersections, relation between dead-end link length and the area they belong to, centrality metrics: closeness, betweenness, straightness	Geometric properties of the networks
[206]	Cambridge and Somerville (USA)	Reach, distance remoteness turns remoteness, intersection remoteness, betweenness centrality	Spatial location choices of retail eating establishments (accessibility, land use, urban form)

The decision must be on made on whether to stay close to competitors, or keep a distance from them. A small proximity between venues may lead to the creation of clusters, which can be defined as an excess or deficit of events across a geographic area measured relatively to a null hypothesis or the expected spatial pattern [117].

As discussed by Krider and Putler, research on clustering patterns of retail activities faces three challenges [138]. First, there are spatial differences in underlying demand density. It may thus be difficult to conclude whether stores attract/avoid one another when their locations are confounded by heterogeneous spatial demands in the population for their services. Second, the data collection and analysis process has historically been so labour intensive that significant simplifications were made [138]:

“at most a handful of store types are evaluated for their attraction-avoidance tendency in any one study. After data collection, the most commonly used analysis methods involve aggregate counts of stores in cells within a two dimensional grid, which is much less accurate, but much easier, than using measures derived from individual point locations.”

Note that this picture has gradually changed with the emerge of geocoding, and the increase in computational power to use GIS techniques. For example, the analysis in the next chapter will *not* aggregate stores in cells, but will instead opt for a more accurate analysis. Finally, there is no agreed upon measure of clustering (like there was none for accessibility). The wide range of methods to measure clustering include the average store counts in contiguous sites, nearest-neighborhood analysis, and the K-function (a density measure of the number of points within a Euclidean distance of another arbitrary fixed point) [248, 123, 175].

Clustering can be examined both from the businesses' viewpoints, and from a customer perspective. Several factors drive a customer's decision on *where* to buy goods, and many such factors can favor clustering (Table 2.7). For example, uncertain customers may prefer

Table 2.7: Overview of factors contributing to agglomeration of retail activities [21, 133].

Attraction forces that encourage homogeneous agglomeration	
Customer-side forces	Firm-side forces
<ul style="list-style-type: none"> • Comparison shopping motivated by consumer purchase • Customer taste heterogeneity • Customer expectations of lower prices • Increased customer awareness of homogeneous clusters • Shopping for entertainment • Transportation savings 	<ul style="list-style-type: none"> • Shared infrastructure • Localized resources • Efficiencies in firm resources utilization • Reduced location choice risk • Follower's traffic interceptor strategy • Demand externalities <ul style="list-style-type: none"> • Beneficial rent contracts for stores that constitute the primary motivation for customers' trips to the center

a cluster of shops (e.g., a mall) to easily contrast what they offer. The aggregation of outlets also influences customers' ability to remember locations: recall and awareness of clusters should be higher than of a single outlet, which in turns increases the trade area of the clustered outlets [138]. There are also arguments to justify avoidance rather than clustering. Spatial differentiation from the competition helps to avoid price rivalry and increase chances for monopoly rents. Similarly, separating increases market coverage, which has historically be shown to play a role when travel costs are important to customers [57] or if demand changes over time [65].

Several studies have analyzed the degree of clustering of fast-food outlets in proximity to schools, although none was conducted in the UK. A New Zealand study revealed a high concentration of fast-food outlets and convenience outlets within 1.5km of schools. Moreover, that clustering was greatest at distances less than 800 meters [59]. These findings are consistent with previous North American studies [11, 209, 223] that have also found a high degree of fast-food and/or convenience stores in close proximity to schools (e.g., a Chicago study found clustering of fast-food outlets within 1.5km from schools [11]).

2.3.4 Beyond graph theory: performing nation-wide analyses

As mentioned in the previous section, graph theory comes with scalability issues [19] and studies have typically used it on one or few cities. When the research goal is to obtain insight about the national scale, different methods have thus been used. It should be emphasized that the number of studies at a large scale is scarce, compared to studies at the local/city scale. Also, the type of ‘insight’ that studies aim for can be different, which is reflected once again in the variety of methodologies.

A 2016 Swedish study used the national register data to generate a comprehensive overview of the changing localization patterns of retail outlets. Concentration (or clustering) was operationalized using the Hoover Index and the redistribution rate [6]. Another large-scale study was conducted in the US and Canada. This research emphasized the use of detailed data on establishment counts by employment-size categories, industry, and geographic classification [115]. Another study on Mexico and Canada used Location Quotients (LQs) to determine patterns of economic development and industrial activity [176]. A 2010 study is particularly noteworthy for its large scale and unique method. Ghosh and colleagues collected data from 4 countries: US, Mexico, China, and India. They used the distribution of nighttime lights data (rather than a GIS-based method) to identify economic activities.

The main theme of these studies is not limited to the distribution of retail activities. The absence of retail activities is also important, particularly when it comes to “food deserts”, which are areas with limited access to food. In a study on food deserts, the authors collected data from 2,275 non-metropolitan counties in the continental U.S., and examined inequalities in food access for non-metropolitan residents [24].

In short, these studies show that national assessments can be done, but *none* has employed a street network approach.

CHAPTER 3

UNDERSTANDING A NATION: USING LARGE-SCALE NETWORK ANALYSIS TO CAPTURE THE FAST-FOOD LANDSCAPE IN ENGLAND

In the previous chapter, we discussed how access to fast-food outlets has been measured. In particular, we showed that graph measures in street networks have been limited to small areas such as cities, whereas studies in larger areas (e.g. national scale) used methods that did not take into account the street networks and thus approximated distances. In this chapter, we present the first study to use the accuracy of street networks at a national scale. Specifically, we compute distances between fast-food outlets and schools in England. This required linking several datasets, and performing computations over a high-performance cluster. The chapter thus describes the datasets, their linkages, and our implementation. Analyses from this chapter are used in the next chapter, which develops simulations.

My contributions consisted of (i) obtaining additional datasets on geographical boundaries and deprivation, (ii) cleaning and linking all datasets, (iii) implementing and running network analyses over a high-performance cluster (HPC), and (iv) visualizing the results. This study was funded by Johns Hopkins University's Global Obesity Prevention Center (GOPC) through a grant to Dr PJ Giabbanelli, who also obtained the location of all fast-food outlets and schools. Computations on the HPC were supported by the Center for Research Computing & Data at Northern Illinois University, with assistance from J Winans on writing scripts.

3.1 Introduction

Road networks are one of the oldest forms of human-made infrastructure networks, preceding power and telecommunication networks. Before network science became a popular approach, geographers devoted several books analyzing road networks, including *Network Analysis in Geography* from the late 1960s [101] and the seminal *The Seminal Logic of Space* in 1984 [112]. While some modern day cities may appear to have a grid-like pattern of roads, many road networks do not result from a central planning process but instead emerge over time as the result of an organic densification/exploration process [18] thus creating structures far more complex than square grids. While local geographical and socio-economical differences may have influenced the process, road networks have nonetheless be found to have commonalities across cities and countries. For example, Buhl and colleagues found similar average degrees [31] while Cardillo *et al.* reported a fractal dimension (per the box-counting method) in the narrow 1.7-2.00 range [38]. For a summary of these commonalities, and a contextualization of findings among other spatial networks, we refer the reader to the review by Barthelemy [17].

Network science has been particularly interested in relating a network's *structure* to its *function*. While there is a myriad of metrics, road networks are often analyzed with respect to betweenness centrality (since they are infrastructure networks and this approximates traffic between all pairs of nodes) and closeness centrality (as a proxy to access). These metrics have been related to various phenomena, such as the presence of specific retail activities. In this paper, we focus on using the structure of road networks to understand the presence of fast-food outlets. Analyses of road networks in the recent years have all been at the city level (Table 3.1). Geographers and economists have also analyzed retail activities at a national level, but without using network-based metrics (e.g. shortest-paths calculations

or centrality). For example, the geographical distribution of retail outlets was investigated at the scale of Sweden using buffers [6], while food deserts were examined in rural U.S. counties using county-level data [24], and economic activities across countries were assessed by imposing a square grid [83]. In this paper, we present the first large-scale analysis of retail activities (focusing on fast-food outlets) using network methods at the national scale.

Our scale is the whole of England. Studying the geography of fast-food outlets at a detailed level (e.g. using network metrics such as shortest-paths distances between outlets) at the scale of England is primarily motivated by the current public health context. In the United Kingdom (UK), based on the National Child Measurement Programme (NCMP) 2013-2014, one third of children aged 10-11 and over a fifth of those aged 4-5 were overweight or obese [107]. The current policy landscape in the UK emphasizes the role of eating patterns in achieving a healthy weight, and fast-food outlets have received particular attention. These outlets play a significant role in children's nutrition: British secondary school children get more food from 'fringe' shops than from the school canteen [210]. In addition, even when there is a stay-on-site policy for lunch, the most popular time to buy food is after school [211]. This situation has led policymakers to increasingly advocate for the regulation of fast-foods as part of an overall strategy of obesity prevention in school neighbourhoods. Between 2011 and 2014, four reports have called for a restriction of fast-food outlets around schools [75, 168, 1, 44]. However, policies have so far differed widely in design as they can restrict fast-food outlets in terms of (i) clustering (e.g. minimum distance between them) or (ii) respectively to schools (e.g. with a minimum distance from schools) [162]. In addition, the impact of fast-foods on obesity and food consumption varies over space, and particularly depending on the deprivation of the area [76]. In this context, The principal contribution of the present work is to take a big data approach to propose the first investigation of fast-food activities based on road networks at a nation's rather than city's level. Specifically, we conduct large-scale network analyses to:

Table 3.1: Network science studies investigating various structures in road networks (sorted by year)

Ref.	Cities	Network Metrics	Phenomena
[178]	Bologna (Italy)	Centrality (closeness, betweenness, straightness)	Retail and service activities
[206]	Cambridge and Somerville, MA (USA)	Number of destinations available in a given radius (i.e. reach) and cumulative number of meters/turns/intersections to reach them using shortest paths; centrality (betweenness)	Retail activities, urban form, and land use
[241]	East Baton Rouge (USA)	Centrality (closeness, betweenness, straightness)	Land use
[177]	Barcelona (Spain)	Centrality (closeness, betweenness, straightness)	Retail activity
[221]	Edinburgh (Scotland), Leicester (England), Sheffield (England), Oxford (England), Worcester (England), Lancaster (England), Catania (Italy), Barcelona (Spain), Bologna (Italy), Geneva (Switzerland)	Centrality (closeness, betweenness, straightness, accessibility), street lengths, intersection angles, areas	Geometric properties
[236]	Neighborhoods of London (England)	Centrality (betweenness)	Gentrification
[202]	Stockholm (Sweden)	Centrality (closeness, betweenness, straightness)	Land use (built-up areas vs green areas)
[54]	Zhengzhou (China)	Centrality (closeness, betweenness, straightness)	Land use (Points Of Interests)
[247]	Cardiff (Wales)	Centrality (closeness, betweenness)	Property prices
[170]	Old cities (Kfar Saba, Raanana, Bat-Yam), new cities (Beer Sheva, Ashdod, Modiin), and hybrid cities (Lod, Ramle) in Israel	Degree, centrality (closeness, betweenness)	Retail activity

- (1) contribute to the evidence base for coordinated regulation at the level of England by analyzing distances (i) between fast-food outlets and (ii) between fast-food outlets and schools, across deprivation levels.
- (2) we investigate the relationship between centrality and the presence of fast-food outlets nation-wide, thus extending the scope of many previous studies employing network centrality mostly at the city-scale.

The remainder of this paper is divided into four sections. In section 3.2, we summarize the different geographical layers in England and the associated datasets used for this study. In particular, we contextualize these datasets with respect to previous studies of food outlets in England, and we explain the different steps to pre-process the datasets. Pre-processing includes assigning fast-food outlets and schools to roads, building the road network, and identifying the deprivation level of each road segment. Our analysis methods (including centrality metrics and their computation) are summarized in section 3.3, with results provided in section 3.4. Results are discussed in section 3.5 in terms of their contribution to the evidence-base for public health in England, and regarding the potential of using large-scale analyses to inform regulations going forward.

3.2 Assembling a dataset

3.2.1 Overview

Our objective was to assemble a data that includes the location of fast-food outlets and schools on the road network, and also provides the level of deprivation. This objective was accomplished in five steps, each involving the use of another dataset. We used a top-down

process (Figure 3.1), starting with the whole of England (thus excluding Wales, Scotland, and Northern Ireland). Rather than obtaining a *single* massive network which would be computationally prohibitive to analyze, we started by dividing England into coarse units. Step 1 divides England into Local Authority Districts (LADs), specified in the 2016 boundary line dataset. In step 2, we added in the 2016 Ordnance Survey (OS) Open Roads containing 3,396,694 roads. Specifically, we found the roads that resided (either entirely or partially) within each LAD. In step 3, we retrieved the location of fast-food outlets and schools from the Points of Interest data (PointX Database Right/Copyright 2016) obtained in January 2016. This dataset aggregates over 150 databases (in the ‘eating and drinking’ category) and has an accuracy ranging from 81% to 100% [76]. Locations for fast-food outlets were added to the street networks. At that stage, we had divided England into 327 LADs, each containing a road network, with fast-food outlets and schools assigned to each road segment. Research has shown that the impact of fast-foods on obesity and food consumption varies over space, and particularly depending on the deprivation of the area [76]. Therefore, we also had to track the deprivation score, Index of Multiple Deprivation (IMD), which takes into account employment, living environment, crime, health, education, income and housing [163]. Tracking this score took two additional steps, because it was provided in datasets using different geographical units.

Whereas LADs are designed based on local governance, most statistics are available in census data, which uses different spatial units. England can be divided using three levels of spatial units, from largest to smallest: Middle layer Super Output Areas (MSOAs), Lower layer Super Output Areas (LSOAs), and Output Areas (OAs). The minimum and maximum number of inhabitants in each of these 3 possible subdivisions is summarized in Table 3.2. In order to most accurately track deprivation levels, we used the most detailed level at which this information is available: LSOAs. It should be noted that LSOAs is a spatial resolution often used in studies of food geography focusing on a *single* city, such as Bristol [76], parts of



Figure 3.1: Our five steps process to combine five (large) datasets into one, specifying the location of fast-food outlets and schools within road as well as the level of deprivation. *The online version allows zooming to see detailed locations and deprivation levels within this sample LAD (Adur).*

Berkshire [245] or the North East of England [32]. However, using them in a *national* study (together with the whole road network) and conducting a detailed *network* analysis are two of the hallmarks of the present study, in contrast with previous work (Table 3.3).

In step 4, we used the latest (2011) census division of England into 34,753 LSOAs (which also included Wales). We removed Wales, and identified the LSOAs to which each road segment belonged. Finally, step 5 cross-referenced the LSOAs with the 2015 Indices of Multiple Deprivation dataset: since we knew the LSOA for each road, and the deprivation for each LSOA, we were able to assign a deprivation level for each road. The summary of datasets involved is provided in Table 3.4.

This five step process required extensive data pre-processing, not only because of the sheer volume of information, but because of numerous challenges in combining the datasets (e.g., missing values, mismatch in geographical units). The operations involved in each

Table 3.2: Minimum and maximum values for each subdivision type in the UK.

Subdivision	Min value	Max value
OA	100 residents or 40 households	625 residents or 250 households
LSOA	1000 residents or 400 households	3,000 residents or 1,200 households
MSOA	5000 residents or 2000 households	15,000 residents or 6,000 households

step are now detailed, each within a dedicated sub-section. All of the scripts necessary to combine and pre-process the data are available within the ‘Pre-processing’ folder at <https://osf.io/gn3f2/>. Note that many of our spatial queries (e.g., to assess whether a road ‘fits’ within a LAD) require the open source library `GeoTools` for Java. As we do not own the data, links within Table 3.4 track data provenance.

3.2.2 Step 1: Dividing England into Local Authority Districts (LADs)

This straightforward step starts our process by using the 326 shape files defining LADs, from the boundary-line dataset. Note that each result is not only a geometry defining the boundaries of the LAD, but a spatial object due to the use of coordinates. It also has a name, which later steps use to double-check linking across datasets.

Table 3.3: Key features of previous studies of fast-food outlets in the UK.

Ref.	Area	Data	Analysis
[55]	Norfolk county	The location of food-related outlets was extracted from the Yellow Pages directory issued in six years (1990, 1992, 1996, 2000, 2003, 2008). Locations were overlaid onto the 2001 electoral ward boundaries for Norfolk (n=205)	Repeated measures analysis of variance (RMANOVA) / multiple logistic regression model
[153]	England and Scotland	The location of McDonald's restaurants (n=942) were obtained from the Yellow Pages directory and overlaid on 38,987 small areas: 6505 'Data zones' in Scotland and 32,482 Super Output Areas in England	One-way analysis of variance
[76]	Avon county	The location of outlets was extracted from the Ordnance Survey Points of Interest in Avon county	Geographically weighted regression
[245]	Berkshire county	The location of outlets was obtained from six local councils, with analyses at the LSOA level	Cross-classified multi-level model with Markov chain Monte Carlo methods
[32]	North East of England	The location of food-related outlets was extracted from the Yellow Pages directory, with analyses at the LSOA level	Correlation analysis, logistic multinomial regression, ANOVA

Table 3.4: Datasets combined for our study.

Step	Dataset	Year	Characteristics
kk	Boundary-Line ¹ products.html	2016	Shape files of polling districts, county and district regions, wards, etc. 1.41Gb in total.
2	Ordnance Survey (OS) Open Roads ²	2016	3,396,694 roads
3	Points of Interest ³	Jan. 2016	Location of 39,374 fast-food outlets and 25,755 schools.
4	Lower Layer Super Output Area boundaries ⁴	2011	Shape file of 34,753 geometries defining LSOAs
5	Indices of Multiple Deprivation ⁵	2015	32,845 rows of IMD score and contributing elements (e.g., income, health).

3.2.3 Step 2: Finding the road segments within each LAD

The input to step 2 consists of the output from step 1 (326 shape files for LADs) and the one shape file that defines roads as a series of segments, where new segments are made everytime a road bends or intersects with another road (Figure 3.2). The output is a road *network*, divided across the LADs. To create this output, we need to (i) identify the (parts of) roads that belong to each LAD, and (ii) convert roads from a shapefile format into a network. For the identification, we go through each LAD, and then through each road. The trivial cases are when the road falls entirely outside the LAD (discarded), or entirely within (assigned to the LAD). The one intermediate case is when a *part* of a road falls within a LAD (Figure 3.3). In this case, we divide the road in two segments: one segment for the LAD it belongs to (assigned to the LAD), and one remaining segment. Note that, while LADs do not overlap, some road segments may be at the border of two LADs. In this case, the segments are assigned to both LADs (i.e. duplicated). For the conversion, each road segment corresponds to one edge of our network, and each node stores the coordinates of the

segment's endpoints as in Figure 3.2. Note that our edges are not a one-to-one mapping of road segments in the road shape file, because some road segments may be sub-divided when they span two LADs.

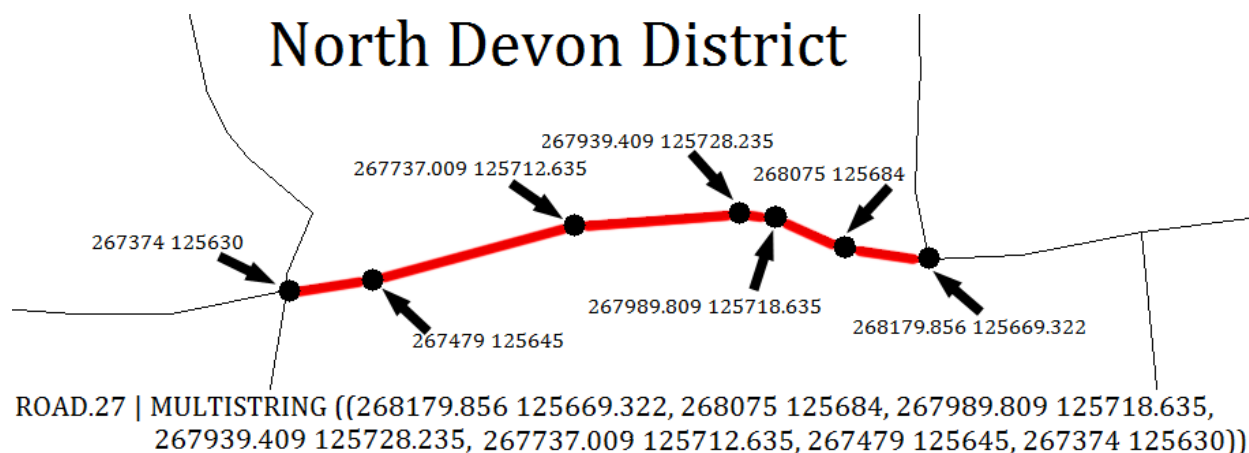


Figure 3.2: Roads are encoded in a shapefile as a series of segments. A segment links two points, specified as coordinates in easting and northing coordinates. Segments are created when a road has an intersection or turns.

After completing this procedure, we have 326 LADs and the road network within them. To ensure the validity of the data, we tested (i) whether the network in each LAD was connected with respect to fast-food outlets and schools, and (ii) whether the network has a large disconnected component even without fast-food outlets of schools. In other words, a school or fast-food outlet that is unreachable would indicate issues with the network data. Similarly, a part of the city that is seemingly inaccessible may indicate issues in pre-processing. We found 6 LADs (less than 2% of the dataset) experiencing one of these two issues. This was mostly due to a misalignment between boundaries for governance (the LADs) and the transportation network (Figure 3.4). For example, one city could be in charge of two areas, but the only road to move between them was within the boundary of another city. The six cases were manually resolved. For Tewkesbury, Windsor and Maidenhead, and Wyre, the small road fragment needed to connect the disjoint parts was re-assigned from the LAD where it fell (Cotswold and Gloucester, Bracknell Forest, and Fylde respectively). For Ashfield and

North East Derbyshire, the parts connected the ‘main’ city to a hamlet were far off, and we thus split each city into two LADs (one for the ‘main’ part and one for the hamlet). Finally, the Isles of Scilly contained roads over five disconnected islands. Since our records indicate that the islands contained no schools and no fast-food outlets, we dropped this LAD from our dataset. We thus had $326 - 1 + 2 = 327$ LADs.

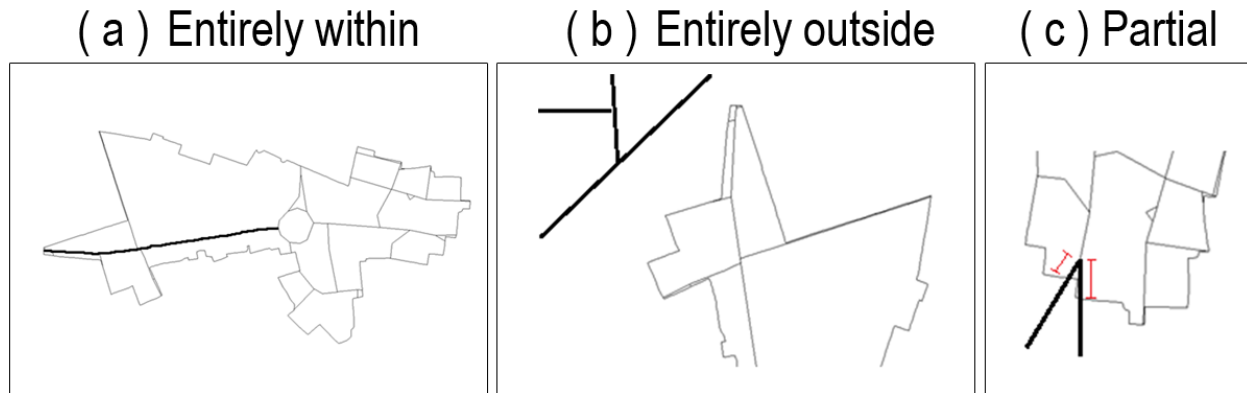
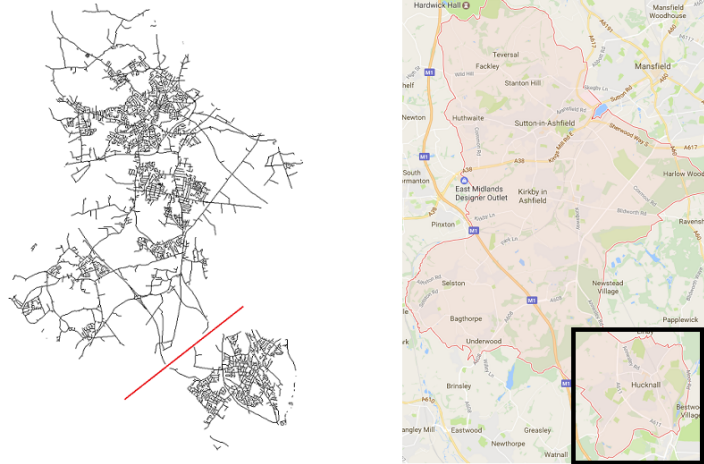


Figure 3.3: Three cases regarding the relationship between a road and an area.

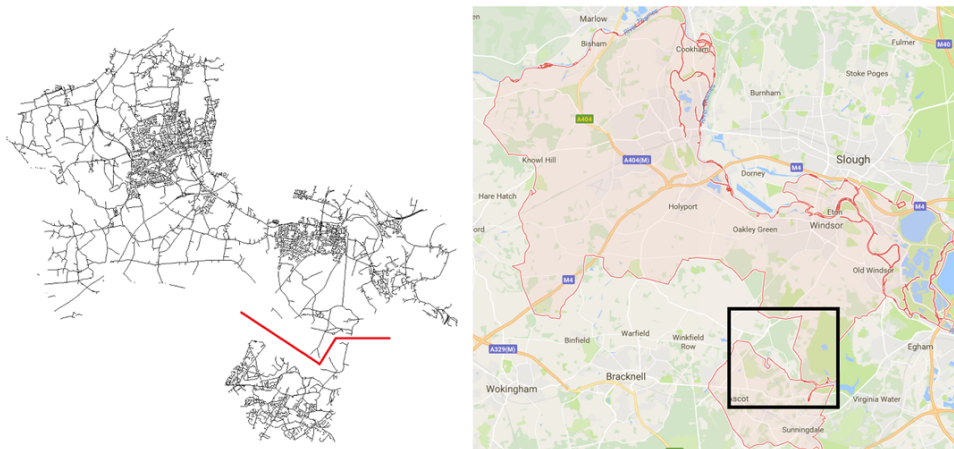
3.2.4 Step 3: Assigning schools and fast-food outlets to road segment

The input to this step consists of the road network divided across 327 LADs, and the Points of Interest data for schools and fast-food outlets. The data includes easting and northing coordinates, the postal code, and a district code. Several entries had missing information, such as incomplete postal codes or no district code. We discarded such incomplete entries, representing only 0.5% of the fast-food outlets and 0.6% of the schools. For the remaining data, we assigned the entities to road segments (i.e., edges of our network) in two steps: (i) identify the LAD based on the district code, and (ii) select the edge closest to the entity. A difficulty of step (i) is that the district code provides the name of a city, and not

Ashfield (City 5) - 25% of nodes Solution: Split graph on two subgraphs



Windsor and Maidenhead (City315) - 16% of missing nodes Solution: Connect two subgraphs



Isles of Scilly (City138) - 44% of missing nodes Solution: Remove from dataset



Figure 3.4: Three situations leading to a largely disconnected road network. Top: hamlet for which the access clearly lies outside the main area. Middle: a very small but critical road section is administratively in another LAD. Bottom: the whole area is formed of islands.

Table 3.5: Hypothetical example of data produced by step 3, showing a network where nodes have coordinates and edges count fast-food outlets as well as schools.

Nodes		
532715 181698		
532742.771 181787.615		
532339.31689 181923.56457		
532308.7005602281 181913.6821562441		
Edges		
532715 181698	532742.771 181787.615	
532339.31689 181923.56457	532308.7005602281	
Outlets/schools		
(532715 181698, 532742.771 181787.615)	0	1
(532339.31689 181923.56457, 532308.7005602281)	2	0

the name of a LAD. In most cases, the LAD had the same name as the city. However, for 36 cases, there was no LAD with the city's name. This occurred for LADs that represented counties, and had several cities (e.g., County Durham includes Durham, Derwentside, Sedgefield, Teesdale, etc). All 36 cases were resolved manually, using Google Maps as geolocation service to find the city in England, and thus identify the LAD that it fits in. After completion of step (i), we knew the LAD for 99.5% of outlets and 99.4% of schools. For each entity within a LAD, we computed its distance to all road segments of that LAD, and we assigned it to the nearest segment (i.e. with minimum distance). The resulting network (Table 3.5) has coordinates on the nodes, and number of fast-food outlets as well as schools on the edges. Note that we do not differ between outlets or between schools, hence we only keep track of their density nearby a given road segment. The distributions of fast-food outlets and schools per LAD follows a similar pattern (Figure 3.5), although we note that there are typically 0 to 150 schools per LAD whereas there is a wider possible range of fast-food outlets.

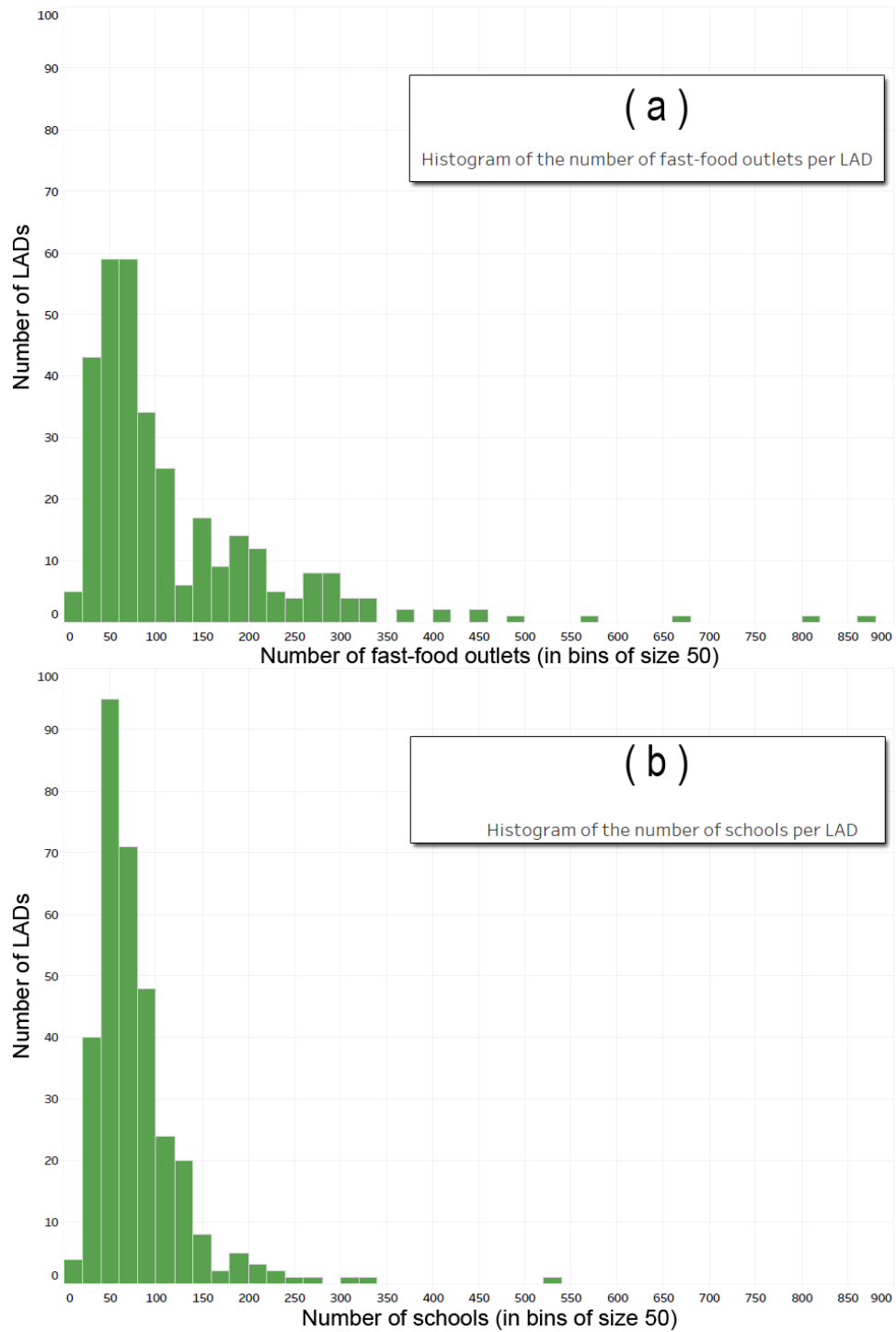


Figure 3.5: Distribution of the number of fast-food outlets and schools (x-axis) across LADs (y-axis).

3.2.5 Step 4: Identifying the Lower layer Super Output Area (LSOA) for each road segment

The LSOA contains statistical information. Identifying the LSOA of a road segment thus provides access to the deprivation level of this road segment. We started by excluding 5.63% of the LSOAs from the dataset because they were entirely outside of England, which is the focus of this work. Then, we identified the LSOA to which each road segment belonged. Because LSOAs were not designed to match the transportation network, we had to operate in the same way as in step 2: segments entirely within an LSOA were assigned to it, while those *partially* within the LSOA were further split. While LSOAs do not overlap, we also noted that several road segments were exactly at the boundary of two LSOAs (53,459 segments or $\approx 0,8\%$ of the data), and we assigned them to both (i.e., a given edge has either one or two LSOAs).

This process resulted in a final network size of 6,549,676 edges and 6,102,863 nodes. This leads to an extremely low network density ($\approx 3.51e^{-5}$), which we expect as a node is most frequently connected to two edges (since a road is stored as a series of lines) and cannot be connected to many others given the practical limitation on the number of roads that can intersect. When outlets were present on a street segment, there were on average 1.28 ± 0.76 outlets. Similarly, when schools were present on a street segment, there were on average 1.03 ± 1.19 schools.

As this is the last step that affects the existence of an edge, we also finalized spatial information about each edge at this step by computing the edge's length (based on the Euclidian distance between its two endpoints). Computing the distance was necessary to later answer questions such as how *far* schools can be from fast-food outlets. The average edge had a length of $59.69 \pm 69.08m$, with the large standard deviation due to the simultaneous

presence of long non-intersecting straight roads as well as extremely small segments (e.g. for tiny portions of roads spanning two LADs, on a road with a strong curvature approximated by many small lines).

3.2.6 Step 5: Adding the deprivation level of each road segment via its LSOA

The Index of Multiple Deprivation (IMD), commonly referred to as ‘deprivation level’ here, is a floating-point number assigned to each LSOA. When a road segment had a single LSOA, we thus assigned it the deprivation level of its LSOA. For boundary roads assigned to two LSOAs, we could not assume that their deprivation would be more like one LSOA or the other, and thus we assigned them the average deprivation level of the two LSOAs. As in previous analyses of the fast-food outlets in England with respect to deprivation [153], we simplified the (continuous) deprivation level into three values: low deprivation, medium deprivation, and high deprivation. As the deprivation level ranges from 1.097 to 92.601, we partitioned this range into three: low from 1.097 to 30.501 (excluded), medium from 30.501 to 61.002 (excluded), and high from 61.002 to 92.601.

3.3 Analytical methods

3.3.1 Overview

The following two sub-sections detail why, and how we computed our results from the network assembled in the previous section. Some notation will be used throughout this

section, and is introduced here. We denote a graph $G = (V, E)$ as formed of a node set V and an edge set E . The number of nodes and edges in the graph is denoted by $|E| = m$ and $|V| = n$ respectively. The ‘cost’ of an algorithm will be expressed in the worst-case, that is, as the peak resources that it needs to complete. Resources are divided into time (i.e. time complexity) and space (i.e. space complexity). The worst-case complexity is expressed using the \mathcal{O} notation, showing how either the running time or space requirements grow as a function of m and n . For example, a space of $\mathcal{O}(m)$ says that we need to store ‘in the order’ of the number of edges for an algorithm (thus omitting constants). For larger networks such as ours, acceptable costs rarely exceed quadratic forms: for instance, $\mathcal{O}(n^2)$ *may* be feasible, but $\mathcal{O}(n^3)$ may exceed available resources. When computing distances, we chose algorithms that provide exact answers at costs less than quadratic. When computing centralities, we opted for approximation algorithms given the high cost of the exact ones. Computations were performed on the shared High Performance Cluster (HPC) Gaea at Northern Illinois University, typically using 5 nodes (each equipped with 2 Intel X5650 processors and 72 Gb RAM). Our scripts for analysis are available within the ‘Analysis’ folder at <https://osf.io/gn3f2/>.

3.3.2 Computing shortest-path distances

The current public health context in England aims at countering the perceived proliferation of fast-food outlets around schools. The Supplementary Planning Document (SPD) can be used by local governments to enact local planning policies that affect fast-food outlets (formally defined as shop types that fall within Use Classes A5 for an SPD). While planning policies can range widely, two specific levers have received increased attention [162]. First, there can be a minimum *distance between fast-food outlets and schools*. Second, there can

be a maximum clustering, by limiting the number of fast-food outlets packed in an area, which consequently would increase *distance between fast-food outlets*. In both cases, policymakers need to decide on a specific value: how close is ‘too close’ to a school? How far should outlets be from each other? In the absence of detailed data, these choices are made on a best-guess basis, reflected by a wide array of values. For instance, Islington Council set a 200 meters buffer between schools and fast-food outlets, while others used a 400 meters buffer (Warrington Borough Council, City of Bradford, Barking and Dagenham, Solihull council) [48, 42, 46, 50, 49]. Similarly, the clustering was set to having no more than 10% of units in an area for Gateshead Council, whereas Barking and Dagenham used a 5% limit, and Solihull imposed a 15% limit. Target areas also varied, with some using zoning to control town centers whereas others targeted specific demographics (e.g., Gateshead Council imposed restrictions in wards where more than 10% of year 6 pupils were obese) [48, 246, 42]. Consequently, a major contribution of our work is to compute the distances used in both policy levers. That is, we compute the shortest distances (i) between fast-food outlets, and (ii) between fast-food outlets and schools.

The *generic* solution to compute shortest-path distances between two objects (i.e., a fast-food outlet and another outlet or school) is typically the Bellman-Ford algorithm whose time complexity is $\mathcal{O}(mn)$. In networks exhibiting desirable properties, more specific solutions can be identified. In our network, edges have a strictly positive *weight*, representing the length of the corresponding road segment. In this situation, Dijkstra’s algorithm is faster due to a time complexity of $\mathcal{O}(m + n \log n)$. While there exists an optimal $\mathcal{O}(n)$ algorithm for planar networks [108] (i.e. which can be drawn without two overlapping edges), the British road network does not satisfy this constraint due to the presence of overpasses (called flyover) including stack interchanges (when roads are above each other on multiple levels). We note that this problem does not affect all LADs: as of 2017, <http://www.cbrd.co.uk/> estimated that there were less than 30 stack interchanges in the UK. Computations may thus be

optimized by processing planar LADs differently than non-planar ones. However, to run distributed computations on the HPC facility, we ensured that the same version of the code was used for all inputs. Consequently, we implemented Dijkstra’s algorithm, and results were computed within approximately 42 hours.

3.3.3 Relating the presence of fast-food outlets to centralities

In this section, we relate the centrality of *nodes* to the number of fast-food outlets. The motivation for this analysis is as follows. Table 3.1 provides a sample of ten studies, all of which investigated betweenness centrality, and most of which also used closeness centrality. Considering a street network as a transport infrastructure, a typical concern is about the flow going through the network. In the absence of real-world data on traffic flows, betweenness centrality provides a proxy to network *flows*. Specifically, it assumes that places passed by a larger number of shortest paths connecting streets are more likely to be visited. This notion has been applied to many large networks [16], and has shown good correlations with important metrics for transportation networks such as congestion [114]. Closeness serves as a proxy to *access*, by identifying how easy (i.e. distance-wise) it is to get from a street to all others. Studies have shown good correlations between closeness centrality and urban elements such as economic activities [177] (and particularly retail stores [54]) or green spaces [202]. Research on food behavior also uses access as one factor driving the choice of a food retail location for individuals [249], highlighting that individuals are more inclined to purchase food sold within up to 1 mile, although other factors such as deprivation mediate this relationship [3]. Our overall process to relate centrality and fast-food outlets is summarized in Box 1, and detailed as follows.

Box 1. Process to relate node centrality and the density of fast-food outlets.

- (1) Remove nodes whose centrality would be zero.
- (2) Approximate the centrality of the nodes.
- (3) Transform the centrality into a ranking of nodes.
- (4) Compute the number of fast-food outlets nearest to each node.
- (5) Correlate (3) and (4).

Betweenness and closeness centralities are formally stated in the two definitions below. Note that they are both centrality *indices*. For instance, for two elements x and y , if the centrality $c(x)$ is at least as much as $c(y)$, then we conclude that x is at least as central as y . As stated by Koschutzki *et al*, “in general, the difference or ratio of two centrality values cannot be interpreted as a quantification of how much more central one element is than the other” [135]. Given that our goal is to correlate the centrality with the presence of urban elements, we do not want the correlation to be biased by wrongly using relative differences in centrality. After computing the centrality of all nodes, we thus normalize it by transforming it into a ranking.

Definition 1 Let $\sigma_{st}(v)$ denote the number of shortest paths between two nodes $s, t \in V$ that contain $v \in V$. Then, the **betweenness centrality** of a node $u \in V$ is given by [135]:

$$c_B(u) = \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (3.1)$$

Definition 2 Let $d(u, v)$ denote the shortest-path distance between two nodes $u, v \in V$. Then, the **closeness centrality** of a node $u \in V$ is given by [135]:

$$c_C(u) = \frac{1}{\sum_{v \in V} d(u, v)} \quad (3.2)$$

Computing betweenness and closeness centralities in a weighted graph takes $\mathcal{O}(n^3)$ time with a modified FloydWarshall algorithm. This can be improved for a sparse graph specifically (as is the case here) by using Brandes' algorithm which takes $\mathcal{O}(n^2 \log n + mn)$ time, but this cost remains very significant for a graph with millions of nodes and edges. We took two steps to improve it. First, similarly to Porta *et al* [178], we excluded nodes whose centrality would be 0, *without having to compute it*. That is, for betweenness centrality, we excluded nodes with a single edge as they act as sinks and no shortest paths go through them (Figure 3.7). Similarly, for closeness, we excluded unreachable nodes (since their distance to others would be infinite and their closeness tend to 0). This approach removed approximately 14% of nodes when computing betweenness, and less than 1% of nodes for closeness. We thus had to use a second step, in which we employ Eppstein and Wang's fast approximation algorithm for betweenness and closeness [68]. The algorithm randomly selects k pivots, and provides the probability that estimation errors are greater than $\epsilon \times (n - 2)$. A higher k or a lower ϵ would lead to more accurate results at the expense of more computation times. We thus have to identify suitable values of the parameters k and ϵ , while noting that these choices are interdependent (Figure 3.8). We set ϵ to a 5% error margin, and we performed a parameter sweep across all 327 LADs and values of k (from 1 to 1000). We identified $k = 109$ as providing a good level of accuracy while keeping computational time small (Figure 3.6).

After obtaining a ranking of nodes with respect to (i) betweenness and (ii) closeness, we had to correlate the ranking with the presence of fast-food outlets. Similarly to step 3 in assembling the dataset, we went through each fast-food outlet and assigned it to the nearest node (instead of the nearest edge as in step 3). Finally, we computed the Pearson correlation between the number of fast-food outlets and the ranking of the nodes, for both betweenness and closeness. Correlation values range from -1 (perfect negative correlation) to 1 (perfect positive correlation).

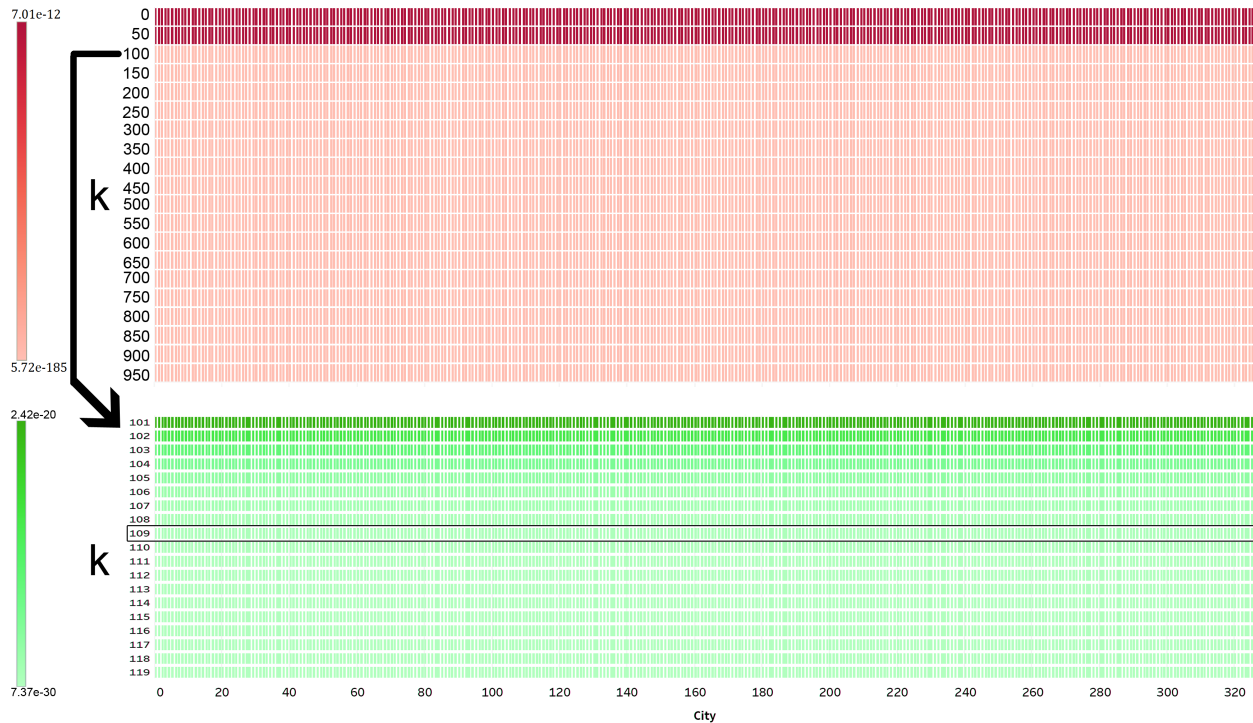


Figure 3.6: After setting ϵ to 5%, approximation errors depend on the number of pivots k (y-axis) and the number of nodes n , which varies across cities (x-axis). We found that the choice of city did not have a noticeable impact. Approximation error became small in the range 100-150 (top), and we chose $k = 109$ (bottom; framed). Due to the wide range of values, note that scales (i.e. colormaps) are different.

3.4 Results

The datasets produced by our analysis (previous section) are available within the ‘Results’ folder at <https://osf.io/gn3f2/>. We computed the distributions of distances between fast-food outlets and (i) the nearest fast-food outlet, as well as (ii) the nearest school. Based on these distributions (Figure 3.9), we can make the following observations:

- Fast-food outlets are very strongly clustered. Most of them are located either on the same spot or within a few dozen meters (60% of the data falls within 0 to 60 meters). Using a 120m buffer suffices to capture almost 80% of the outlets.

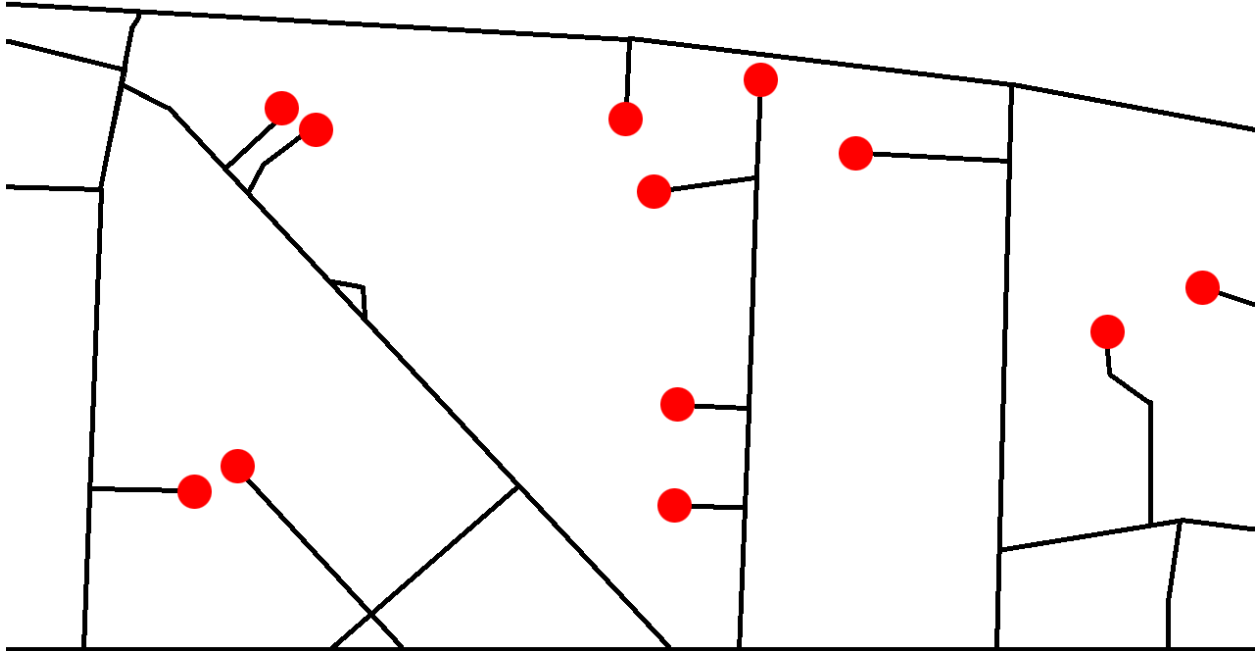


Figure 3.7: Nodes with a single edge are easy to identify, and removed before computing the betweenness centrality. This example in Adur City shows how 12 nodes (red circles) can be removed, as no path goes through them.

- While outlets are strongly clustered around each other, they are much less clustered around schools. Less than 5% of outlets are found within 60 meters of a school (compared to 60% with respect to other outlets), and less than 20% of outlets are found within 120m of a school (compared to 80% with respect to other outlets).
- The widest buffer of 600m around a school would capture over 80% of existing outlets, while the other classic buffer of 420m would capture about 65%. This shows that doubling the buffer does not double the number of outlets included.

We further investigated the relationship between distances and the fraction of fast-food outlets, in general as well as across levels of deprivation. After transforming our discrete distribution into a continuous one, we fitted different curves. We found that a logarithmic relationship had the best fit, ranging from $R^2 = 0.76$ to $R^2 = 0.87$. We also tried polynomials of degree 3 but found that they over-fitted the data ($R^2 = 0.99$) and thus retained the

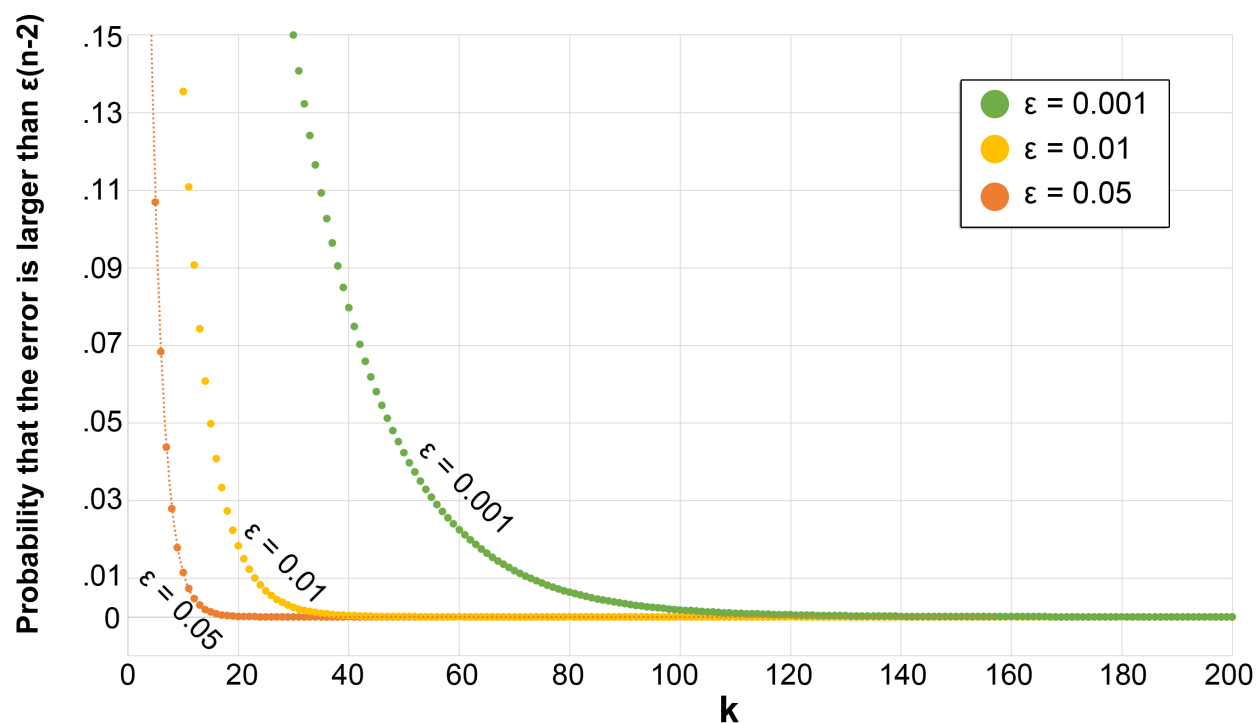


Figure 3.8: Probability that the error exceeds a target (depending on ϵ) for different number of pivots k . Computations were performed for Cornwall.

logarithm. The equations and corresponding fit are summarized in Table 3.6 while three examples are shown in Figure 3.10. Distributions appear lower as deprivation increases because most outlets were found in the least-deprived areas, followed by areas of medium deprivation, and then most deprived areas.

The correlation between centrality and the presence of fast-food outlets is shown in Figure 3.11. We observe that almost all of the data falls within the range $[-.1, .1]$ in which we conclude to the absence of a correlation. While three points fall outside this range, they are still at a very low level of correlation and may be outliers.

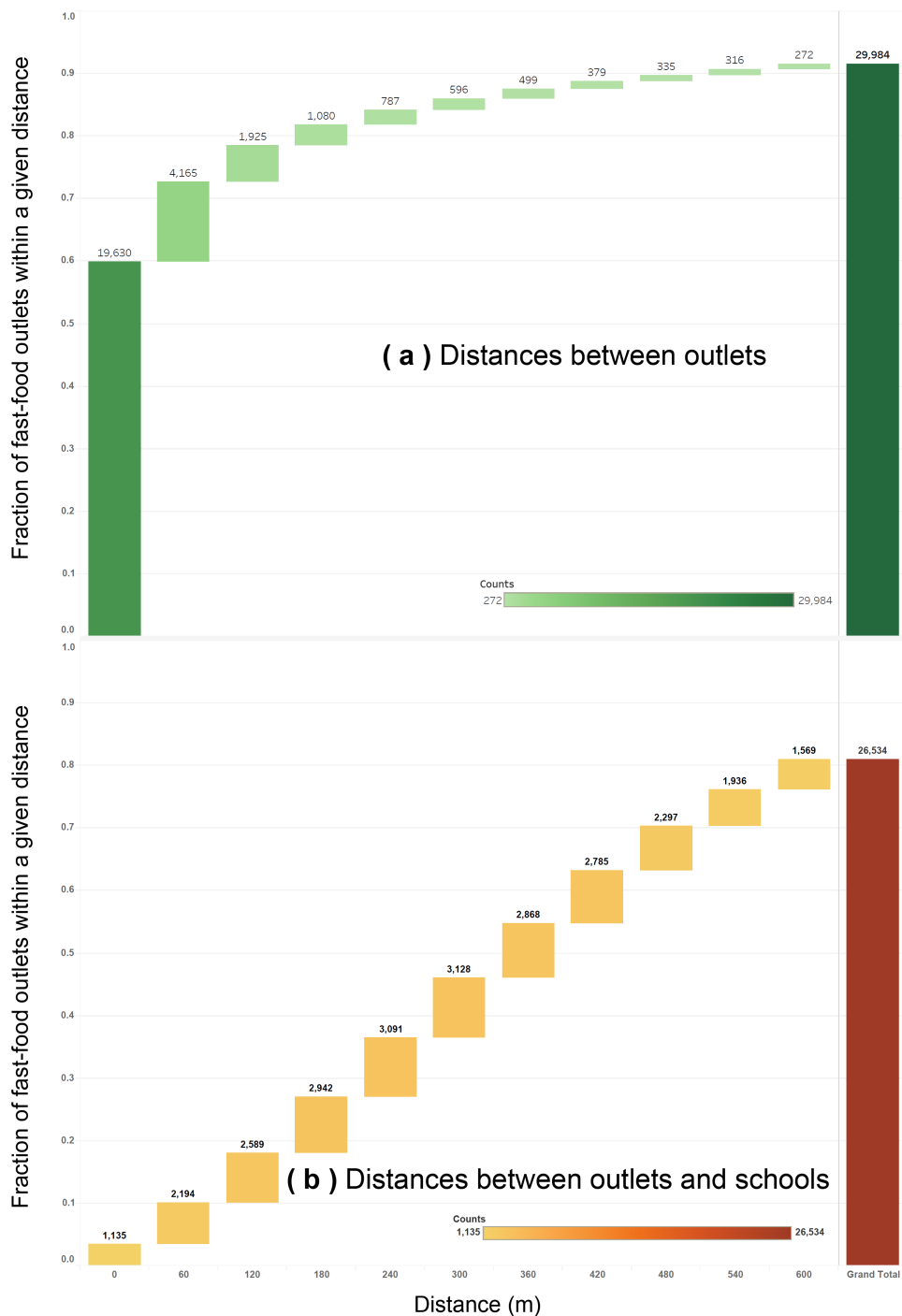


Figure 3.9: Distribution of distances (in meters) between outlets (a) as well as between outlets and schools (b). The a-axis goes up to 600m as it is the largest value encountered in current zoning policies regarding fast-food outlets and schools in England.

Table 3.6: Equations and fit (R^2) of the distributions of distances d for fast-food outlets with respect to other outlets, or to schools. All p-values are lower than 0.0001.

	With respect to outlets	With respect to schools
All deprivation	$0.109115 \log d + 0.181773$	$0.265617 \log d - 1.09389$
	$R^2 : 0.83$	$R^2 : 0.78$
Low deprivation	$0.0625031 \log d + 0.110917$	$0.15776 \log d - 0.653455$
	$R^2 : 0.82$	$R^2 : 0.79$
Medium deprivation	$0.0399222 \log d + 0.061546$	$0.0918827 \log d - 0.370456$
	$R^2 : 0.84$	$R^2 : 0.76$
High deprivation	$0.00695667 \log d + 0.00856996$	$0.016772 \log d - 0.0746996$
	$R^2 : 0.87$	$R^2 : 0.81$

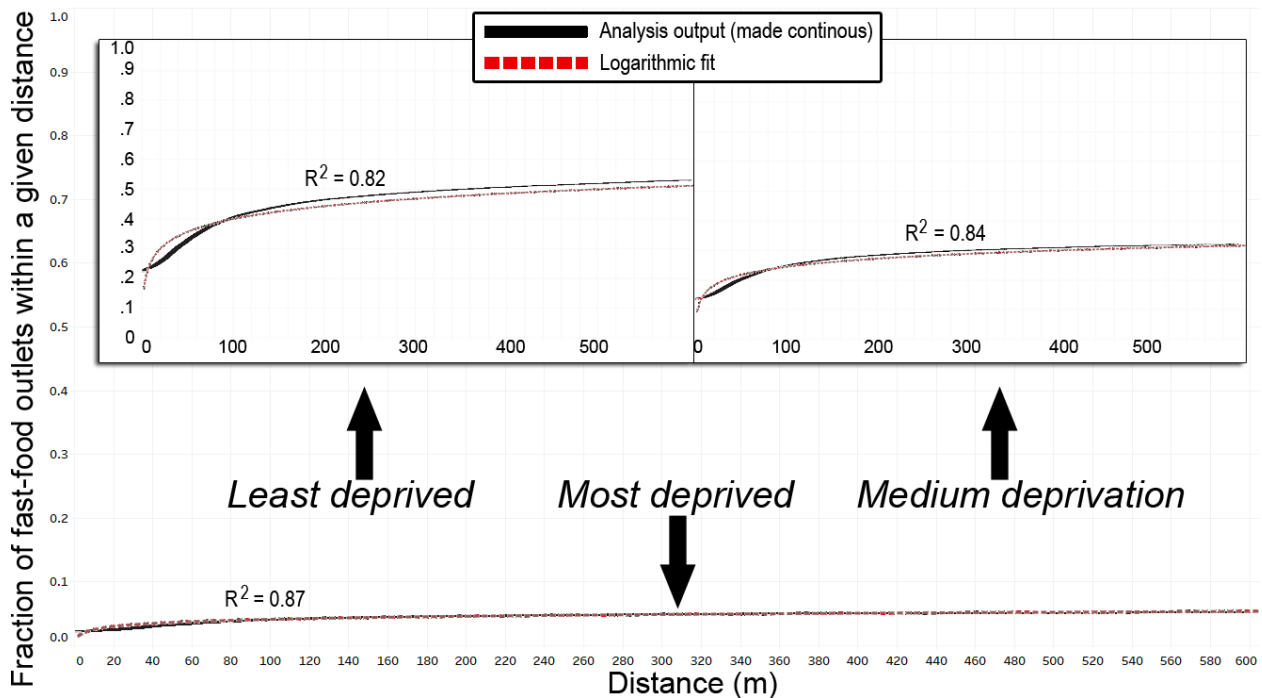


Figure 3.10: Fit between the analysis output (transformed from discrete to continuous) and logarithmic curves, across three levels of deprivation, for distances between fast-food outlets. Equations are provided in the first column of Table 3.6.

3.5 Discussion

While network analyses of retail activities have been performed at local scales (Table 3.1), our study is the first to do it over an entire nation. This was made possible by obtaining and

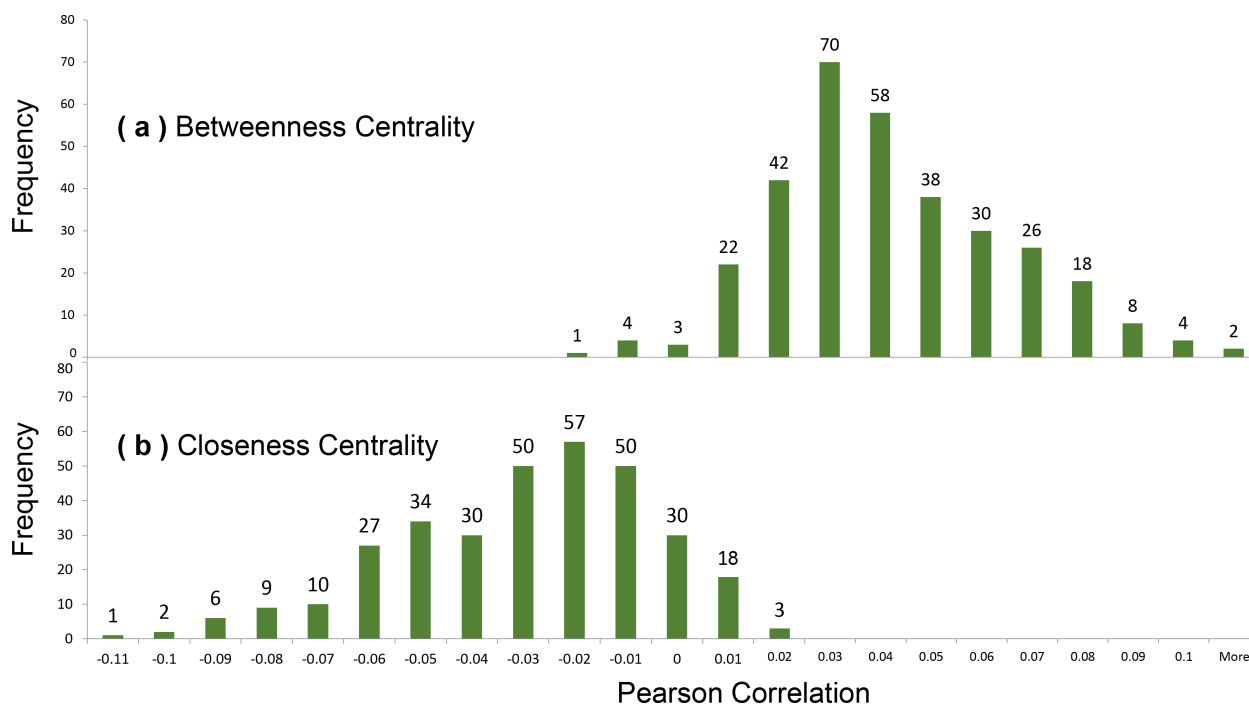


Figure 3.11: Distribution of correlations between the density of fast-food outlets and (a) betweenness centrality or (b) closeness centrality.

linking very detailed datasets, including the position of all outlets as well as the complete road network. Our focus is on fast-food outlets, and their relationship with schools. Research has suggested that this relationship is mediated by the level of deprivation, which we have included in our dataset to examine our findings across levels of deprivation.

Our first research question was to identify the distances between fast-food outlets and (i) other outlets as well as (ii) schools. This was motivated by the pressing need for a national evidence base to either (i) increase distances between fast-food outlets by limiting clustering, or (ii) create a buffer around schools. The 2011 National Institute for Health and Clinical Excellence (NICE) guidance recommended that local authorities regulate the number of fast-foods in specific areas, such as within walking distance of school [75]. The 2013 Academy of Medical Royal Colleges' report advocated to "reduce the proximity of fast food outlets to schools, colleges, leisure centres and other places where children gather" [168]. However,

neither could say exactly by which distance to reduce it, and what number of outlets would be affected, as this analysis was not previously available. Following these recommendations, several local authorities have started to use planning as a tool to address childhood obesity. As summarized by Peter Wright, an emerge view is that improving nutritional quality

“is not an issue that will be satisfactorily resolved by voluntary improvement, education, advice or any other “easy” intervention. Without political will and a determination to limit the proliferation of takeaway food businesses we are unlikely to make any meaningful impact on the impact of poor diet on significant parts of the population.” *Peter Wright, Gateshead Council, Centre for Diet and Activity Research (CEDAR), ‘Neighbourhood food environments, diet and health: research policy meeting’, Nov. 4th 2014, Cambridge, UK.*

Given the reality of having to address childhood obesity, local authorities have thus had to make assumptions about what distances were the right ones and what effect would be obtained. This illustrates the two unknown: what distance should we use, and how many fast-food outlets would it capture? The Takeaways Toolkit, considered to be one of the reference documents to assist with designing regulations, has previously emphasized the need for more evidence since such planning measures “have not yet been evaluated, and the impact on obesity and other health issues remains unknown”. This study contributed to the creation of robust evidence through our national-scale analysis of distances. We found strong spatial clusters of fast-food outlets (Figure 3.9-a): most fast-food outlets were within a few dozen meters from each other, and 80% of them were within 120 meters. However, clusters around schools were significantly weaker (Figure 3.9-b): less than 5% of outlets were within a few dozen meters from schools, and going as far as 120 meters captures less than 20% of them (compared to 80% when using other outlets as referential). This finding is in contrast to previous studies finding strong clusters around schools. This difference

may be explained partially by context, as previous analyses were conducted in Scotland, New Zealand, or the United States instead of the United Kingdom [11, 59, 66]. Our data can also inform authorities having implemented buffers around schools about the average fraction of outlets that may be captured: the 200 meters buffer for Islington Council [49] may impact a third of the outlets (based on national averages), while the 400 meters used by others [48, 42, 46, 49]. This suggests that increasing the distances between fast-food outlets may create more disruptive changes in the foodscape. However, like many upstream interventions, being disruptive can be both an opportunity (to avoid concentrated obesogenic environments) and a challenge (as many actors are concerned and a high political capital may be needed to enact such changes). Our last contribution regarding fast-food outlets and schools is the finding that their relationship may follow a logarithm (Table 3.6), which grows slower than a linear relationship. The practical policy implications is that increasing the buffer around schools (e.g., doubling its distance) provides a progressively lower level of return on the number of fast-food outlets affected (e.g., *less* than double the number of outlets). As we previously concluded in this situation, “moderate interventions would yield benefits, but stronger interventions may only be of limited further benefit” [85].

Our second research question was to investigate the relationship between network centrality and the density of fast-food outlets, thus taking previous local studies (Table 3.1) to a national scale. While previous studies found strong correlations between centrality and economic activities ($R^2 = 0.61$ [177], or $R^2 = 0.651$ [54]), we found no correlation: the correlation was close to 0 for 324 out of 327 areas, and only marginally beyond -0.1 or 0.1 for 3 areas (Figure 3.11). This suggests that, either at the national scale or at the scale of our areas, closeness or betweenness centrality were not a sufficiently strong factor to explain the location of outlets.

While our study combines large datasets from the national mapping agency with other governmental sources, there are nonetheless limitations to this work. First, the location of

outlets and schools is high but may not be perfect, as previous analyses have found the accuracy of the location database to range from 81% to 100% [76]. This creates a small margin of uncertainty on our results, but would not affect our broad conclusions on the lack of correlation between fast-food outlets and betweenness/closeness centrality or the much stronger clustering between outlets compared to outlets and schools. Second, while we used the most common forms of centrality from previous studies, there are many other forms. In particular, authors have also proposed using straightness [202, 241, 177, 178, 54, 221], or less common notions such as the cumulative number of turns or intersection crossings to reach destinations [206]. These metrics could also be approximated from our dataset, since each intersection of turn led to divide a road into another edge. However, the scale of our dataset raises the problem of efficient algorithms, and not all centrality metrics are supported by approximation algorithms (whose approximation factor is well-known or controllable). In addition, as there are dozens of centrality metrics [135], implementing and trying many would be a significant endeavour while not being necessarily the most informative. Indeed, it may be that several metrics taken *independently* exhibit low or no correlation, but *together* they may be more informative. In our future work, we plan to explore the combination of metrics that best explain the location of fast-food outlets. In addition, while this work provides national evidence regarding the strength of the *association* between schools and fast-food outlets, it cannot be used to make inferences about *causation*. Our next study will focus on causation, examining how different factors may successfully replicate the location of fast-food outlets.

CHAPTER 4

SIMULATING A NATION: A DATA SCIENCE APPROACH TO DEVELOPING A MODEL OF THE FAST-FOOD ENVIRONMENT IN ENGLAND

The previous chapter focused on the very first analysis of road networks at the national scale. In this chapter we simulate the spatial patterns of allocating new fast-food outlets and potential factors, which contribute to retailers' decisions. We conduct series of spatial simulations to assess which factors affect locational decision and we analyze the results with respects to different cities' characteristics. A factor is considered as being part of locational decision-making when using it in a simulation produces a distribution of outlets similar to real-world data. Our results suggest that proximity to schools may not be the main driving force behind the location of fast-food outlets, and that network centrality also has a weak predictive ability. The datasets described in chapter 3 have been used in this chapter as well.

All of this chapter will be submitted to the 2018 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (ACM SIGSIM PADS).

My contributions consisted of (i) assessing at the national scale whether fast-food outlets do preferentially target locations around schools and (ii) assessing which factor can best predict the distribution of locations for fast-food outlets.

4.1 Introduction

Understanding how social factors interact with the environment is essential to develop more effective interventions for complex problems, such as obesity. One particular environmental component that is receiving increased attention is the availability of local area fast-food outlets. In the United Kingdom, the consumption of food away from the home has increased by 29% during the last decade, paralleled by a dramatic increase in the number of takeaways or fast-food outlets. For these reasons, modifying the distribution and density of takeaways in cities and neighborhoods became very important in the policy agenda for both the UK and US[33]. The government and local authorities seek to empower individuals in making healthy choices by developing interventions to limit exposure to unhealthy foods. Despite their efforts, there is a lack of efficient plan to reduce and manage the issue of overweight and obesity. Therefore, studies exploring spatial trends in the growth of new outlets are essential to understand the changing foodscape and support policymakers in developing effective interventions.

In the early years of spatial allocations studies, the most dominant approach was based on intuition and experience of the owners. Additionally, stores tended to be separated from one another to maximize coverage and reduce competition [217]. However, the growing popularity of geographic data and development of spatial simulations caused a shift in locational decision-making. Controversially, close proximity to similar stores often gives more benefits to retailers than risks, encouraging them to cluster within small areas [123]. New factors that greatly contributed to this change of thinking include the transportation cost, which supports an agglomeration of many stores, for example at shopping malls.

Existing evidences of the usage of geographical methods for decision making in retail are very limited and refer to old trends of retail locational decision making. Simulation models

for fast-food outlets are nowadays especially important because (i) policymakers are already making decisions, and (ii) there is currently no tool to let policymakers try different policies virtually before implementing them in the real-world. In the process of developing simulation models that represent new policies on fast-food outlets, the first step is to develop models whose rules can at least replicate the current foodscape. Therefore, more studies involving network analysis performed at large scale are required to analyze the complex nature of spatial allocation of retailers' sites and factors that greatly impact them such as separation from competitors. We have investigated some of them as well, as presented in previous chapter. Specifically, our study is focused on road networks, which can give insights on the importance of roads based on an adopted measure i.e. different centrality measures (Figure 4.1). Overall this chapter aims to:

1. Verify whether fast-food outlets are preferentially located in close proximity to schools.
2. Determinate which factors have the greatest impact on retailers' locational decisions.

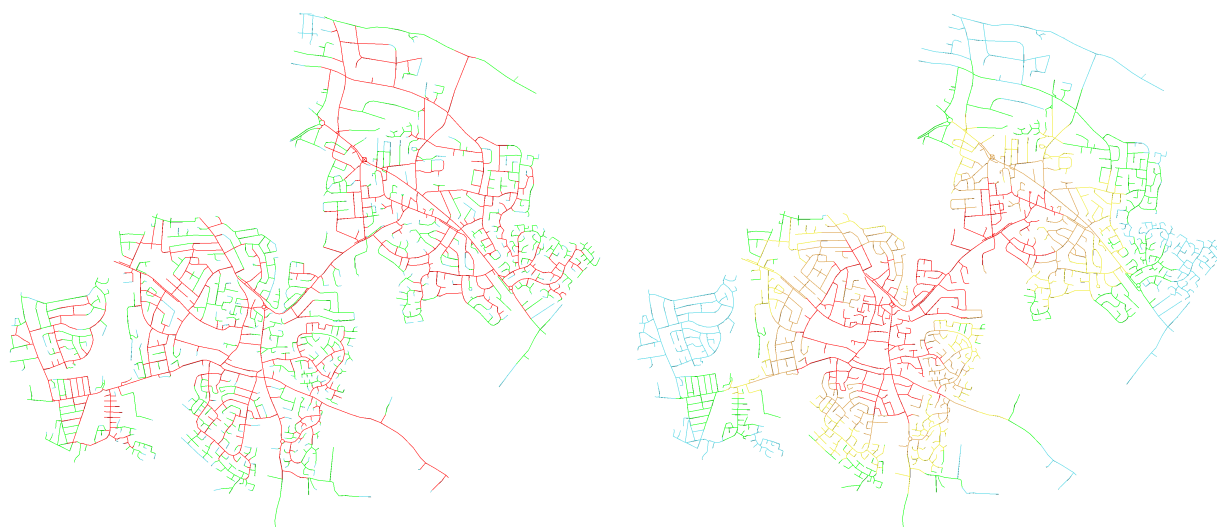


Figure 4.1: Example of centrality measures applied to the road network of Oadby and Wigston, left: betweenness centrality, right: closeness centrality.

The rest of the paper is organized as follows. In section 4.2 the most common methods used to analyze spatial allocation mechanisms are discussed based on previously developed models. Researchers recognize GIS data as a supportive tool, which allows to effectively track attractiveness of locations and proximity to competitors. The clustering pattern of outlets is also presented along with possible reasons behind retailers' decisions to reduce rather than maximize distance to other premises. In section 4.3, we detail the design of our spatial simulations. Developing and validation rules for the spatial distribution of fast-food outlets helped us to understand which factors contribute to locational decision-makers of retailers'. Rules used to obtain simulated allocation of restaurants include network measures such as degree distribution, closeness centrality, betweenness centrality, proximity to schools, and a purely random spatial distribution (used as baseline to evaluate the quality of other rules). Section 4.4 presents the most intriguing findings, according to which none of the analyzed factors had a significant impact on spatial allocation of outlets. In section 4.5 we discuss our results given our assumptions and the current literature. We also elaborate on limitations of our approach and future directions.

4.2 Background

The analysis of the existing distribution of fast-food outlets, presented in the previous chapter, describes the availability of unhealthy foods to the population of England. However, it does not tell how locations of restaurants will change in the future. The matter of where to open new premises has been a subject of many studies for the past 100 years [173, 181]. A historically prevalent assumption is that owners of food premises seek the areas which would maximize demand. For instance, this translated to choosing retail locations as close to the

clients as possible [217]. This assumption also highlights the importance of proximity to other economic activities, rejecting the idea of clustering as limiting the chances for monopoly:

“By contrast, he [retailer] desires to locate as far away as possible from his direct competitors, in order to enjoy as much of a monopolistic advantage as possible.” [217]

While this assumption dominated throughout the nineteenth century, the last decades have brought technological changes that transformed the theory behind economics. The reliance on cars resulted in establishing shopping centers, where numerous stores compete over overlapping control zones, thus running contrary to the theory of retailers’ attempting to stay away from potential competitors. Technological changes do not only concern how clients identify, and commute to stores, but also how the stores themselves optimize their locations. The expansion of data-driven decision-making and computing power have supported a shift in methods used by business organizations to select new sites. A survey conducted by Hernandez and Bennison in 1998 investigated the usage of different decision support tools by retailers in 8 sectors of UK retailing [109]. Findings showed a growing interest and use of GIS techniques as decision support tools. Grocery stores were also found to use a wide range of techniques, such as gravity modeling. Nevertheless, human judgments were still essential in the allocation processes. Similar results were obtained by Byrom *et al.* in their large-scale postal survey of retailers, conducted in 2000 [36]. In the end “common sense” and retailers’ intuition were again found to be the primary approach to locational decision-making, although authors highlighted the growing reliance on GIS tools. Additionally, knowledge about locations of competitors to improve strategy planning has been pointed out by responders as one of the main applications of geographical data. According to the results of this large scale survey, 46% of businesses are conducting competitor analysis on a regular

basis. This supports the theory that separation from homogeneous stores plays a role in the spatial allocation of outlets.

Clustering patterns are not limited only to close proximity of homogenous stores. Our investigation of the current distribution of fast-food outlets in England (see chapter 3) allowed to discover an existing association between locations of fast-food outlets and schools. Results showed that indeed fast-food outlets favor sites near educational institution, thus increasing the availability of high-caloric food for children.

4.2.1 Theoretical models supporting clustering

The phenomenon of spatial agglomeration of similar stores has been widely studied, providing many interesting ideas that can justify clustering patterns of some retailers' businesses. Models introduced by Konishi incorporate taste uncertainty into the theory of spatial competition [133]. Assumptions made in this study point out two important aspects that can explain clustering patterns of stores within shopping malls:

1. Consumers are not fully aware of their own expectation towards stores and available products within them. Therefore, a wider variety of competing stores have a greater potential to satisfy unsettled tastes of customers. Moreover, in terms of food, one place containing many restaurants, but offering different cuisines is a more desirable option for someone with unspecified preferences by allowing to *choose* which type of food to eat and this also encourages to return in the future.
2. Consumers calculate the expected benefit of choosing shopping malls by taking commuting costs (understood as both, financial and time, expenses) into consideration. From the perspective of transportation spending, shopping malls are also a preferable

option, because they offer many possibilities in only one place. Commuting costs are reduced while the time available for shopping is increased.

Building on these two assumptions, the authors also proposed as an additional factor the **consumers' expectations of lower prices** in areas with high concentrations of stores. Close proximity to retailers operating in the same market creates competition and consumers perceive it as a logical reason for stores' owners to lower their prices to attract consumers to come to their store.

Another model that partially supports previous findings is based on the simplifying assumption that the new stores will appear in the system at locations that **maximize profit** [123]. Sites within this model are characterized by four factors, which together capture the market share of a given retailer: position, the number of businesses in the established neighborhood, the variety of offered items, the price of the products sold, and the cost of commuting to a given location. Authors recognized that a customer's willingness to visit a particular site is proportional to the expense of getting there, which highlights the importance of spatial analysis. This approach pointed out the role of geographical grouping of stores as a main drive for increasing customers chances of finding a more desirable product and of reduced price. Furthermore, the paper suggests that existing distribution of retailers premises is a result of a well thought out plan of business owners rather than a random effect.

One more interesting idea of how retailers locate their businesses was proposed by Bester [21]. His model links together spatial distribution of economic sites with perceptions of product price and quality. The core idea is that low prices signal poor quality, unless a store is located within a cluster of other, homogeneous businesses. Therefore, clustering patterns can be explained as a **mechanism to justify lower prices** and results from market imperfections rather than strategic decisions of retailers. In other words, customers

have different knowledge about product quality among retailers, and are more willing to trust economic sites offering their products for cheaper, which are located among similar ones.

The evidence presented in this chapter indicates that clustering is an important pattern in studies of spatial distribution of economic premises. Many models have been created to explain this phenomenon, and many theories have also been suggested. Our own network analysis, presented in chapter 3, suggested that the distribution of fast-food outlets *may* also be following this pattern, since retailers' businesses tend to gather in close proximity of not only each other, but also of schools. Given the lack of a clear set of rules driving strategic spatial allocation, we decided to build on our associational study in the previous chapter, and develop in this chapter a study a causation. The next section details the design of our model and experiments.

4.3 Design of the Model and Experiments

4.3.1 Model boundaries

This model operates at the national scale of England, which is a novelty among network-based models and analyses for fast-food outlets. Each area of the model represents a city (also referred to as LAD), which contains:

1. **Road network.** The simulation will identify the location of fast-food outlets on this network. Features of the network (e.g. centrality) may be used to find locations, depending on the model's rule (as explained in the next sub-section).

2. **Location of schools.** This does not change during our simulations: we use the real-world data on the location of schools, and simulate the new locations of fast-food outlets.
3. **Number of fast-food outlets.** Our simulation will consist of taking each one of the fast-food outlets known to be in the area, and give them a location based on a simulation rule. The of a rule will compare the simulated locations with the real-world locations of the fast-food outlets.

Each model is a simplification of the reality and the one presented in this thesis is no exception. In the simplification process, several aspects were not retained for our model, but may be candidates for future extensions and adaptations of our work:

1. **Zoning regulations.** The need for zoning regulations has been recognized by several local policymakers. Furthermore, several local councils have implemented their own policies, as seen in official documents (Supplementary Planning Documents). Our model does not take into account these policies. The rationale is that the policies are currently restricted to very few communities, and our model operates at the national scale. That is, our rules allow for fast-food outlets to choose locations regardless of possible zoning regulations.
2. **Directionality of the streets.** We represent streets using an undirected network, that is, we assume that all streets can be traveled both ways. This is primarily due to our dataset, as the shapefiles used for the road network code them as sets of lines and do not provide directions or meta-data about the types of roads. Ignoring directionality may alter the results of our centrality-based rules, but it would not alter the results when rules use exclusively the presence of schools or fast-food outlets to determine new locations.

3. **Adjacency of cities.** Our simulations were parallelized by dividing the nation into independent cities. That is, each city was run on its own, and results were combined. Relationships between fast-food outlets and schools belonging to different cities were not included. For example, if a school is 500 meters away from a restaurant in a different region, but 700 meters away from a restaurant in the same region, then the ‘closest’ restaurant is 700 meters per our simulation. We do not expect these boundary effects to significantly alter the conclusions of our simulations, as neither schools nor fast-food outlets appeared to be typically located at the edges of cities (per our manual investigation of the dataset).

While designing the simulation model, we made the following assumption: the **distance between two premises on the same street segment is taken as zero**. Each street segment can contain multiple fast-food outlets or/and schools. Coordinates were used to assign premises to the corresponding road segment. That is, the smallest unit of analysis becomes the street segment. We cannot tell where outlets or schools are located *within* the street segment, thus we assign them a distance of 0 if they are both in the same segment. Segments have different length, thus this approximation does not create a uniform error throughout the simulation. We limited this issue by providing results in groups of distances (e.g. how many locations are within 0 to 60 meters away) rather than claiming a more accurate result.

4.3.2 Simulation rules

The simulation ran on described model places a corresponding number of *virtual* stores, based on simulation rules whose quality we seek to test. Each of the following five rules

states the selection mechanism by which a fast-food outlet chooses a street segment within a city:

- **Random selection** - business sites are chosen uniformly at random. This rule represents the method used as a baseline for comparison with other approaches. Factors which have a real impact on fast-food outlets distribution would outperform the random selection;
- **Preferential attachment**, a method in which some roads have a greater chance to be selected than others, according to one of the following four characteristics:
 - **Degree distribution**: the more favorable are the nodes connecting multiple street segments indicated by degree of given node (number of connections with other nodes). Degree distribution has been previously used in studies of the world-wide airport network [99] and coauthorship networks [2];
 - **Closeness centrality**: results of network analysis include a closeness centrality measure (defined in 2.3.2) for each node within a road network. Using this information, the impact of this measurement could be obtained by making roads that are more “close” in the network more likely to be chosen as a new site for outlets. This measure has been previously used to analyze scientific collaborations [2], although it is less commonly used than degree distribution or betweenness centrality;
 - **Betweenness centrality**: similar to the previous case, betweenness centrality (defined in 2.3.2) is calculated for each node in the road network and was used to differentiate roads and choose the ones that are more “between” as new locations for food premises. This approach has been applied before to model traffic

flow [114], airport network growth and distribution [99] and research collaboration networks [2];

- **Proximity to school:** This case has a great importance for presented research, because it focuses on the analysis of a potential causal effect between the presence of schools (as a customer base) and the opening of fast-food outlets. The goal is to prove that locations closer to schools are more likely to be selected. Results of this simulation help to answer the question if retailers are indeed clustering around places primarily because they are easily accessible to the youth, and by this create a greater exposure to unhealthy foods for children ¹;

4.3.3 Strategy for model validation

4.3.3.1 Running the simulation model

Simulations with randomness (i.e. outputs of stochastic models) need to be performed multiple times, in order to approximate the underlying distribution of the output. We thus need a method to determine the appropriate number of times that a simulation should run for the same set of parameter values, also known as the amount of *replicas*. Few different methods exist that can be used to approximate the right number of experimental replicas. The approach adopted in this work is based on boundaries of probability called Confidence Intervals. This statistical methodology estimates the range within which the real mean

¹ These types of experiments require extra computations of shortest distances between each street and closest school, and were made using Dijkstra's algorithm

average is expected to be false [195]. Therefore, the narrower the interval the more precise the estimate is expected to be:

$$CI = \bar{X} \pm t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} \quad (4.1)$$

where:

\bar{X} = mean of the output data from the replications

S = standard deviation of the output data from replications

n = number of replications

$t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$ = value from Student t-distribution with n-1 degree of freedom and a significance level $\alpha/2$

The required number of replications can be determined by rearranging the confidence interval formula:

$$n = \left(\frac{100St_{n-1, \alpha/2}}{d\bar{X}} \right)^2 \quad (4.2)$$

where:

d = the percentage deviation of the confidence interval about the mean

As an example, number of replications for Brentwood city using random allocation of fast-food outlets is equal to:

$$n = \left(\frac{100 * 0.254176396 * 2.2622}{10 * 0.374704462} \right)^2 = 235$$

The simulation process is based on three phases. Initially outlets are allocated to new sites, following the chosen rule. Then, distances between newly located restaurants and

schools are computed, using Dijkstra’s algorithm. Finally, correlation between the real distribution and the simulated one is obtained. This correlation is the output of a simulation, and thus we repeat simulations until the correlation falls within a target confidence interval. Specifically, we choose a 95% confidence interval with $t_{n-1,\alpha/2} = 2.2622$ for 10 initial runs (i.e. we perform a minimum of 10 runs so that we can compute the mean and standard deviation necessary for equation 4.2). Equation 4.2 concludes that we need more than the initial 10 runs, we automatically repeat the simulation for the additional number of runs. The pseudocode for a single simulation run is presented in algorithm 1.

Algorithm 1 Simulation algorithm for a single city

Require: Rule for allocating R new outlets AND evaluating using the real-world distance distribution D

- 1: $L \leftarrow \emptyset$ // L is a list of correlations used to compute right number of replicas
 - 2: **for** $n \in 10$ **do**
 - 3: allocateNewOutlets(R)
 - 4: $S \leftarrow$ distanceDistributionDijkstra() // S is the simulated distribution of distances between fast-food outlets
 - 5: $C \leftarrow$ computeCorrelation(D, S)
 - 6: $L \leftarrow L \cup C$
 - 7: $R \leftarrow$ computeReplications(L)
 - 8: **for** $r \in R - 10$ **do**
 - 9: allocateNewOutlets(R)
 - 10: $S \leftarrow$ distanceDistributionDijkstra()
 - 11: $C \leftarrow$ computeCorrelation(D, S)
 - 12: $L \leftarrow L \cup C$
-

4.3.3.2 Strategy for model validation

The core idea behind the experiments is to compare the real distance distribution (analyzed in chapter 3) and the one acquired from simulation. The correlation measure is used to

investigate the mutual relationship between distance distributions, which is known to reflect the strength of statistical dependence between data:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (4.3)$$

where:

X = first variable

Y = second variable

σ_X = the standard deviation of X

σ_Y = the standard deviation of Y

μ_X = the mean of X

μ_Y = the mean of Y

4.4 Results

Our investigation of potential factors influencing allocation of new fast-food restaurants relies on comparing the existing distribution of outlets (obtained from the data) and the simulated distribution of outlets (created under certain conditions). Computed correlation was averaged across all simulation runs for that city (to obtain at 95% confidence interval), and then averaged again in respect to analysis types (i.e. the correlation of a rule is the average of using that rule across all cities). Arithmetical mean obtained for particular simulation conditions reflect the influence of tested factors on distribution of new fast-food outlets. Random distribution was used as a baseline to measure the effectiveness of other factors. Our

findings, presented in Figure 4.2, suggest that none of the analyzed circumstances have a potential to significantly replicate current spatial patterns: random selection (**0.6264**), degree distribution (**0.6233**), and closeness centrality (**0.5900**). Betweenness centrality (**0.6454**) is the only condition that shows a slightly better result than random selection. The results from the least significant group, the proximity to closest school (**0.3617**) is not a statistically significant causation to support the theory of retailers' tendency to locate fast-food outlets around schools. These findings suggest that in England an existing clustering trends around school are an unintended effect rather than a strategic business decision.



Figure 4.2: The average (main blue bar) and standard deviation (orange sub-bar) for correlations from all simulation runs for each city, in respect to analysis type.

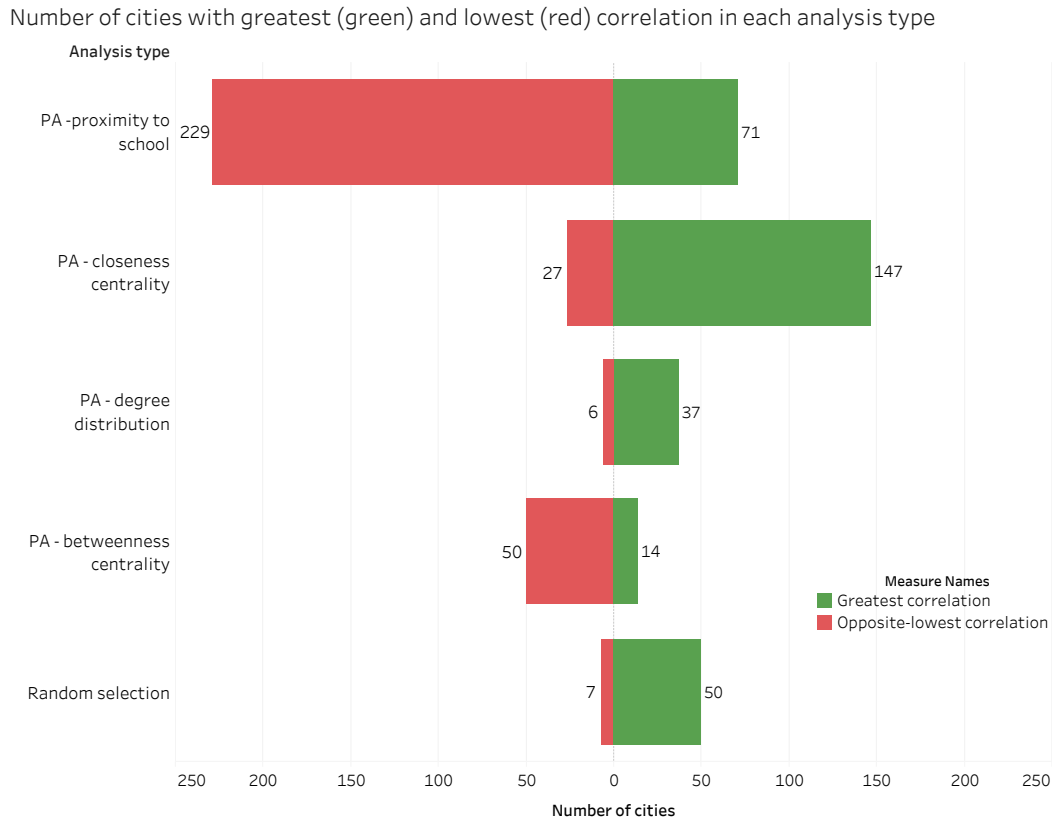


Figure 4.3: Number of cities with greatest and lowest correlation in each simulation category.

Statistics for both the lowest and greatest correlation coefficients of each city supports our claim of lacking evidence to prove takeaways spatial dependency on schools (Figure 4.3). Although in 71 cities proximity to schools was the aspect with the highest correlation, in 229 cities it was also pointed out as the factor with the least importance (lowest correlation to existing distribution). This might suggest that cities in England differ in terms of locational strategy and different factors impact retailers decisions.

The range of values for correlation coefficient can be also divided on few subgroups to better reflect the intensity of the relationship between two measures. In our analysis we adopted 5 categories 4.1. Analysis of number of cities in respect to these groups shows that

most of the values belong to the strong correlation category for all simulation types except the factor of proximity to school. Categories of very small correlation, small correlation, and medium correlation differ slightly among each other, while no correlation to existing distribution was found only in a few cases, as showed at Figure 4.4.

Table 4.1: The division of correlation coefficient values into five categories.

<i>Category</i>	<i>Range</i>
No connection	[-0.1 , 0.1]
Very small relation	[-0.3 , -0.1) & (0.1 , 0.3]
Small correlation	[-0.5 , -0.3) & (0.3 , 0.5]
Medium relation	[-0.7 , -0.5) & (0.5 , 0.7]
Strong correlation	[-1.0 , -0.7) & (0.7 , 1.0]

In the next phase of analyzing results, we focused on specific characteristics of cities. One of the characteristics was the deprivation level, divided into low-medium-high as in the previous chapter. Comparison of average correlations among different types of simulations showed that closeness centrality has been the most related measure to the real distribution of outlets, as presented on Figure 4.5. However, results for random distribution, degree distribution and betweenness centrality are very similar to one another, and do not vary much from closeness centrality. One outstanding factor, which has the lowest score in all deprivation levels, is proximity to schools, similar to previous findings.

We have also compared the importance of each tested factor among various deprivation levels. Performed simulations fit best for the real allocation of stores in wealthier cities, since the correlation coefficients for all except one factor were higher than 0.75. In contrast, values of correlation for most deprived areas are relatively smaller, in any case exceeding 0.5. One possible explanation for these findings is that cities differing from each other by wealth and prosperity take diverse factors into consideration in spatial decision making.

Comparison of correlation distributions for each analysis type

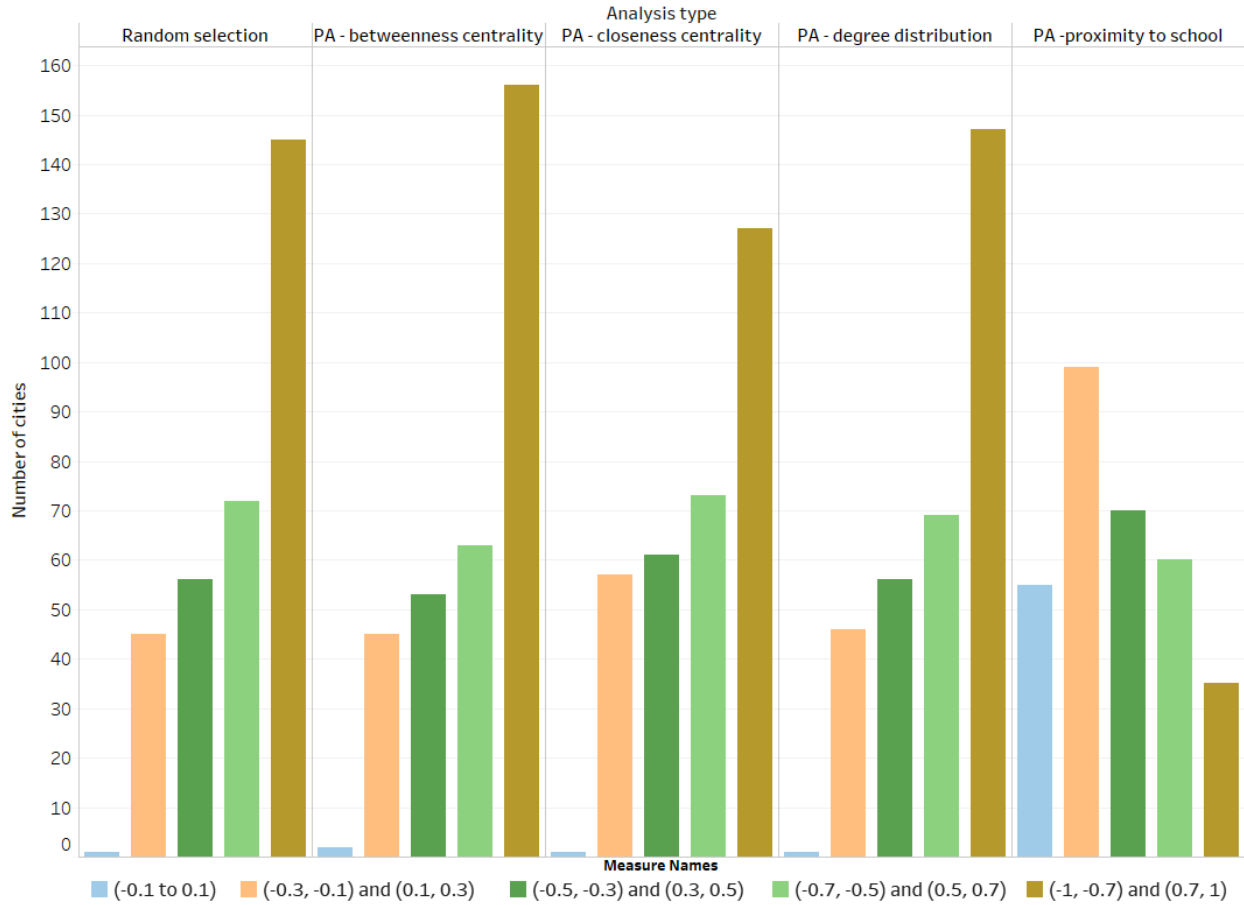


Figure 4.4: Numerical comparison of number of cities in each correlation coefficient bin in respect to type of performed analysis.

It seems that factors not included in this study (e.g., urban planning, land use, property prices [54, 241, 247]) play a significant role in more deprived areas.

Another characteristic used to analyze results of simulations is the number of outlets located within cities, presented on Figure 4.5. Again, we group values into 4 categories to gain some insights of the effects of particular factors on the distribution of outlets. This approach allowed us to discover that closeness and betweenness centralities, and degree distribution and random distribution have similar average correlations in all groups. Additionally, higher values of correlation were noted for cities with a larger number of fast-food outlets (over

100), while areas with less aggregation of businesses were significantly less related to the real distribution. Based on these results, it might be concluded that areas with greater and smaller number of food restaurants follow various factors in relation to spatial placement of new outlets and different approaches should be developed for them.

Similar findings were obtained when comparing average correlation among different population levels (Figure 4.5). Consistently with previous results, only proximity to schools has very little impact on the distribution of restaurants. Moreover, results for cities with a greater number of citizens (over 172000) have much higher correlation coefficients than other groups. This would suggest that analyzed factors have an impact on only some part of cities, while smaller ones adopted other guidelines for spatial decision making.

4.5 Discussion

4.5.1 Principal Findings

The network analysis, presented in 3 is consistent with results of previous studies from England, since we found that fast-food outlets tend to form clusters, and stay rather close to one other. Other relevant determinants, which greatly impact spatial strategy planning are residential and workforce population, transport intensity and nearby non-retail attractions such as transportation hubs and schools [138]. However, our spatial simulations proved that factors investigated in many network studies (i.e. centrality) or central to the current policy landscape (i.e. clustering around schools) are actually not a sufficiently strong driving force to explain the current locations of fast-food outlets. These results should be nuanced, because different cities follow different directions. Each characteristic considered in the analysis contains a spectrum of results where some groups of cities show very good correlation with

existing distributions, while others are only slightly related. Results for highly populated, wealthy cities with a great number of outlets shows greater values than in smaller and poorer areas. While network metrics tended to be in-par or slightly above baseline, proximity to school seems to have no importance, which is a very important finding for public health policies aiming at changing the dynamics of the fast-food landscape to reduce childhood exposure to unhealthy foods..

4.5.2 Limitations

Our study is conducted on the scale of the whole country, which is one of its key strengths for generalizability (in contrast to previous, small-scale studies), but it also faces many challenges and limitations. First, the focus of this work is on finding new locations for fast-food restaurants. Although this process of allocating new premises constitute the important part of any business, it is not the only locational decision available for retailers. Few studies pointed out the set of approaches called *6 R's*, which describes increasingly important approaches used by spatial strategy planners [36, 109]

- Roll-out/extension - Refers to process of opening a new store or altering the existing one by expanding occupied space;
- Relocation - Changing the location of the store due to competitors or availability of a more desirable site;
- Rationalisation - Closure of an individual store or disposing of divisions;
- Refascia - Changes in the appearance or name of stores, which might vary among different premises of the same business;
- Refurbishment - Altering the physical parts of an existing store;

- Remerchandising - Updating the range of offered products to better fit local customers' tastes;

The process of forecasting potential sites for new or relocated stores is even more complex because priorities given to these activities are not equal. Reynolds and Wood reported [192] there is less interest among retailers in managing existing stores and analyzing the distribution of outlets for the rationalisation process. The main focus is on opening and relocating outlets to more attractive locations. The scope of this work follows this trend and investigate the factors that drive the movement of stores, thus ignoring other possible locational decisions.

Another limitation is our use of one rule to assign all fast-food outlets, whereas different retailers may use different rules, and each rule itself may be a mix of different personal and organizational preferences. Studies exploring adopted methods, by locational decision makers in UK, found the whole set of approaches used at different levels of business management. Grocery sector alone takes advantage of combining together various techniques including statistical analysis, GIS support tools and gained experience [109]. Geographical solutions can be divided also into few sub-groups adding another set of possibilities to an already broad set of available methodologies. This work emphasizes the role of geographical data in analyzing and predicting future trends of food premises allocation. Network simulations have been conducted before, but on significantly smaller areas (i.e. city). Studies incorporating the data at the national scale suggest other GIS methods that can be used to explore big datasets. Table 4.2 contains a few examples of such cities.

Table 4.2: Studies performed at national scale using different methods than network analysis

<i>Study</i>	<i>Year</i>	<i>Methods</i>
“The geographical redistribution of retail outlets in Sweden 1998-2008” [6]	2016	The relocation of shops between localities of different sizes and between different kinds of positions within them. Two simple measures were used: Hoover index measures the general level of geographic concentration and redistribution rate
“Spatial distribution of economic activities in North America” [115]	2004	Number of measures used to explore the composition of economic activities across space: location quotient (LQ), locational Gini coefficient, clustering of establishments based on EG (Ellison and Glaeser) index, and measure of urbanization based on specialization of establishments
“Retail Concentration, Food Deserts, and Food-Disadvantaged Communities in Rural America” [24]	2007	Classification of census block groups as high retail access areas or low retail access area, based on zip code and a ten mile buffer around locations. Descriptive analysis: employment of both maps and tables to describe food desert counties.
“Shedding light on the global distribution of economic activity” [83]	2010	The creation of the disaggregated map of total economic activity: <ol style="list-style-type: none"> 1. Estimation of estimated total economic activity for each administrative unit by multiplying the sum of lights of each administrative unit by a unique coefficient; 2. Spatially distributing the estimated total economic activity of each administrative unit into 1 km² grid cells based on the percentage contribution of agriculture, the nighttime lights image, and the LandScan population grid;
“Location matters: comparing the distribution of economic activity in the Canadian and Mexican urban systems” [176]	1999	Used methods: classification based on city size and distance, grouping observations into comparable city-size classes and comparable distance classes, location quotients used to describe location patterns, for various sectors of the market, correlation coefficients to compare measures from different countries

4.5.3 Future work

Network simulation has a great potential to become an inseparable part of locational strategy planning by supporting decision makers and retailers in expanding their businesses. This chapter focuses on exploring determinants of spatial allocation of outlets and many further directions should be explored in the future. Analyzed conditions for opening new restaurants include two types of centralities, betweenness and closeness. Studies performed on a much smaller scale (i.e. city) applied also other kinds of measures, not included in this work. Further research should investigate influence of straightness centrality ², and accessibility ³ on distribution of premises. Another approach used in studies performing network analysis is based on measuring centralities at different scales. Porta proposed to calculate chosen measures both at a global level (entire area) as well as for local parts using buffers with an established radius, for example 800 meters [177]. Such division would focus more on particular parts of cities, instead of treating them equally.

From the perspective of city layout, one more characteristic stands out. Literature focusing on spatial location choices of retail establishments points city urban form and land use as an important factors, due to their complexity. Attributes of urban form include building's geometry, aesthetic dimension, capacity of the chosen environment to adopt change, and spatial accessibility [206]. The last aspect is especially attractive to researchers, who attempt to understand the competition between the retailers seeking for the best locations according to their needs and preferences. A study performed by Venerandi *et al.* investigated land use as an indicator of morphological patterns in terms of urban form and geographic

²Describes the extent to which the shortest paths from source node to all other nodes within network resemble straight Euclidean paths. The formal definition can be found at [178]

³Reflects the number of neighboring nodes reachable by imaginary agent randomly traversing the network, compared to overall number of nodes that belong to the neighborhood [221]. It is increasingly used in simulation models of food behavior such as [249]

locations [236]. Another study analyzed land use separately in terms of built-area and green areas [202]. However, these studies face some challenges as metrics of urban form are hard to capture. Therefore, more research is required to explore obscure relationships between urban form features and the locational decisions of retailers.

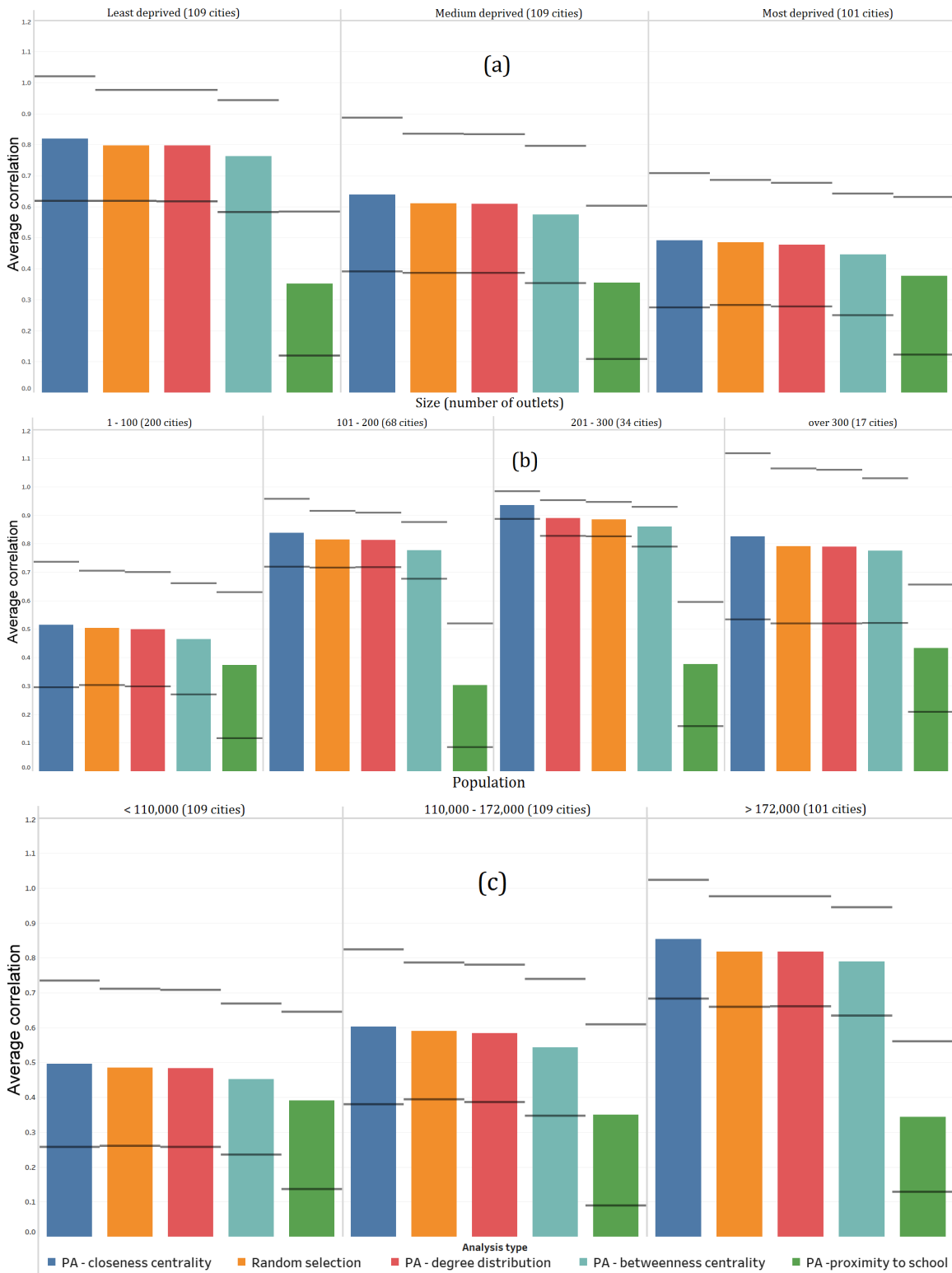


Figure 4.5: Comparison of results based on (a) deprivation level, (b) number of outlets, and (c) population. Average correlation is represented as bars and standard deviation as dark line marks.

CHAPTER 5

A NOVEL VISUALIZATION ENVIRONMENT TO SUPPORT MODELERS IN ANALYZING DATA GENERATED BY CELLULAR AUTOMATA

The third aim of this thesis is to develop novel visualizations to assist modelers in managing data produced by discrete simulation models. This aim is accomplished by designing, implementing and evaluating an environment focused on the visualization of two-dimensional cellular automata (CA) with square cells, which can intuitively be thought of as grids of colored cells. This chapter focuses on the design and prototype, with its main contribution being the use of a temporal clock glyph to show the successive states of each cell on the same display. We illustrate this approach on three classical models include epidemics. Based on feedback collected from trained modelers in this chapter, the next chapter refines the design and evaluates our environment.

All of this chapter was published in the following peer-reviewed conference article [89]:

- PJ Giabbanelli, GJ Babu, & M Baniukiewicz. A novel visualization environment to support modelers in analyzing data generated by cellular automata. Proceedings of the 2016 Human-Computer Interface (HCI) International Conference, Lecture Notes in Computer Science (LNCS) 9745, pp. 529–540.

My contributions consisted of (i) co-designing the software with PJ Giabbanelli; (ii) leading the implementation, with GJ Babu as contributor; and (iii) applying the software to three case studies.

5.1 Introduction

In the ‘big data’ era, a lot of attention is devoted to processing massive datasets about humans (e.g., Medicare data, hospital discharge data, police calls), by using machine learning or by calibrating and validating digital human models. These models also produce massive datasets to analyze. In particular, they typically produce time series capturing changes from baseline to the end of a hypothetical intervention. While only the last point is seen as the “final result”, both modelers and field experts often need to pay close attention to trends in the series. This can inform modelers of potential bugs in the implementation (e.g., identical consecutive pairs may indicate that results are mistakenly registered twice), while informing experts about the human dynamics (e.g., by observing cycles).

Consequently, many interactive visualizations have been developed for time series generated by simulations (Figure 5.1). In this setting, time tends to be either linear (i.e., an ordered collection of time points) or branching (e.g., a simulation splits into ‘branches’ when there are several possible outcomes or competing hypotheses) [4]. While sliders can straightforwardly navigate through time, they lead to issues such as change blindness (i.e., some differences from one time point to another may be missed). Pixel visualizations [128] or glyphs [242] allow to visualize multiple time series on the same space. Several temporal glyphs have been designed (Figure 5.2) and experimentally evaluated for tasks such as detecting peaks of trends [79]. While such innovative visualizations have adopted for simulations in engineering [134] (e.g., automotive, flows), there is a relative paucity of visualization environments for data generated by digital human models.

In this chapter, we focus on digital human models implemented as cellular automata (CA). Intuitively, a cellular automaton is a collection of coloured cells on a grid that updates over a period of discrete, fixed time steps based on certain rules defined around neighbouring

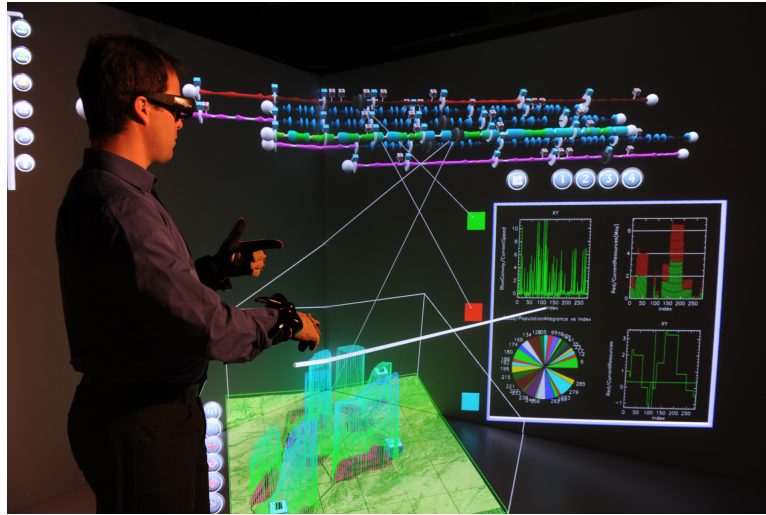


Figure 5.1: The EXP V2 environment [159] allows to explore a simulation by hand gestures. *Reproduced with permission from Defense Research and Development Canada, who holds all intellectual rights.*

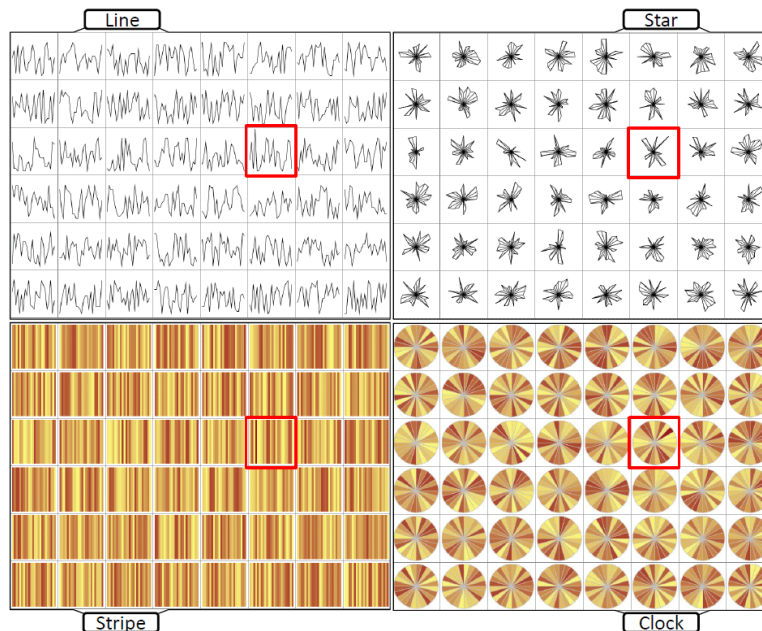


Figure 5.2: Fuchs and colleagues compared different temporal glyphs for a dataset with continuous values [79].

cells [152]. CA grids and cells can be of different types and shapes. Square and hexagonal cells are most common. CA models are generally one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D), but can have more dimensions as well. There is a vast quantity

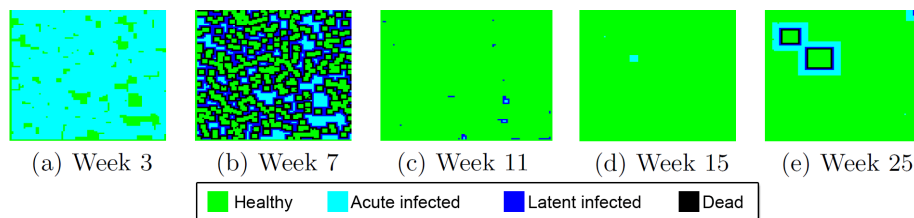


Figure 5.3: Cells within the body were modeled via a two-dimensional cellular automaton to study the spread of HIV [188]. Each figure corresponds to CA at a specific time step. To visually explore disease progression, the modeler would use a slider and go through the weeks, displaying the CA of each week one after the other.

of research using CA, as it can be applied to study any situation where individual units or cells affect others surrounding them [180]. In this chapter, we focus on two-dimensional cellular automata with square cells, while noting that the same principles would apply to other shapes of cells. Such CA are typically visualized by using a slider to move through the grid of states at different time steps (Figure 5.3).

Our main contribution is the design and prototype implementation of temporal glyphs for cellular automata. This allows to see multiple time steps rather than going through each one via a slider. Our hypothesis is that this new visualization environment can contribute to providing better analytical capabilities, particularly when designed for the specific needs of modelers.

This chapter is organized as follows. In section 5.2, we introduce our visualization environment and explain how the data is rendered. In section 5.3, the environment is illustrated for three well-known simulations (i.e., epidemics, sandpile, burning forest). Our hypothesis regarding the usefulness of this framework for modelers is discussed in section 5.4 based on the feedback obtained from trained modelers. Finally, concluding remarks are provided in section 5.5 together with a brief overview of future work.

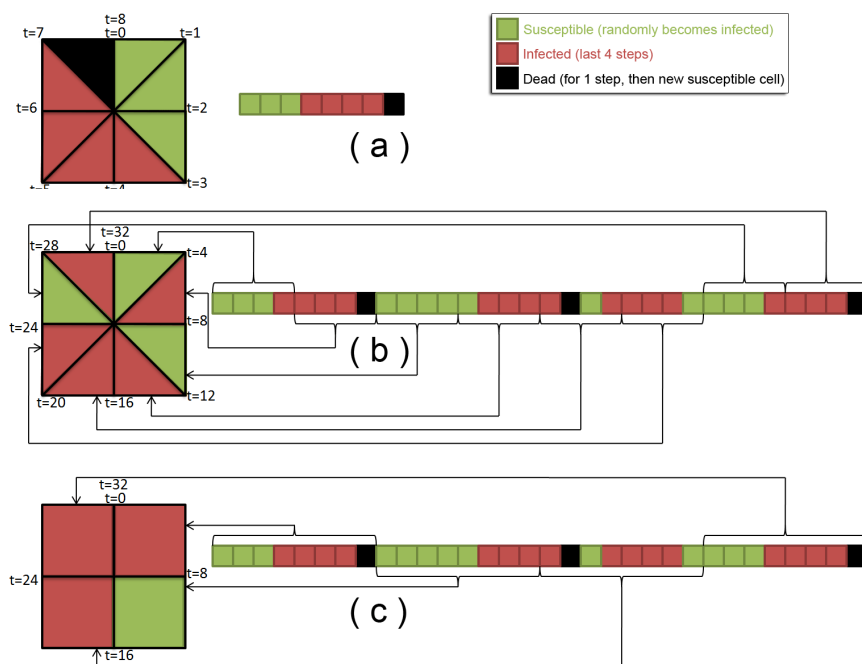


Figure 5.4: Data produced by a disease model (top-right: states and transitions) may be entirely visualized if the model is ran for a few time steps (a). As the number of time steps grows, they are aggregated into each of the 8 (b), 4 (c) or 16 possible segments.

5.2 Designing the visualization environment

Since CA have categorical values, line or star glyphs (Figure 2; top) are not suitable. Either the stripe or clock glyphs could be used. They encode data values through colours (Figure 2; bottom), and differ only in their encoding of time as either position (stripe glyph) or angle (clock glyph). Recent experiments found that the ‘clock metaphor’ helps with chronological orientation, thus proving better than linear layouts to detect temporal locations, and triangular shapes performed better than rectangular ones to encode colours [79]. This suggests that temporal glyphs with a ‘clock’ ordering and triangular shapes have potential to support visualizing CA. Since users of CA are particularly familiar with square cells, we used square cells as clock glyphs (Figure 3).

To understand the design challenges in using clock glyphs, we can think of the well-known pie chart. Having too many slices in a pie chart turns it from a meaningful visualization into an abstract pattern. The number of slices is thus best kept small, for example by collecting small slices within an ‘other’ category. A clock glyphs with too many data points would thus be like a pie chart with too many categories. This problem is particularly salient in our situation, since each glyph would have to represent the successive states taken by a cell over all time steps of the simulation, and there may be more time steps than could even fit within a circle (i.e., 360 slices). To address this problem, we limited the clock glyph to have either 4, 8 or 16 equal-sized partitions and each partition attempts at displaying the most relevant state within the corresponding segment of the data. This is illustrated in Figure 5.4. In Figure 5.4(a) we use 8 partitions and there happen to be exactly 8 time steps in the simulation, so each value is mapped to one partition. In Figure 5.4(b) there are more values than partitions, so each partition represents the most frequent¹ state among multiple data points. Consequently, the visualization depends on the number of partitions (4, 8, or 16) and on the aggregation method (e.g., most frequent value). Both are set by the user in the current prototype.

Research suggests that “multiple views are particularly helpful in analyzing time-oriented data” [4]. Consequently, another design consideration was to allow working across multiple data representation. Given that the glyphs need a significant amount of space to display each cell, our goal was to have a complementary representation that takes limited space and provides a higher level of abstraction. This was fulfilled by using a flow diagram as secondary view. Flow diagrams are the most common depiction of cellular automata models; that is, a modeler using CA would immediately recognize and know how to interpret a flow diagram. In short, a flow diagram shows each state, and possible transitions between states. Formally,

¹If the top frequency is found in multiple states, then ties are solved by picking the first one. For example, if there are 1 ‘dead’, 2 ‘susceptible’, and 2 ‘infected’ then ‘susceptible’ would be picked.

a flow diagram is a directed graph where each state corresponds to a node, and an edge exists from node a to node b if a cell can transition from state a at time t to state b at time $t + 1$. Our prototype automatically generates the flow diagram from the trace file (i.e. dump of simulation data).

5.3 Application to classical cellular automata models

5.3.1 Epidemics

In compartmental modeling, the population is divided into several groups or ‘compartments’, and then transitions or ‘rules’ specify the flows. The underlying mathematics are described in details by Hethcote [110]. Compartmental models of epidemics are typically named after the transitions between compartments: for example, in the SIS model an individual starts susceptible, can become infected, and eventually becomes susceptible again; similarly, in the SEIR model, an individual starts susceptible, can be exposed to a disease, then becomes infectious, and eventually recovers.

While a compartmental model represents a population, it can be ran on a cellular automaton where each cell stands for an individual. In this case, the rules reflect how infections can be passed on between neighbouring cells. This approach has been widely used. For example, there are cellular automata models of the classical SIR model [243] or SIS model [80]. In this example, we used the SIR model where an infected cell has a probability $p_i = 0.4$ of transmitting the disease to a healthy cell, and an infected cell has a probability $p_r = 0.5$ of recovering.

Visualizations of simulation traces from this model are shown in Figure 5.5, with different grid size (10 by 10 or 25 by 25) and different number of segments per cell (4, 8, or 18).

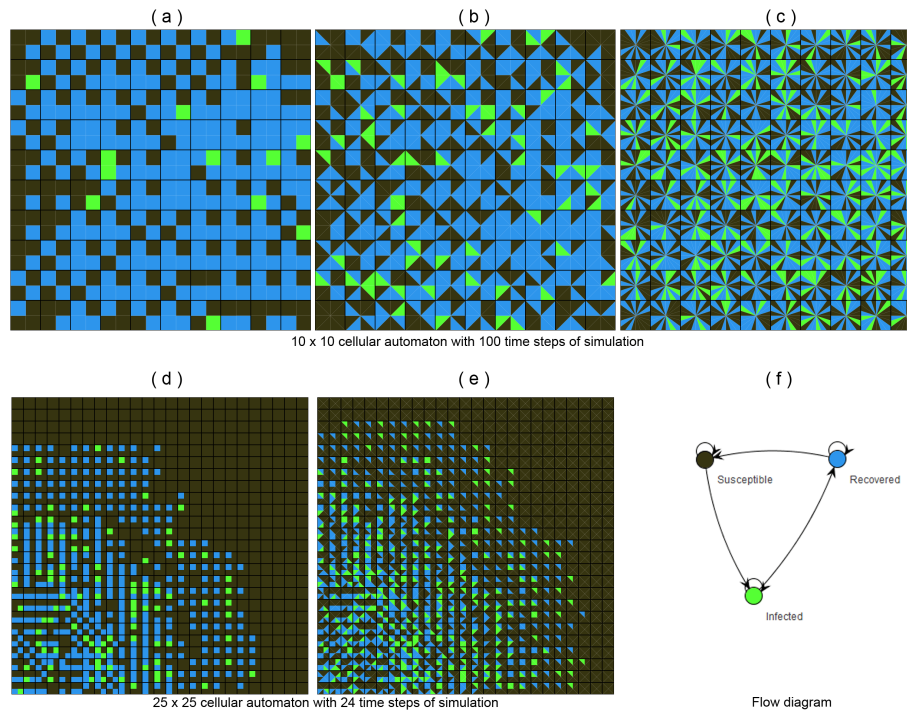


Figure 5.5: Visualization of an epidemic over a 10x10 CA where each cell has 4 segments (a), 8 segments (b) or 18 segments (c). The same model is visualized over a 25x25 CA with 4 (a) or 8 (b) segments. The same flow diagram applies to all simulations and is automatically generated (f).

The flow diagram is automatically generated by the visualization environment, and names or colours can be changed by the user. A consequence of displaying the most frequent state within each segment is that, as the number of segment decreases, some transitions are not visible. For example, we can see that cells get infected multiple times with 18 segments (Figure 5.5-c), less so with 8 segments (Figure 5.5-b), and not at all with 4 segments (Figure 5.5-a). Similarly, some states may not be visible: using 4 segments (Figure 5.5-d) instead of 8 (Figure 5.5-e) would tend to under-estimate the spread of the disease as peripheral cells that were recently infected do not yet display this infection.

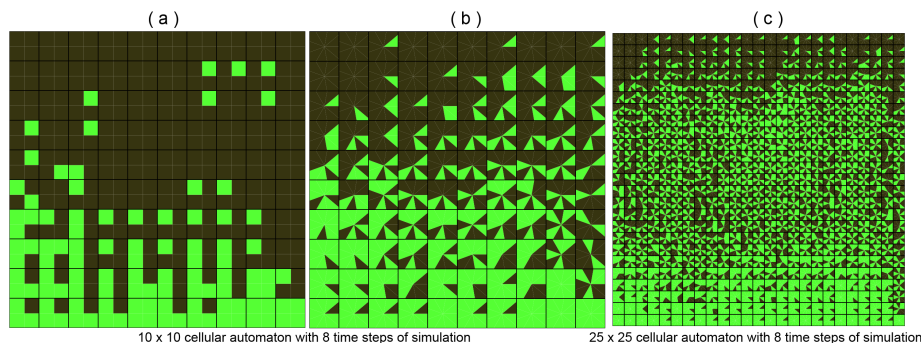


Figure 5.6: Visualization of a sandpile over a 10x10 CA where each cell has 4 (a) or 8 (b) segments. The same model is visualized over a 25x25 CA with 8 (c) segments.

5.3.2 Sandpile

The sandpile model is a vehicle to illustrate the theory of Self-Organized Criticality (SOC), that is, the idea that large interactive systems self-organize into a critical state and that small perturbations in this state trigger chain reactions. Informally, one can build a pile of sand by adding one grain at a time, until reaching a critical point where adding a single more grain causes an avalanche. This model was introduced by Bak and colleagues in 1987 [14]. We implemented the Sandpile model as described by Athanassopoulos and colleagues [10]. There is only one parameter p which applies when two grains are above two empty cells: the configuration either remains as such (with probability p), or both grains fall in the cells below (with probability $1 - p$). In our example, we used $p = 0.5$.

Visualizations of simulation traces from this model are shown in Figure 5.6, with different grid size (10 by 10 or 25 by 25) and different number of segments per cell (4, 8, or 18). All visualizations display that grains gradually fall and a stack of filled cell increases from the bottom. This would appear to be too simplified with 4 segments, and perhaps excessively detailed with 18 segments. The visualization with 8 segments could thus offer an interesting trade-off.

5.3.3 Fire spread

The mathematical principles of fire spread were summarized by Rothermel in 1972 [200]. Cellular automata have since been abundantly used to model fire spread within a spatial context, captured using either square [5, 126] or hexagonal cells [67, 229]. In its simplest version, each cell has three possible states: empty, tree, or burning. Initially, each cell is empty with probability $p = 0.3$ or a tree with probability $1 - p = 0.7$; the fire is started by picking one cell as burning. At each time step, a tree burns if at least one neighbour is burning, or has a probability 0.001 of spontaneously burning. A burning tree turns into an empty cell after 1 time step, and an empty cell can turn into a tree with probability 0.1.

Visualizations of simulation traces from this model are shown in Figure 5.7, with a grid of 25 by 25 and different number of segments per cell (4, 8, or 18). In this simulation, no large component formed, thus there were random sporadic and isolated fires. Since the fire lasts only time step and only the most frequent state is displayed, the fire is never visible. Thus, it is implicit that a cell transitioned from light green (tree) to dark green (empty) because fire occurred. Having the most segments (i.e. 18) shows that almost all cells have been occupied by a tree at some point, which is gradually lost as the number of segments decreases.

5.4 Feedback from modelers

Two trained modelers were contacted to provide feedback on the prototype. On the positive side, the overall idea of avoiding a slider was well-liked. One modeler stated: “I like it a lot because it simplifies visualization of states across timesteps”. On the negative side, a modeler reported that it became quite hard to read with a large number of cells or segments

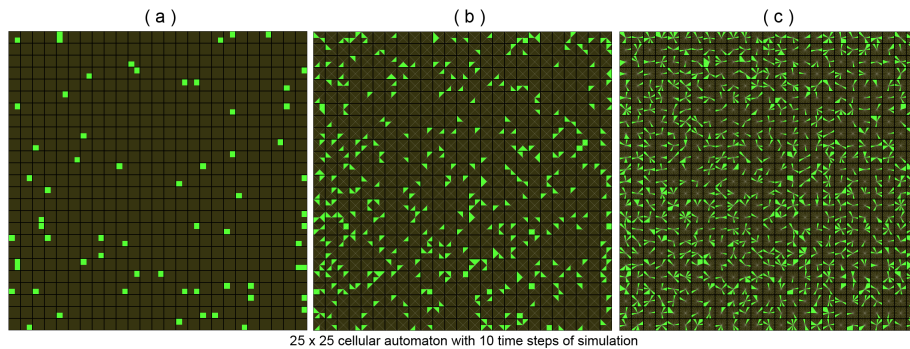


Figure 5.7: Visualization of a forest fire over a 25x25 CA where each cell has 4 (a), 8 (b) or 18 (c) segments.

per cell. Having 8 segments was considered the most readable version. Several improvements were suggested, falling into three categories. First, even if the motivation for this visualization was to avoid sliding through time, a slider was deemed useful to allow modelers to narrow the range of time steps that are displayed. In other words, modelers agreed that a slider to move through a *single* time step was not as effective as our visualization, but they recommended being able to move through a *range* of time steps. This is in line with the visual information-seeking mantra of starting with an overview, and then having details on demand: narrowing the time range would increase the detail of the cells since their segments would represent a narrower set of values. Using range sliders to narrow the data of interest was also done in the 2-d matrix-based interactive visualization by Song *et al.* [216].

Second, modelers appreciated the flow diagram and offered several ways to better link it with the main visualization. For example, hovering over a state or transition in the diagram should highlight all the cells that include that state of transition. Conversely, selecting a cell or group of cells should update the diagram to show only the transitions relevant to the selected cell(s). The idea that selections in one view would affect another view is known as ‘brushing’, and is essential to work across multiple data representations. Other (interactive) data representations were suggested, such as a stacked bar chart showing the number of cells in each state, which would also update when selecting specific cells.

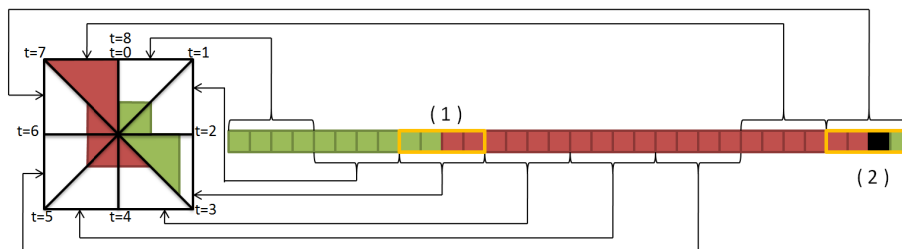


Figure 5.8: Segments could have a different length to represent another feature of the data. In this simple example, the length encodes the number of different states present within each segment. All segments are low (since they only have one state) but segments (1) and (2), which encode 2 and 3 states respectively. This encoding helps finding where changes happen in the data.

Finally, the visualization currently displays a summary of the successive values within each cell but leaves it to the user to find relationships between these values. One modeler suggested going further by displaying trends among the values as additional features: “the simulation will generate a temporal data streams (each cell will end up generating a stream of data points), so change analysis (shape, direction and velocity of changes) can be performed to understand how the whole system has impacted the individual cells”. Since the colour of the segment already encodes information (i.e. the most frequent state) and all segments must have equal width (as they represent the same amount of time steps), the main possibility to encode additional information is to use the segment’s length. This is illustrated in Figure 5.8.

5.5 Discussion

There is a growing interest in using visualizations at different stages of the modeling process, ranging from the early conceptual stage [94] to experimentation [159] and the analysis of results. There has been a particular interest for visualizations for cellular automata [137], as it is a widely used modeling approach. In this chapter, we focused on the analysis of data generated by a two-dimensional cellular automaton. We presented a visualization in

which the successive time steps of the simulation are aggregated and displayed all at once. The resulting visualization allows to see key properties of the models, such as grains of sand falling in a sandpile, or an epidemic spreading. Nonetheless, the visualizations had a number of shortcomings, mostly stemming from the aggregation method and/or the number of segments used within each cell. Two approaches should be explored in future work.

First, we could introduce a customizable weighting, allowing to under- or over-weight certain states for display. For example, consider an epidemic in which individuals start as healthy, get infected, and either recover by being healthy again or die. This scenario has three states (healthy, infected, dead) but they may not be equally important. Indeed, if we are concerned with the spread of the disease, we may want to underweight healthy individuals, give a neutral weight to infected individuals, and over-weight dead individuals. Similarly, in a forest fire, we may be less interested in seeing empty spaces than we are in seeing burning trees. In addition to allowing users to customize the weighting, an interesting research avenue would be to automatically set the weights based on the dynamics of the data. The simplest way would be to perform the equivalent of a histogram normalization, where very frequent states are under-weighted while rare states are over-weighted. However, finding a weighting scheme that best helps modelers understand the dynamics would require performing change analysis on the data as well as structural analysis on the flow diagram.

Second, we could create a large databank of visualizations in which each dataset is visualized using different aggregation methods and number of segments. Then, modelers would assign a score to each visualization based on how informative they find it for a given task. Tasks would be chosen by their relevance to modeling, and by their heterogeneity in terms of the perceptual notions involved. Example of tasks could include identifying cells whose final state is the initial one, localizing a spread, finding clusters of cells in the same state, etc. For example, consider the epidemics described in section 3.1: that same dataset may be rendered with 2 different aggregation methods and 3 different number of segments. For

each of the $3 \times 2 = 6$ visualizations, modelers would assign a score from 0 (least useful) to 5 (most useful) expressing the usefulness of the visualization for localizing disease spread. This would generate a relational database consisting of properties of the dataset (e.g. number of states), aggregation method, number of segments, and mean score per specific task. To understand how visualization parameters (i.e. aggregation method and number of segments) affect task performance for a given dataset, we could then mine the relational database by building classifiers [52, 53, 84]. We acknowledge that assembling a dataset where modelers judge a large number of visualizations is labour intensive. Nonetheless, having the target audience evaluate the visualizations for a set of task is a routinely performed procedure.

CHAPTER 6

ANALYZING SPATIO-TEMPORAL AND MULTI-RUN DATA PRODUCED BY SIMULATIONS FROM CELLULAR AUTOMATA

The prototype in chapter 5 only dealt with spatio-temporal data, which is one facet of the scientific data produced by CA. In this chapter, we also deal with multi-run data, since simulations on CA are often repeated (e.g. to account for randomness in the model or to study the impact of different parameter values). To do so, our expanded visualization environment includes additional linked views, and represents variability within the glyphs. We conducted an empirical evaluation of this new environment to (i) assess whether important tasks for modelers can be performed efficiently with this environment, (ii) examine how performances are influenced by key simulation factors, and (iii) identify whether modelers can use the familiar slider-based visualization together with our new environment.

All of this chapter was submitted in the following peer-reviewed journal article:

- PJ Giabbanelli & M Baniukiewicz. Analyzing Spatio-Temporal and Multi-Run Data Produced by Simulations from Cellular Automata. *Under review.*

My contributions consisted of (i) performing the empirical evaluation and processing its results (including extracting confidence, correctness, and error patterns from the videos), (ii) reviewing the literature of multi-faceted scientific data visualization, and (iii) implementing the design. PJ Giabbanelli designed the software and the empirical evaluation. J Salim assisted with processing results.

6.1 Introduction

Visualization is a typical component of Modeling and Simulation (M&S) studies. It can be used in the early stages of M&S studies to design or refine a model, for example, by identifying data patterns that should be included in a model or by highlighting discrepancies between a model and available data [94]. At a later stage, visualization can be employed to represent and analyze simulation results [227]. Such results are often multi-faceted scientific data, as they may include both spatio-temporal data (e.g., when a grid representing a physical space is updated over a set of time steps) and multi-run data (e.g., for stochastic simulation models which are run multiple times to estimate the distribution of the output). Simulation models creating spatio-temporal and multi-run data can be found in applications ranging from climate research [69, 208, 164] to geo-engineering [155, 113]. Despite their relative ubiquity, visualization methods have often focused on only one facet [127], through systems that can assist in the exploration of *either* spatio-temporal simulation outputs, *or* multi-run/ensemble data. In addition, advanced systems have been developed by the visualization community, but may not have made their way into practices within the simulation community. In this paper, we design, implement, and evaluate a new set of visualizations to explore the spatio-temporal, multi-run data generated by discrete simulation models.

The models under consideration are Cellular Automata (CA), which have been widely used for decades in simulation. Recent examples of their application include geophysical systems (see [230] and references therein for a brief review) and pharmaceutical modeling [22]. In short, a CA has a regular division of space (e.g., a square grid or hexagonal lattice) in elements known as *cells*. Cells have a state, which is updated over a period of discrete, fixed time steps based on rules involving neighboring cells. The temporal aspect is important for CA. For example, individuals infected with HIV have to take treatment for life, and the

virus may develop resistance to treatment due to mutations occurring daily (when the virus replicates). CA-based simulations of HIV may thus run for many time steps [188]. There can also be a significant number of runs to reach typical targets (e.g., 5% confidence interval for a simulation output). In the context of HIV, massive variability between individuals is such a hallmark of the disease that its presence can be used to validate a simulation [188]. Our focus is on visualizing data generated by CA with multiple time steps and several simulation runs. While the size of the CA (i.e., number of cells) is also a factor for visualizations, our system is primarily designed for small to medium grids, and possible extensions to larger sizes are examined in the Discussion section.

The *typical* visualization for a 2-dimensional CA consists of showing the cells on a plane, with colors standing for states (Figure 6.1), and to navigate across simulation steps using a slider. That is, analysts move through snapshots of the simulation, seeing all states but for only one time step. This visualization has been witnessed in many contexts, ranging from geo-social dynamics (e.g., migration [56] and insurgencies [180]) to biological ones (e.g., spread of HIV within a host body [188]). While this approach is typical, reviews in visualization have suggested that “visual analysis of a larger number of concurrent data volumes requires more sophisticated methods” [127]. Consequently, the main question here is *not* whether visualizations built for spatio-temporal, multi-run data can be more effective than those who are not built (but commonly used) for this purpose. Rather, our research questions are (i) whether important modeling tasks can be performed correctly and confidently with our proposed environment, (ii) how performances depend on simulation parameters (e.g., when there are more time steps or more variability), and (iii) whether the familiar slider-based visualization can be used together our new environment.

The remainder of this paper is organized as follows. In section 6.2, we summarize how the problem of visualizing results from CA and related structures has been addressed so far, in part through our early prototype. Then, section 6.3 presents the design of our environment,

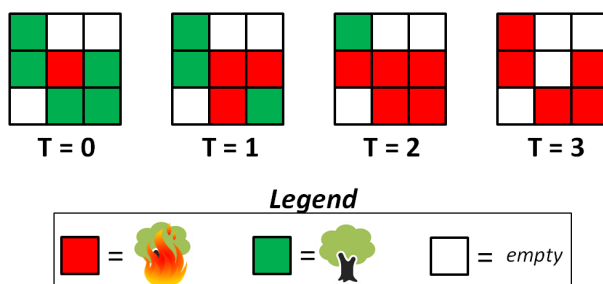


Figure 6.1: This cellular automaton has 9 cells representing trees, burning trees, or empty spaces (e.g. clearing in the forest). It changes over time as the fire spreads, and trees eventually burn entirely.

significantly extending our previous work, by handling replications and offering details on demand as well as filtering tools. In section 6.4, we explain the design of our empirical evaluation including the tasks that participants were asked to perform and the surveys that they filled. In order to provide full disclosure, all documents for this study (including our source code, script for participants and all their recordings) can be freely accessed at <https://osf.io/brjqr> while a video presenting the problem, the software, and the tasks can be seen at <https://www.youtube.com/watch?v=s5jQ2Z6iIf8>. The results of our evaluation are provided in section 6.5 and discussed in section 6.6. Finally, we offer concluding remarks in section 6.7.

6.2 Background

As detailed in the introduction, there are two facets to the data considered here: it is spatio-temporal (with an emphasis on the temporal dimension over space in this paper), and it has multiple simulation runs (also called ‘ensemble data’). Visualizations for simulation data have often been designed with respect to only one facet, and they are thus briefly reviewed in turn.

6.2.1 Visualizing time-series of simulation outputs

Many interactive visualizations have been developed for data generated by simulations over time. In this setting, ‘time’ tends to be either linear (i.e., an ordered collection of time points) or branching (e.g., a simulation splits into ‘branches’ when there are several possible outcomes or competing hypotheses) [4]. Sliders have been used for both types, as they provide a straightforward way to navigate through time. One such example is the EXP V2 environment produced in partnership with Defense Research and Development Canada, where a slider is the essential link to interacting with 3D simulation results via hand gestures [159]. Despite being intuitive and commonplace, slider-based navigation leads to issues such as change blindness. As the user moves to the next snapshot of the simulation, the whole visualization may be refreshed. Thus, users may not only miss differences from one step to the next, but experience difficulties in identifying trends over time. Pixel visualizations [128] or glyphs [242] can address this difficulty by displaying several time steps on the same space. Several studies have employed glyphs, either to determine their empirical effectiveness for temporal tasks [79] or for specific application contexts [134]. Different glyphs may be used [242] depending on the type of data (e.g., numerical as in line or star glyphs, vs categorical) or how time is visually encoded. For instance, one can use a linear time axis [106] or a cyclic time axis [228]. Comprehensive design and usage guidelines for glyphs can be found in [196, 29], although we note that some of the recommendations (e.g., avoiding asymmetry and visual ambiguity [131]) are more applicable to 3D data than the two dimensional grid considered here.

A clock glyph provides a ‘clock metaphor’ whereby the data starts at the position for noon on a clock, and is wrapped-up clockwise. Experimental evaluations found it efficient on several temporal tasks [79]. Consequently, in 2016 we proposed a visualization for CA

based on clock glyphs [89]. Intuitively, if the CA only runs for 8 time steps, then all steps can simply be wrapped up and shown all at once (Figure 6.2). However, CA usually run for much longer. We cannot subdivide the clock into as many time steps as we generated, because this would be similar to producing unreadable pie charts with tiny divisions (and it would not work beyond 360 time steps). Consequently, we face the classical problem of visual scalability, which can be addressed either by blending (i.e. mixing colors proportionally to their overlap) or aggregating (represent several data points through one color) [20]. Our solution used aggregation. The set of all time steps was divided into 8 regions, and only the most frequently occurring state was represented for each region. While we also explored dividing cells into 4 or 16 regions, feedback from our early system suggested that 8 divisions provided a good balance [89] between heavily simplifying (4 divisions) or producing potentially overwhelming renderings (16 divisions). While other parameters may influence the

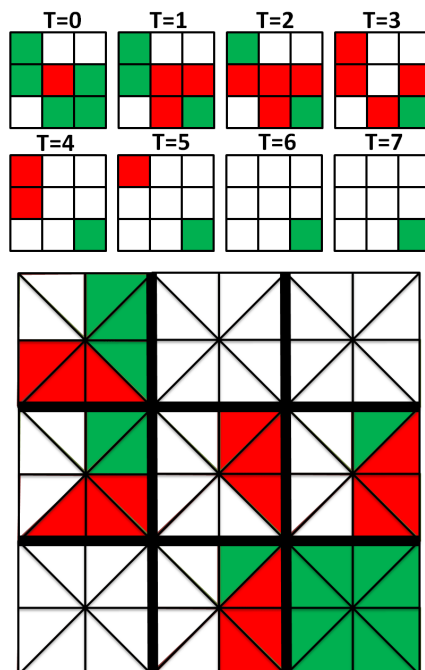


Figure 6.2: Eight consecutive time steps of the simulation (top) displayed on the same space (bottom) using a clock glyph. That is, each cell of the CA is replaced by a clock glyph showing its consecutive states [89].

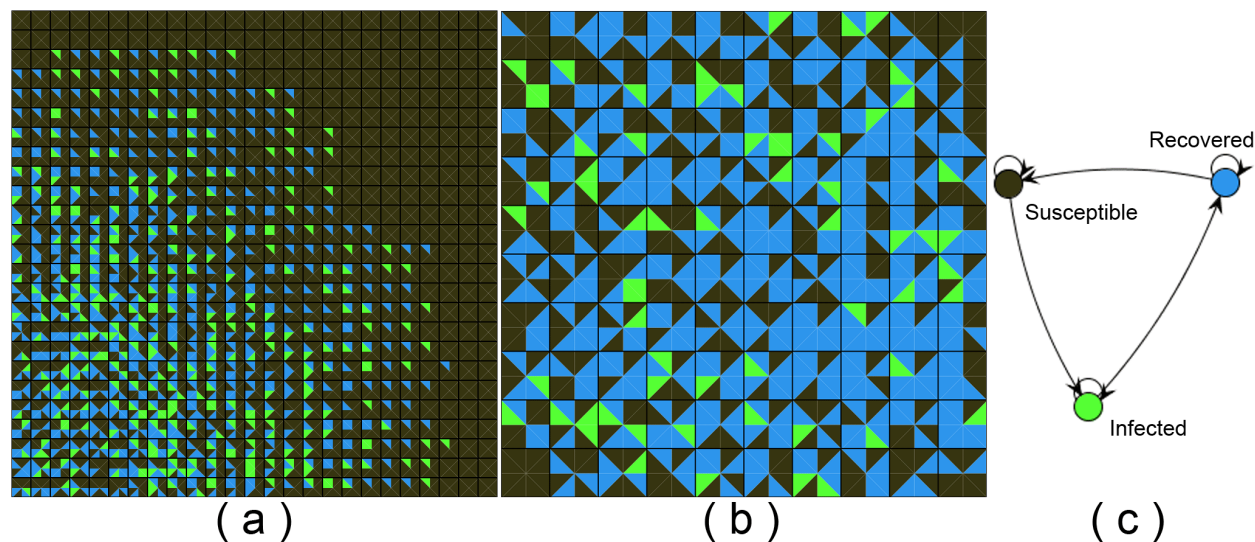


Figure 6.3: Main visualization when the simulation unfolds in a 25x25 CA over 24 time steps (a) or in a 10x10 CA over 100 time steps (b). Flow diagram for a simulation of an epidemic using the Susceptible-Infected-Recovered (SIR) model (c).

quality of the visualization (e.g., differences in size or time steps as shown in Figure 6.3a-b), these were not explored.

Given that “multiple views are particularly helpful in analyzing time-oriented data” [4], our early prototype included a flow diagram (Figure 6.3c) together with the main view (Figure 6.3a-b). The flow diagram is automatically constructed from the simulation data to show the different states (circles) and transitions (directed arrows). The two visualizations were independent from one another. Our early study engaged in discussions with modelers to examine how to progress toward a complete visualization environment. Modelers suggested using interaction schemes, so that additional linked data representations can offer details on demand or complementary information.

6.2.2 Visualizing multi-run simulation data

Multi-run simulation data refers specifically to the various datasets that might be obtained when *one* model is run with *one* set of parameter values (i.e., an ‘experiment’). It is implied that the model has some stochasticity, since running a purely deterministic model several times with the same parameter values would produce the same output. We are thus interested in visualizing a set of different simulation outputs, produced by the same experiment for the same model. If one instead wished to compare data originated from different models and experiments, then different views would be needed, as explained by Unger and Schumann [233].

Our context is to visualize simulation output produced by a cellular automata, in which cells have fixed positions throughout the simulation and may only change states. This is different from Agent-Based Models, where the agents may be moving across the space, and capturing these movements is a key objective of the visualization [8]. The different possibilities in our setting have been summarized in a review by Kehrer and Hauser:

“Multi-run data can, for example, be represented as families of data surfaces or spaghetti plots. However, it is often not practical to directly visualize such data since they can consist of multiple co-located volumes of spatio-temporal (and often multi-variate) data. Consequently, some approaches compute summary statistics from the multiple runs, which are represented by glyphs or box plots” [127]

A recent example of using glyphs is provided by Kothur *et al.* [136], in which glyphs are composed of three nested squares which represent the uncertainty of a set of ‘input squares’ in the data. The size of the median one represents the median uncertainty, the outer one denotes the lowest uncertainty, and the inner one shows the higher uncertainty. This is a typical case

of data aggregation for multi-run data, in which characteristics of the underlying data are calculated and only these characteristics are rendered. One potential caveat of computing an aggregate is that it may not preserve outliers, or infrequent states [165]. While this can be acceptable when the analyst seeks to see the overall trend, it may be less desirable in tasks such as bug detection.

6.3 Design of the visualization environment

Our environment has four components (Figure 6.5): three on the left, and one occupying most of the space (which we refer to as ‘the main’). All four are explained in this section. The **main visualization** builds on the principles from section 2.1 and our previous experience with modelers [89] to represent the whole CA at once, where each cell is divided into 8 equal segments whose color represents the main state during the corresponding part of the simulation. However, this would only address temporality, and not the multi-run facet of the data. Consequently, and in contrast with our early prototype, our main visualization *also* represents replications through the length of each segment. That is, the length of a segment is proportional to variability across replications. For example, Figure 6.4 shows the same cell over three runs of the simulation, where some of the runs do not agree about the majority state (highlighted in yellow). The extent of this disagreement is reflected in the corresponding segments’ lengths.

All three visualizations on the left implement brushing, that is, they are linked to the main visualization. Their objectives and the nature of that link are now detailed in turn, from top to bottom. The top-left visualization displays **states and transitions** which are automatically inferred from the simulation data, as in our prototype (section 6.2). While this visualization did not allow for interaction in our prototype, our newer environment links

it with the main visualization for filtering purposes. That is, clicking on a *set* of states will highlight the segment of cells that contain all of these states. For example in Figure 6.5, after users click on the green state (0) the main visualization highlights segments containing the green state. If the user then clicks on the brown state (1), the main visualization will highlight segments containing *both* the green and brown states (which will be fewer or equal to those containing just green). The new **time division** visualization (left-center) is also a filtering tool. The user selects one or more time divisions, and only these will be visible while others will be made dark. This tool thus allows to focus on patterns within the divisions of interest.

Finally, the new **prevalence** visualization (bar chart; bottom-left) offers details on demand for the division of a cell. Specifically, for the part of the simulation results covered by that cell and division, it shows the prevalence of each state *and* their standard deviation across simulation runs. For example, in Figure 6.5, the selection has on average more brown (9) than green (6) with a standard deviation of ± 2 in both cases. This visualization gives an intermediate between the aggregate of the main visualization and raw simulation runs.

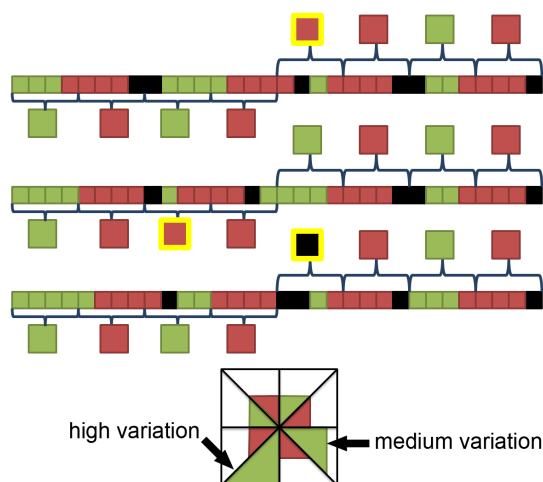


Figure 6.4: The length for each of the 8 segments encodes the variability across replications, where more width means more variability.

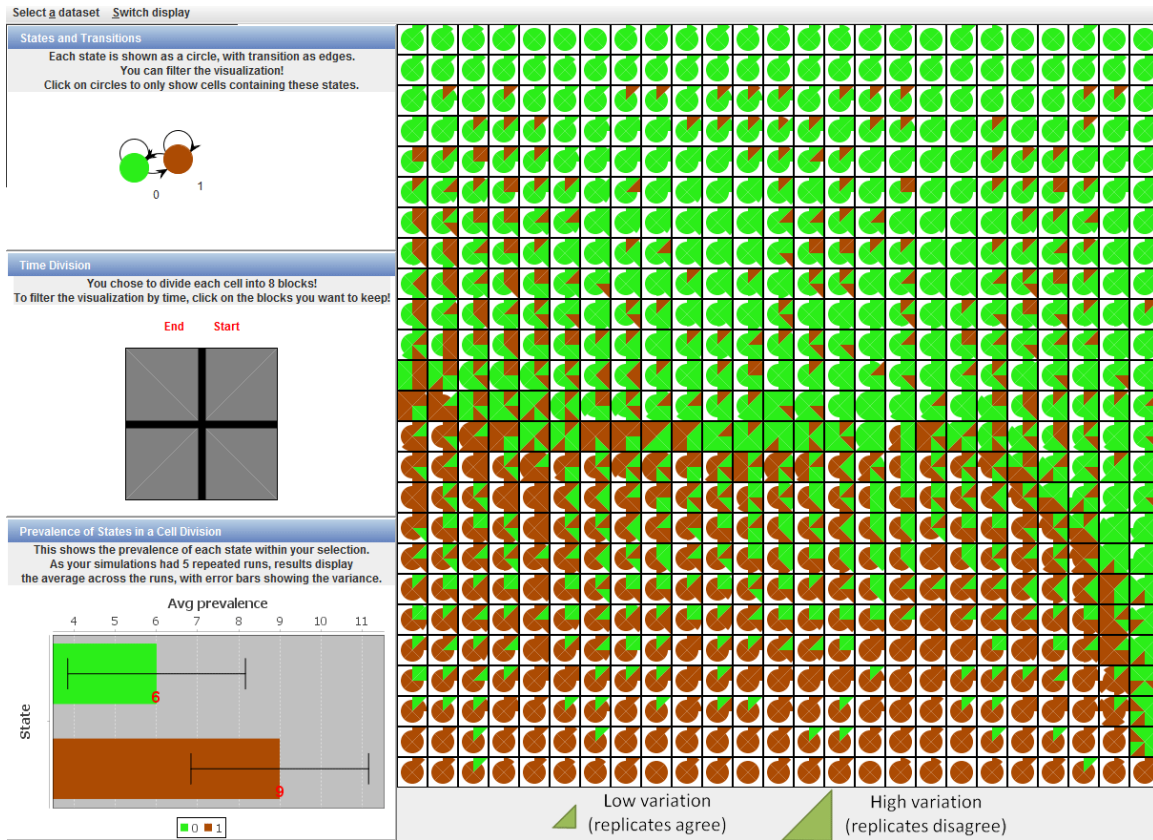


Figure 6.5: Our new visualization environment with an update main view also providing information on replications, and two additional visualizations (left-center, left-bottom) to provide filtering and details on demand, respectively. This example displays a sandpile simulation, where grains (brown) eventually settle due to gravity. Most of the uncertainty (i.e. notable differences across simulation runs) happens around the ridge of the sand dune, which is the interface between the sand and empty cells (green).

6.4 Design of the empirical evaluation

6.4.1 Objectives and overview

Our pilot study (section 6.2) gathered some of the requirements for the full environment presented here, and concluded that 8 divisions were an appropriate choice. We did not explore what factors influenced the participants' experience and whether our environment

Table 6.1: Study components, process, and recent studies employing the same process

Study Component	Where /how	Item	References
Pre-study questionnaire	At home	Gather participants' personal data (age, sex)	[143, 148, 100]
		Gather participants' experience with the visualization or the data being visualization	[143, 148, 130, 100]
Overview video		Explanations on how to use the software	[143, 140, 100, 189]
Observational study	In lab with trained facilitator	Measure reaction time for each experiment	[145, 240]
		Use the "Think-aloud" method to record comments	[143, 140, 240, 189]
Short interview		Ask about positive and negative aspects of each interface, as well as overall impressions.	[148, 240]
Post-study questionnaire	In lab	Assess the amount of distraction, and effectiveness on each interface	[143, 148, 130, 100]

supported them in better performing typical modeling tasks. Consequently, the objectives of our empirical evaluation are:

- (1) Examine whether participants could correctly and confidently achieve common modeling tasks through our proposed visualization,
- (2) Assess how participants' experiences were influenced by simulation factors, and
- (3) Explore the possibility of combining the commonly used slider-based visualization with our proposed visualization.

As our goal is to develop a visualization environment to effectively gain insight from CA with many time steps *and* multiple replications, our empirical evaluation uses two tasks corresponding to these two aspects. The first task focuses on time: users have to find the cells that mostly end the simulation in the same state as they started. In other words, the

temporal pattern of interest is a cycle. The second task focuses on replications: users have to find where there is significant variation in the simulation, both spatially (i.e. which cells) and temporally (i.e. which divisions of cells). In other words, users need to find areas where replicates disagree a lot on the simulation outcome. Both tasks are common in modeling. The first task is routinely carried out as we need to find if, and where, the simulation stabilizes. The second task typically serves to assess whether more simulation runs are needed to get a clear consensus on the outcome.

We also had to the slider-based visualization in our experiments, not only to see its potential for combined use (objective 3) but also to serve as a referential for performances (objective 1). Indeed, we wanted to observe whether using of our proposed environment would take more time than users would normally spend via the slider-based visualization. Consequently, we also programmed this visualization within our software (Figure 6.6). As explained in section 6.2, the slider-based visualization shows the exact state of all cells for one time step, and allows the user to refresh the whole screen via a slider controlling the time step currently displayed. To allow user to navigate several replications within a slider-based visualization, we provide a left panel in which users can simply click on what replication to visualize. In sum, our two tasks were performed using both our proposed visualization and the ‘simpler’, slider-based visualization.

For objective 2 (role of simulation factors on visualization performance), we included three simulation factors as potentially impacting the participants’ experiences: the length of the simulation, the number of replicas, and the phenomenon being simulated. Doing a simple ‘sensitivity analysis’ that varies one factor while fixing the others has been described as statistically inefficient, and does not account for interactions [118]. For example, it would not reveal whether users find it harder to navigate CA that have many time steps *and* many replicas. Consequently, we used a Design of Experiments (DoE) known as a 2^k factorial design, which captures interactions. We have a total of 4 factors (phenomenon, replicas,

time steps, common vs. new visualization) thus there are $2^4 = 16$ combinations. We asked participants to perform both tasks, for all 16 combinations. We will refer to these combinations as our *experiments* from here on. The order in which experiments are performed may create bias. For example, imagine that a participant is given a simulation of a spread (phenomenon) over 25 time steps with 50 replications and has to find cycles using our proposed visualization. Then, if the participant was given the same phenomenon, same time steps and same replications but using the simpler visualization, there would be a potential for the answer to reflect the insight actually gained using the proposed visualization. This effect may also be reciprocal, whereby insight from the simpler visualization may be used to solve the problem using the proposed visualization. Consequently, we randomized the order of the experiments for each participant.

6.4.2 Detailed Study Design

The previous section provided a brief overview of our two objectives, the two tasks that they translate to, and the 16 experiments to perform for each task. Table 6.1 lists what participants had to do, what part of our study it belonged to, and where/how they had to do it. In order to highlight common practices and allow for comparison of study designs, we also provide references to other studies incorporating the same aspects in their experimental designs. The overall script for this study can be obtained at <https://osf.io/brjqk> under *Experimental Setup*, including the template of emails sent to participants as well as subject information and consent.

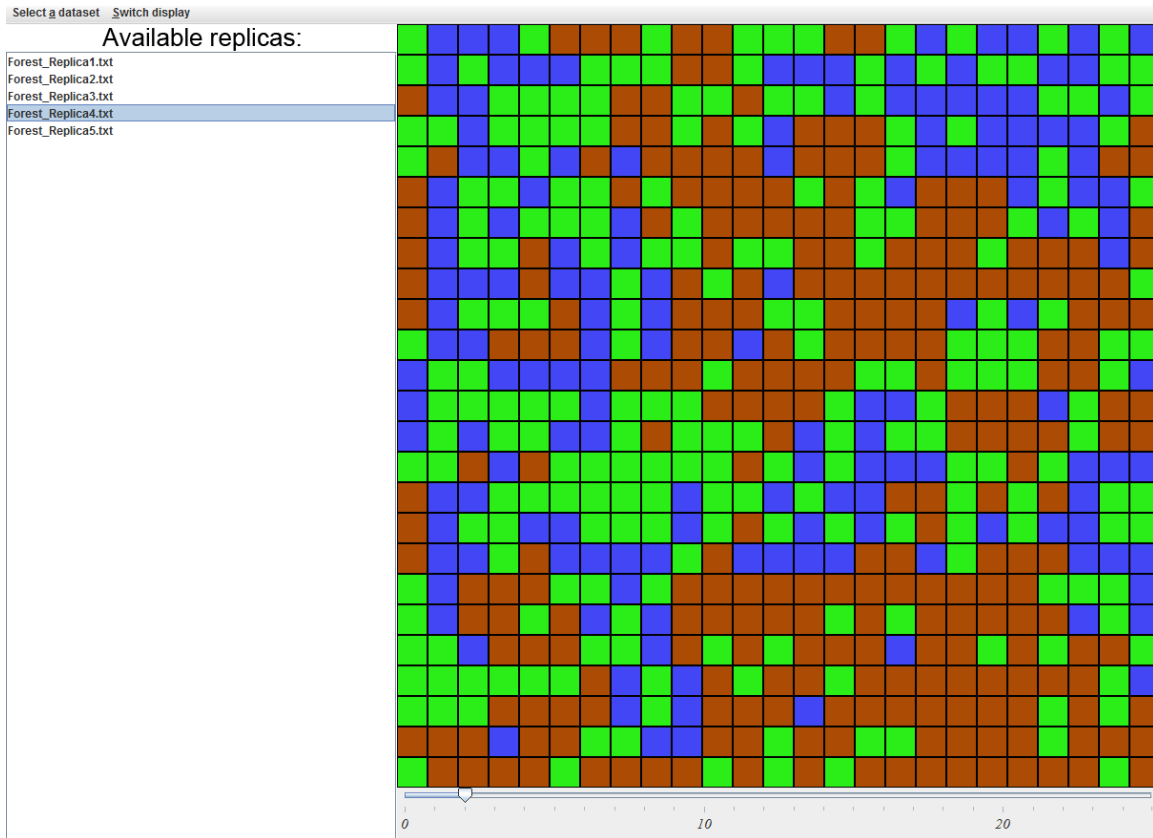


Figure 6.6: Simple visualization environment allowing to select a simulation run (left) and move through its time steps using a slide (bottom).

6.4.2.1 Pre-study questionnaire and overview video

A pre-study email was sent to all participants, asking them to complete a pre-study questionnaire, available at <http://bit.do/NIUpresurvey>, and to watch a 22 minutes overview video, available at <https://www.youtube.com/watch?v=s5jQ2Z6iIf8>. The video explained the basics of cellular automata, the different visualizations of our software, and briefly stated what tasks participants may be asked to complete. The questionnaire asked for the first name, last name, participant ID, age, and gender. It also assessed previous experience in terms of visualization, modeling and simulation, or visualization data from simulations. One day prior to the observational component, participants received an email reminder about

their appointment and, if they did not complete the questionnaire by then, this was specifically mentioned in their remainder. To check whether participants did watch the video, they were asked questions about cellular automata upon meeting the facilitator in the lab for the observational study. All answers were judged satisfactory, thus none of the participants was asked to watch the video again.

6.4.2.2 Observational study and short interview

The first author served as facilitator for this study. Upon meeting with participants in the lab, and after checking whether they watched the video, the facilitator emphasized the “think aloud” protocol, asked participants to behave naturally, and informed them that overall impressions would be discussed *after* experiments would be completed. The audio and video recording started by confirming whether each participant agreed and consented to participate in the study. In order to limit sources of bias [90], the randomized ordering of the experiments was not known *a priori* to the facilitator. Instead, a random order for the experiments was generated as the recording started (the code is available at <https://osf.io/brjgk> under *Experimental Setup*).

For the 2^k design of experiments that we employed, our four variables were limited to binary values. The phenomena modeled (i.e. what the simulations were depicting) consisted of a sandpile and a burning forest. The idea of a sandpile is that grains are dropped from the top, go down under the effect of gravity, and possibly cause an avalanche when they accumulate in unstable ways. That is, the CA shows the system from a side view, as if one was taking a slice of how a dune builds up (Figure 6.5). The specific sandpile model that we used to create the data is from Athanassopoulos and colleagues and uses a probability p (set here to $p = 0.5$) to deal with the situation where two grains of sands are above two

cells [10]. The principles of a burning forest (also known as a ‘fire spread’) is that the system is composed of clearings, trees, and burning trees. The fire originates at one tree, and spreads through two rules: either one tree has a neighbour that is burning and catches fire from it, or a tree spontaneously starts to burn with probability $p = 0.001$ (representing what may be carried across the system by the wind) [126, 5]. In contrast with the sideways perspective in the sandpile, a burning forest looks at the system ‘from above’, as if one had an aerial view of the forest (Figure 6.6). The two models also represent broadly different types of phenomena, between a mixing model (sandpile) or the dynamics of a spread (burning forest).

The three other variables were the type of visualization (either the simple one with the slider or our proposed one), the number of time steps (few at 25 or more at 50), and the number of replicas (few at 5 or more at 50). The rationale for the numbers is that we expect ‘few’ to be manageable by hand, and ‘more’ to start becoming overwhelming to get a correct answer. For example, one may browse through 5 replicas to see if they agree, but browsing through 50 may lead to a different outcome.

In total, the observational component generally lasted one hour. Then, a short interview took place. Participants were asked for overall thoughts on the simple visualization, and then on the full visualization. They were prompted to reflect on which task they found harder, and if anything could be done to either visualization that would be helpful. The interview component generally took about 10 minutes.

6.4.2.3 Post-study questionnaire

Immediately after the interview, participants completed the post-study questionnaire available at <http://bit.do/NIUdoneSurvey>. As in [100], the questionnaire drew from the NASA Task Load Index (NASA-TLX) questions for task evaluation. We did not include the

questions ‘How physically demanding was the task?’ and ‘How hurried or rushed was the pace of the task?’ as neither were applicable. All other questions were included and touched on mental demand, performance, effort, and frustration. In addition, we asked participants to share whether anything distracted them during the experiments, and asked for feedback about the visualizations (which was already also part of the interview component).

6.4.3 Participants and Apparatus

The rendered images were generated on a professional workstation with 2 Intel Xeon processors (E5-2650 v3 at 2.3 GHz). They were displayed on an Asus VS248H-P 24-inch LCD monitor, using a full screen window (1920 by 1080 pixels). Participants sat approximately 75 cm away from the monitor. Participants were not fixated into a chin rest and were instructed to behave naturally. Our participants were students from the computer science department at Northern Illinois University. Participants received neither financial compensation nor extra grades. We video-recorded the sessions using FlashBack Express version 5.24.0.4208.

We recruited a total of 16 participants, aged 18 to 31 (mean 24). Only two were female, reflecting the skewed demographics for computer science students at our institution. Seven participants (43.75%) reported having prior experience in simulation (Figure 6.7), while only one reported formal training on visualizations. Four participants were excluded from the analysis (Figure 6.7): three left before completing all experiments in the observational study, and one had technical issues (unresponsive keyboard). After excluding these participants, the mean age remained 24, there were still 2 females, half of the participants reported having prior experience in simulation, and there was still one with formal training on visualizations.

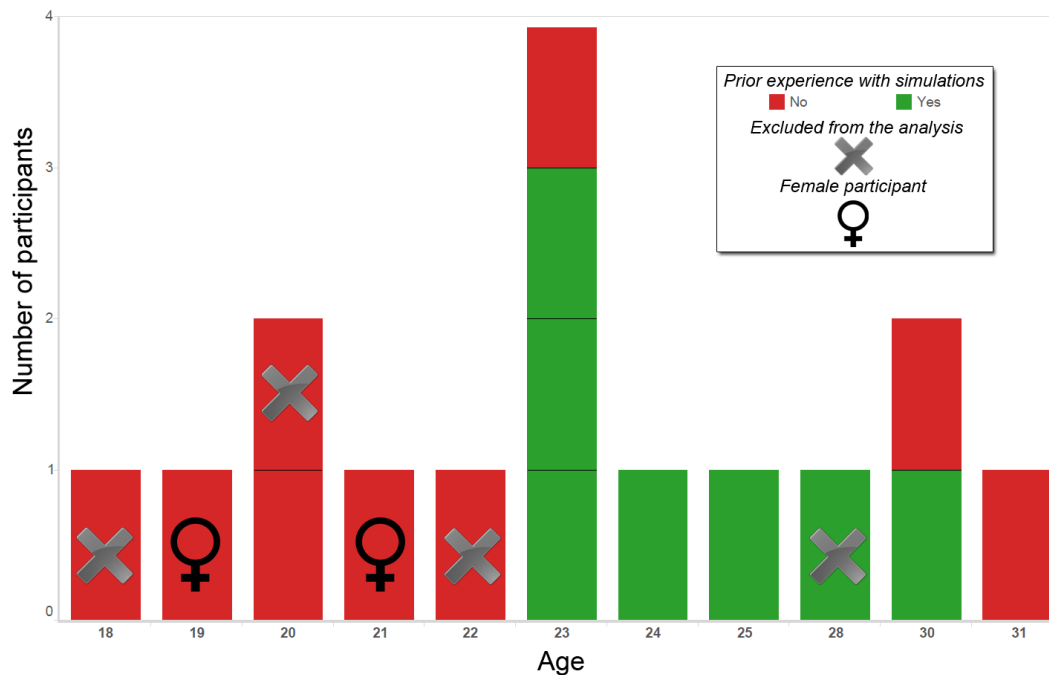


Figure 6.7: Age, gender, and prior experience of participants. Four were excluded from the analysis.

6.5 Results

6.5.1 Overview

The compressed video recordings of the computer screen for all 16 participants are available at <https://osf.io/brjqk> under *Experiments (videos)*. The pre- and post-surveys as well as all data measured from the video are available as series of spreadsheets under *Experimental Results*. The same folder contains the verbatim transcripts of the participants' interview. The participants' experience for each experiment was primarily measured through 3 variables:

- Confidence, using the categories 'yes', 'no' and 'maybe' based on the participant's verbal feedback. For example, "it would be hard to say" and "I really have no way

to tell” were coded as ‘no’, while “it’s obvious that...” and “ I am pretty confident that...” were coded as ‘yes’, and “It seems like something...” as well as “It might be...” were coded as ‘maybe’.

- Total number of errors, out of 4 possible errors. Three of the errors consisted of (incorrectly) generalizing by looking at (i) entire regions rather than cells, (ii) a few replicas, or (iii) a non-systematic sample of time steps¹. For example, if there are 25 time steps and 50 replicas, and the user decides to randomly look at time steps 2–5–40 for only three of the replicas, then there are two errors. The fourth error is the failure to accomplish the experiment, either by being unable to find any of the cells requested, or by just guessing the cells.
- Total time spent, from the moment the dataset was loaded by the facilitator, to the moment when the participant concluded.

Our analysis is available online under the *Analysis* folder. The next section shows the results of the quantitative analysis based on the three metrics mentioned above. Then, we provide findings from the qualitative analysis based on the verbatim transcripts. Finally, we discuss the outcome of the post-survey.

6.5.2 Quantitative analysis of the observational study

6.5.2.1 What affects the participants’ experience?

We have four individual factors (phenomenon modeled, number of time steps, number of replications, visualization employed) and six possible pairwise combination of factors. We

¹We counted errors of time steps for task 2 but not for task 1, because looking only at the first and last time steps suffices to correctly conclude whether a cell’s final state was the same as its initial state.

performed a factorial analysis to identify the key contributors to our three metrics, for both tasks. In short, a factorial analysis shows how much of the variance is caused by each of the 10 (4 individual and 6 pairs) potential contributors. The larger the percentage, and the more important the contribution. Note that the ‘importance’ only tells us that it affects the results, but not in which way. Complementary analyses (after the factorial analyses) thus explore what creates a *better* experience.

Results are shown in Figure 6.8 and are now described, from top (confidence) to bottom (reaction time). In both tasks, confidence is primarily driven by the type of visualization (82.75% for task 1 and 76.90% for task 2) and is also affected by the phenomenon modeled (9.59% in task 1 and 17.08% in task 2). All other contributors are negligible as they account for less than 3% of the total variance. The type of visualization plays an even bigger role in the number of errors (94.93% for task 1 and 93.86% for task 1), with a very small role for replicas either directly or in interaction with the type of visualization. The picture is more nuanced when it comes to the reaction time. On task 1, participants were highly impacted by the type of visualization (73.43%), and moderately by the interaction of replications and time steps (14.76%). On task 2, the visualization had a negligible impact, and instead the key driver was the number of replications either as sole factor (25.46%) or in combination with other factors (49.75% of the variance through interactions). Note that the reaction time in task 2 has a much large error (9.70%) than in any other task or metric, suggesting that some of the variation was driven by differences between participants. Results for the reaction time in task 2 may thus be interpreted less precisely than for other metrics or for task 1.

Through previous reviews on visualizing simulation data, it was expected that performances would be higher using a visualization specifically designed for spatio-temporal, multi-run data. These results shed light on the scale of the improvement, rather than its existence. Our proposed visualization makes a significant difference on helping participants to either

Category	Factors	Task 1	Task 2
Confidence	Errors	• 0.58	• 0.66
	Phenomena & replicas	• 0.15	• 2.94
	Phenomena & time steps	• 2.40	• 0.50
	Phenomena & vis	• 1.35	• 0.82
	Phenomenon	■ 9.59	■ 17.08
	Replicas	• 1.84	• 0.25
	Replicas & time steps	• 0.94	• 0.01
	Replicas & vis	• 0.34	• 0.25
	Time steps	• 0.04	• 0.09
	Time steps & vis	• 0.04	• 0.50
	Vis.	■ 82.75	■ 76.90
	Errors	Errors	• 0.08
Phenomena & replicas		• 0.00	• 0.08
Phenomena & time steps		• 0.10	• 0.01
Phenomena & vis		• 0.02	• 0.08
Phenomenon		• 0.00	• 0.08
Replicas		• 2.16	• 3.22
Replicas & time steps		• 0.10	• 0.08
Replicas & vis		• 2.42	• 1.49
Time steps		• 0.10	• 1.01
Time steps & vis		• 0.10	• 0.02
Vis.		■ 94.93	■ 93.86
Reaction Time		Errors	■ 3.88
	Phenomena & replicas	• 0.10	■ 20.75
	Phenomena & time steps	■ 4.46	■ 3.85
	Phenomena & vis	• 2.28	■ 6.91
	Phenomenon	• 0.00	• 0.00
	Replicas	• 0.16	■ 25.46
	Replicas & time steps	■ 14.76	■ 21.72
	Replicas & vis	• 0.21	■ 7.28
	Time steps	• 0.16	• 1.18
	Time steps & vis	• 0.55	■ 2.06
	Vis.	■ 73.43	• 1.10

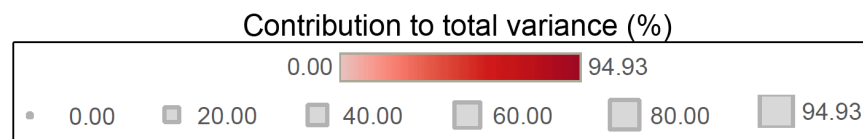


Figure 6.8: Contribution of independent factors and pairs of factors to the total variance, for each task, and for each metric (i.e. participants' time, participants' confidence, and participants' error level).

avoid errors or be confident about their findings. As shall be discussed in the qualitative analysis, familiar phenomena may also boost a participant's confidence. For example, in sand dunes, participants may be certain that a grain of sand at the bottom of the screen will stay there for all future time steps because of gravity. However, dealing with a more 'familiar' phenomenon does not change the number of errors. Finally, the reaction time should be interpreted with caution: participants overwhelmed with data and feeling incapable to navigate it may quickly generalize and move on, thus ending an experiment early. The reaction time thus also captures whether participants actually felt able to fully perform the task.

6.5.2.2 How are participants affected by the visualization?

The previous section showed that the type of visualization significantly impacts participants' confidence and errors, while its impact on time is more nuanced. We now focus on how results on the three metrics are improved or worsened by the type of visualization. The metrics are examined in the same order as in the previous section.

The distribution of confidence score for both visualizations and both tasks is shown in Figure 6.9, with the number of participants as y-axis and the number of times an answer was used as x-axis. For example, the top-left bar shows that 2 participants had the answer 'no' once. These distributions show that participants were uncertain using the simpler visualization as we observed that the 'no' confidence dominates. Furthermore, all participants (n=12) did not feel confident in at least one experiment for task 1, while the majority did not feel confident (n=9) in at least one experiment for task 2. The opposite results are found using the proposed visualization where we observe that 'yes' dominates, and only 1 (for task 1) or 2 (for task 2) participants did not feel confident in an experiment. Note that the factorial analysis from the previous section showed that the type of visualization was an



Figure 6.9: Distribution of participant's confidence in their answers for both the simpler (top) and proposed (bottom) visualization, on the first task (left) and second task (right).

important contributor but not the only one, given that the phenomenon modeled played a role too. In other words, just changing the visualization is not powerful enough to create perfect results across experiments: the proposed visualization shows more confidence, but many participants felt uncertain ('maybe') at least once. Further improving the results may thus require adapting the visualization to the phenomenon modeled.

The distribution of errors for both visualizations and both tasks is shown in Figure 6.10. The *absence* of an error (i.e. a perfect answer) is shown as having committed 0 errors on the y-axis. Using the simpler visualization for task 1, we see that all (n=12) participants were unable to perform some of the experiments. Many also looked at only a few of the replicas, thus assuming what other simulation runs would do without looking at them. They also commonly generalized from regions rather than tracking individual cells. As concluded in the factorial analysis, the type of visualization has a massive impact on errors. Using

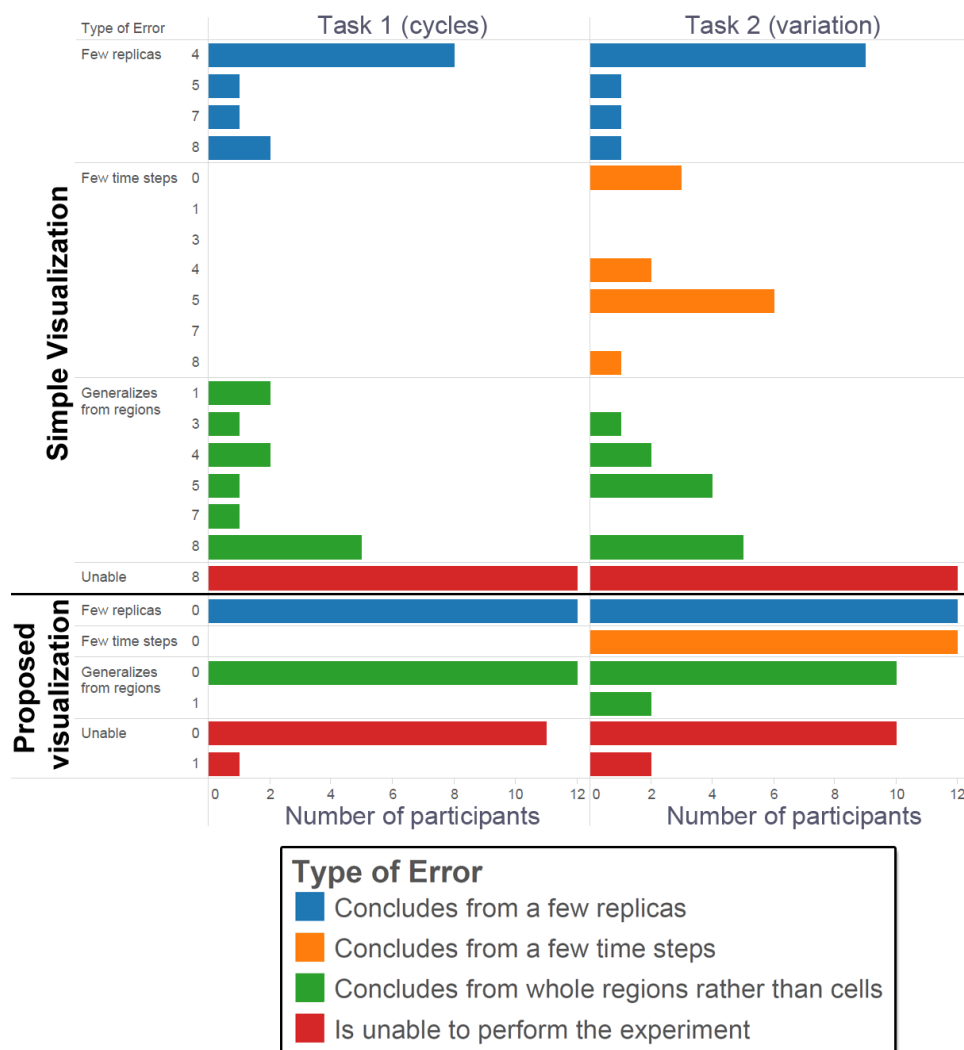


Figure 6.10: Distribution of participant's errors in their answers for both visualizations and both tasks. The y-axis shows the number of errors (where 0 is the absence of any error) and the x-axis shows the number of participants having committed these many errors.

our proposed visualization, all participants but one were able to perform the experiments. Their conclusions correspond to what was found in most simulation runs ($n=12$) and they were able to track individual cells ($n=12$). Task 2 also included the un-systematic browsing of time steps as an error. Findings are similar to task 1: all participants had difficulties completing some experiments using the simple visualization, while all but two were able to perform perfectly using the proposed visualization. Errors are most commonly made by

ignoring replicas via the simpler visualization, which did not occur using the proposed one. Un-systematically sampling time steps also occurred with the simple visualization but not with the proposed one. Generalizing from region occurs using both the simple and proposed visualization, although to a much lower extent with the proposed visualization. Overall, the proposed visualization leads to a significant reduction in errors although a handful still occur.

Finally, we investigated how much time participants spent using our proposed visualization compared to the simpler one. Differences were small and going in different directions from one task to the other, in line with the factorial analysis showing that the type of visualization impacts time much less than confidence or errors. For task 1, participants spent an average of 10.5% *less* time with the proposed visualization than the simpler one. For task 2, participants spent an average of 8.3% *more* time using our proposed visualization.

6.5.3 Qualitative analysis of the interview

We performed a thematic analysis to reveal the main themes from the interview component. As this is a form of interpretive research, biases can occur. We considered that having directly spoken to the participants or knowing who they were when reading the transcripts could bias the interpretation. Consequently, the first author facilitated the experiments, created verbatim transcripts, and anonymized them. The analysis was then performed by the second author on the anonymized transcripts. Three themes emerged.

First, participants discussed how the two visualizations could actually be *combined*. The fact that combination was a theme across interviews, rather than an anecdotal mention, suggests that there is potential to use both our proposed visualization and the classic one. The specific joint use was described by a participant as follows:

“I would probably start with... with this view [points at our proposed visualization], and you know try to make conclusions and assumptions, and then if there was some use-cases or one-off-cases that seemed very unique I would probably try to pull up the same dataset using the, the more simpler visualization, and try to see... and try to drill down into the specific timesteps and see if I can identify what’s going on.”

Similarly, another participant pointed out that to “be able to switch between the two is very key”, which was echoed by a third as wishing to have a shortcut to easily switch “instead of activating one, deactivating the other”. This suggests that future work should explore the design of a more integrated visualization.

Second, participants mentioned their perceived *obstacles* to navigate the data across time steps and replicas. This theme was expected, as such obstacles prompted the design of this study. Some perceived that obstacles were traded between visualizations, which again provides contextualization to a possible joint use of the two. The slider-based one required many interactions and memorization (high cognitive load) while the proposed one had a steeper learning curve. In the words of a participant, it is “a lot more simpler to use just an original [slider-based] visualization, it is just not as efficient.”

Third, participants shared their experiences on functions of the proposed visualization *other than the main view*. It was noticeable that the main view was discussed very little beyond its ability at aggregating information, and that instead participants focused on other tools or interactions. They occasionally started suggesting possibilities and then found that they already existed. For example, they realized that they could zoom while describing how they would do it, or suggested the ability to click on a tool and noticed that it was clickable. While it is reassuring that participants discover interactions naturally when looking for them, it also suggests the need for a tutorial to be more aware of what can be already be achieved.

Several participants pointed out that they did not use the States and Transitions tool (top-left), which may be partly due to ignoring how it filters, or because there were only few states and transitions in both phenomena modeled. It would thus be of interest to offer tools depending on characteristics of the simulation. For example, the States and Transitions may be displayed when there are many such states and transitions, while the prevalence graph (bottom-left) is most useful when there are variations.

6.5.4 Mixed methods analysis of the post-survey

The post-survey asked participants if they were distracted during the study. None reported an external distraction. While several participants shared additional impressions about the visualizations, these were also stated in the interview component and would not change the outcome of the thematic analysis. Table 6.2 provides the median scores for questions judging the degree of failure, hardness, stress, and mental demand to accomplish both tasks using the two visualizations. Scores range from 1 (best) to 5 (worst). For example from top to bottom, 1 stands for perfect success, very low work, very low stress/irritation, and very low mental demand. We observe that the proposed visualization systematically produces better scores.

These results are in line with the analysis in section 6.5.2.2 regarding the effect of the two visualizations on task performance. The thematic analysis in section 6.5.3 suggested that both visualizations could cause significant mental demands. On one hand, the slider-based version creates a high cognitive load by having to mentally keep track of states which participants avoided simply by not looking at the data and instead forming assumptions (“I don’t know how about other replicas, but I am assuming they’re similar”). On the other hand, the proposed one requires navigating multiple tools and looking at much more of the

Table 6.2: Results of the NASA Task Load Index (NASA-TLX)

Metric	Task	Proposed vis.	Slider-based vis.
Failure	1	2	4
	2	2	4
Hardness	1	2	4
	2	1	5
Stressed	1	2	4
	2	1	5
Mental demand	1	1	4
	2	2	5

data at once (“it is busy in there, there is a lot going on”). The fact that participants see the slider-based visualization as causing a higher mental demand thus suggests that they saw memorization as a much larger issue.

6.6 Discussion

Inspecting simulation output can be challenging when the simulation lasts for many steps or uses multiple runs. In this paper, we focused on the output generated by 2-dimensional cellular automata with square cells. This modeling approach is used in a wide variety of fields ranging from geophysics [230] to biology [188] and occasionally social phenomena [180]. In a pilot study, we examined whether a clock glyph could be used to provide a visualization environment in this context [89]. Modelers suggested that the idea had potential, and contributed to identifying important visualization parameters such as the number of divisions for the clock. The present study then developed a more complete system including interactive tools (e.g., to provide filtering or details on demand) and evaluated it with 16 participants, of whom 12 completed the protocol. Our evaluation examined their experience when performing two typical modeling tasks: identifying temporal trends (with a cell ending as it started) or assessing variability (where different simulation runs do not agree on the

output). We measured their experience through their confidence, the number of errors, and the time they took. A 2^k factorial design of experiment allowed us to assess the impact of several parameters on these three metrics, both as independent parameters and through second-order interactions (i.e. pairs of parameters). While previous studies suggested that there should be some improvement when using our multi-faceted visualization instead of the classic slider-based alternative [127], we found the difference to be significant as the choice of visualization made a major difference on confidence and errors. The phenomenon being modeled also impacted the participants' confidence, but did not play a role in other metrics. Using our proposed visualization instead of the classic resulted in a small difference in time (of up to 10%), although the direction of this change depended on the task. That is, participants were faster with our proposed visualization (by 10%) to identify temporal trends but slower to find variation (by 8%). A thematic analysis revealed that participants would like to combine the two visualizations, which would allow them to spot potential trends using the aggregation offered by our proposed visualization, and then confirm the trend by using the 'microscopic' view of the slider-based visualization. Both visualizations create mental demands. In the words of a participant,

“The more complex visualization is definitely better and then tools helps to compare the first and the last state. But it is harder to see... to actually visualize what is happening.”

Conversely, the slider-based visualization created a high cognitive load by mentally tracking states across time or simulation runs. In the post-study questionnaire, participants expressed that the proposed visualization exerted low or very low mental demand, while the slider-based equivalent had a high or very high mental demand. Consequently, memorization was much more of a barrier than navigating our new aggregate outputs and its interactive tools. In addition, providing aggregated information about the whole dataset at once provides a

systematic way to explore the data. When participants explored it without our aggregation, they were often un-systematic, and (incorrectly) assumed that what they would see in a simulation run would just be the same in one or all others.

Our evaluation was limited to two typical tasks. Other tasks may require the development of additional interactive tools. For example, modelers may want to ensure that some transitions never occurred (when debugging). These transitions may be more complex than forbidding a dead cell to become alive again: for example, there may be a minimum time requirement. While motif mining is a well-studied approach to find sequences in time series, with multiple supporting visualizations (see Hao *et al.* for a recent example [104]), limited research has been performed on motif mining in data from cellular automata. Identifying efficient ways to express and display patterns in the context of cellular automata (particularly with multiple runs) would thus be beneficial.

Two other avenues are of particular interest for future work. First, our evaluation found a need to identify which interactive tools to use depending on the context. Ma's suggested integration of machine learning and visualization could be one step to realize this objective [150]. Specifically, we can collect information on an analysis session and characteristics of the data, and then use machine learning to automatically adapt the visualization to the person, task, and dataset. The information can be stored in a structure format, on which different data mining algorithms can straightforwardly be applied [52, 53]. The main difficulty is thus to collect enough data to create a comprehensive repository, and to extract the right information from a session and dataset.

Second, our visualization focuses on the temporal and multi-run dimensions, while limiting the size of the model to a small or medium number of cells. To broaden the application of this work, it is necessary to accommodate much larger models. This is an important theme for the modeling and simulation community. In a 2013 panel, 'big simulation' was presented as a grand challenge [224], and many studies have recently been published on interactive visu-

alizations for data produced by ‘big’ or ‘large-scale’ simulations. For example, in 2016 alone, visualizations were proposed for the CODES simulation built on the large-scale Rensselaer’s Optimistic Simulation System (ROSS) [198], for an ocean simulation built using the Model for Prediction Across Scales (MPAS) [169], or for molecular dynamics simulations performed via the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [194]. Cellular Automata can be produce big simulations as well. For instance, cells in a fungal network may be at the micrometer (because many important interactions in the organism happen in the 10-15 μm range), but the overall network may cover kilometers [58], which leads to a massive number of cells. Our suggestion is to combine our work with approaches for dimensionality reduction. While several exist, not all can work in our context. For instance, Self-Organizing Maps (SOMs) have been used for grouping and arranging spatial distributions and temporal variation profiles according to their similarity [9]. However, approaches such as SOMs [9] and others [60] that rely on re-arranging elements would be difficult in the case of a CA where the position of each cell within the grid already has an important meaning to modelers. Potential approaches would have to be based on grouping/clustering cells, rather than re-arranging them. Many such approaches exist, including hierarchical clustering techniques from Bordoloi *et al.* [28], time-varying partitioning [208], or Janicke’s information theoretic method which identifies regions with different temporal behavior [120]. Future studies could thus combine these methods and empirically evaluate their effectiveness for selected modeling tasks.

6.7 Conclusion

Our proposed visualization efficiently supported users in performing key modeling tasks, by aggregating information (across time steps and simulation runs) as well as through inter-

active tools for filtering and details on demand. Additional work is needed to also aggregate many elements (via multi-resolution techniques), find specific patterns (by adapting motif mining to cellular automaton), or tailor the environment to features found in the simulation run (such as the number of states and transitions).

CHAPTER 7

NAVIGATING COMPLEX SYSTEMS FOR POLICYMAKING USING SIMPLE SOFTWARE TOOLS

The third aim of this thesis (realized in the previous two chapters) was to develop novel visualizations for simulation data. In this chapter, we posit that simulation models themselves (rather than the data they generate) would also benefit from visualizations. This is important for policymaking, to support a systems science approach that goes beyond thinking of disconnected ‘inputs’ and ‘outputs’ to instead emphasize interrelatedness and loops. Similarly to chapter 5, we present a new environment, emphasize its key functionalities, their relevance to policymaking, and why they have not been met by other software packages.

All of this chapter was submitted in the following peer-reviewed book chapter:

- PJ Giabbanelli & M Baniukiewicz. Navigating complex systems for policymaking using simple software tools. In *Advanced Data Analytics in Health*, VK Mago, PJ Giabbanelli, E Papageorgiou (eds.), Springer, 2018.

My contributions consisted of (i) finding and reviewing the functionality of software for argumentation, visualization, and modeling (Table 7.1); (ii) leading the development of the software based on an earlier prototype [86]; and (iii) creating the data used for demonstrations. Functionalities and design principles were produced by PJ Giabbanelli, while the supporting online video was made by N Rosso.

7.1 Introduction

“As we enter an era marked by more complex drivers of population health and by diseases with multifactorial roots [...], it will be more useful to have in our population health armamentarium the capacity to model the potential impacts of different manipulations of the multiple factors that produce health.” Galea et al. [82]

Studies at the intersection of systems science and population health have demonstrated the usefulness of computational models for conditions driven by multiple interacting factors. For example, agent-based models of obesity have allowed to reconcile peer influences on food and physical activity behaviors with environmental (e.g. built environment) and individual (e.g. stress and depression) factors [249, 88, 92]. Such models are particularly useful for policymaking, as they allow to evaluate the health impacts of complex interventions [41], or can answer ‘what-if’ (also called ‘what happens if’) scenarios [82] in which policymakers explore the consequence of several policy levers either independently or in a synergistic fashion. Given the support that computational models can offer, they have become an increasingly popular tool in public health. In the case of obesity, there were only a handful of models in the 2000s [13, 116] but the recent years have seen so many models that they were the subject of several dedicated reviews [144, 207].

The emphasis in a computational model has historically been to capture the salient characteristics of a phenomenon. For instance, our early model of obesity focused on environmental and peer influences on changes in body weight. It thus had to capture how peers and the environment come together in affecting one’s physical and food behaviors, which in turn can affect one’s weight depending on physiological factors (e.g. metabolic rate) [88]. Policy levers are limited to a subset of variables in the model: the gender or

metabolism of individuals cannot be changed by policies, but social norms may be amenable to changes [193, 218]. Running ‘what-if’ scenarios then consists of assigning different values to these variables, either independently (which is common yet statistically inefficient) or using Design of Experiments techniques [118]. We recently discussed two issues with this approach to ‘what-if’ scenarios [91]. First, policy levers may not be independent. For example, a high density of restaurants might lead to increased market competition in part through larger portion sizes. A ‘what-if’ scenario may artificially set values for density of restaurants and portion size, without realizing that the value taken by one affects the other. Second, a policy is not merely an abstract concept: it eventually has to be implemented by coordinating across sectors or jurisdictional boundaries. This involves many stakeholders, which have to work at different time scales. Consequently, inputs to a model may not be directly set to a value but instead change gradually, at different speeds for different inputs. In sum, rather than freely manipulating a collection of disparate inputs, policy-relevant models should include the essential interactions between inputs. This can be achieved by adding another layer to models, which captures the interactions between inputs for policy purposes. Any ‘what-if’ scenario would be done in this layer, and then passed to the model [91].

There are three key steps in driving a policy-relevant model through a layer accounting for inter-dependencies between the model’s inputs. First, we need the layer itself, that is, the set of relationships between inputs along with relevant metadata (e.g., causal strength or time scales). Many suitable layers may already exist, but there is currently no comprehensive database or repository where researchers or policymakers would find them for a population health context. In obesity research, many projects have now produced comprehensive ‘maps’ of the relationships between factors related to weight or well-being. It is difficult to use one of these maps as one may not know where to find them, cannot compare them, or may not know the context in which they were produced. In addition, maps were produced using different techniques, which means that they may not have the same type of metadata, or

provide the same level of trust in the data. Causal networks are formed of directed weighted relationships [235, 63], to which Fuzzy Cognitive Maps add an inference engine allowing to simulate the consequences of these relationships and thus test their validity [95, 97]. System Dynamics go even beyond, by including time effects and lags [238, 222].

Second, the target audience of a policy-relevant model is not the modeler(s) but the participants, such as policymaker(s) and/or community member(s). Their views are essential to select the inputs that are modified in ‘what-if’ scenarios, and to identify relevant outputs. When the interactions between inputs start to be accounted for, this information can be provided to participants to support their decision-making processes. However, a set of interactions defines a network, and handling networks can be a much more demanding task for participants than handling seemingly disconnected inputs. Our usability study with policymakers in British Columbia (Canada) found that navigating even a medium-size network could be a very challenging task [86]. The Foresight Obesity Map (Figure 7.1) exemplifies the difficulty of using a network for the target audience. As pointed out in Hall *et al.* [102] (emphasis added), “the *complexity* of the obesity epidemic is graphically illustrated by the *web* of interacting variables”. This is echoed by Siokou et al [213]:

“With 100 or so causal factors, and 300 or more connections linking each cause to one or more of the others, the Foresight diagram is a complicated, almost incomprehensible web of interconnectedness that depicts the drivers of obesity prevalence and the ways in which they depend on each other. The diagram is brilliantly useful in demonstrating the complexity of factors driving the current obesity trend, but the scale and number of interactions in the diagram make it difficult to see how one might use it in any practical way to develop systemic approaches to obesity prevention.”

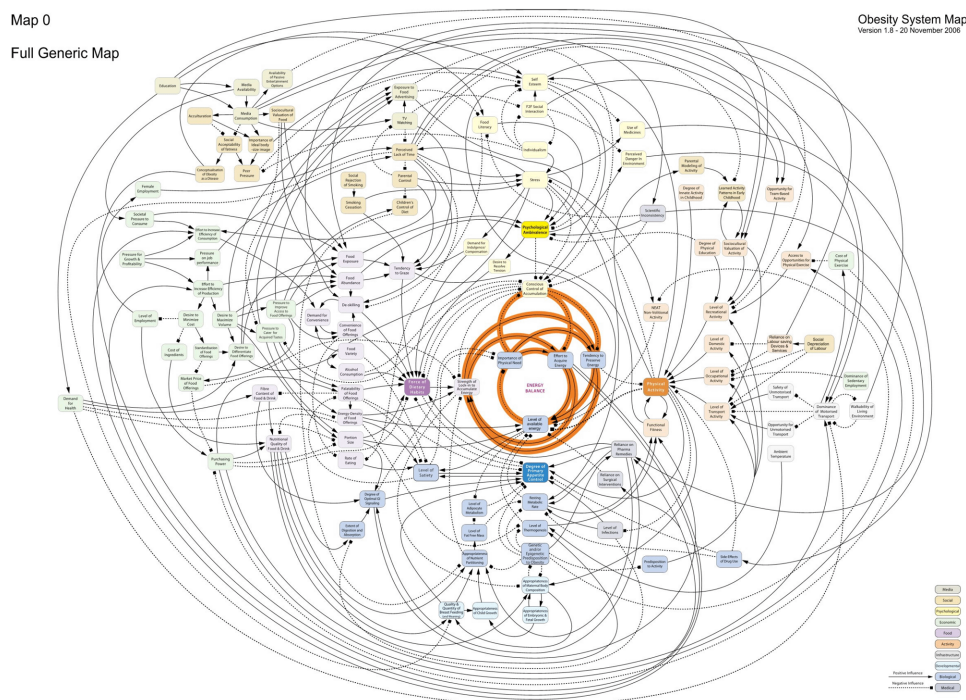


Figure 7.1: The Foresight Obesity Map is a well-known network representation of interdependencies between weight-related factors [235]. The use of this representation as a herald of ‘complexity’ rather than a decision-support tool [102, 213] exemplifies the difficulty of using network-based representations for participants who are not modelers [86].

Once an appropriate layer has been identified or developed, and after participants have been supported in working with this layer, the last step connects the layer with the model. This connection may not be trivial, as discussed elsewhere [239, 132, 93]. At first, it may also appear that this step could be done *before* or *irrespectively* of how participants navigate the layer. However, a layer may start as a very comprehensive map (e.g. the Foresight Obesity Map). Participants will find the subset of that map that is necessary to capture relationships between inputs involved in a what-if scenario (Figure 7.2). Connecting only what is necessary may save time, compared to connecting everything and then deciding on what actually needs to be used.

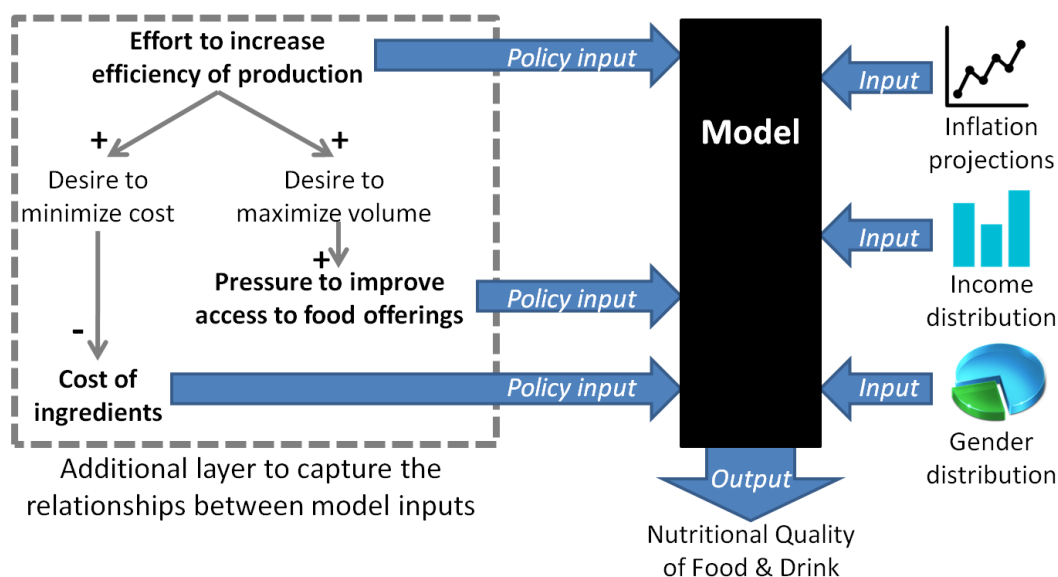


Figure 7.2: A policy may be targeting three inputs of a model (efficiency of production, access to food offerings, cost of ingredients). Using the relevant part of the Foresight Obesity Map, we can include the relationships between these inputs in a layer separate from the model.

Our aim is to improve the second step, such that existing maps (e.g., Foresight Obesity Map) can be used as guidance tools for the design of policies rather than as symbols of complexity. Our three contributions in this chapter are to:

- Identify the key functionalities that software need to support in order to navigate networks with a policy focus.
- Contrast these features with existing network and modeling software.
- Propose a new open-source software, and demonstrate it on existing maps.

The next section presents five required functionalities, grounded in experimental studies and key concepts of systems thinking in public policy. Then, we contrast these functionalities with those supported in existing software for visualization, argumentation, or modeling. Having identified unique needs for a new software, we introduce our open-source solution `ActionableSystems` in section 3. A brief demonstration is given in section 4, with additional

examples at <https://youtu.be/OdKJW8tNDcM>. Finally, we briefly discuss research directions in public health informatics, and we provide concluding remarks.

7.2 Functionalities to navigate networks with a policy focus

7.2.1 Functionalities required

The idea of an ‘input’ to a model is often rooted in a simple cause/effect reasoning: a change in one of the input factors triggers a changes in the model, which are reflected in a *different* set of factors labeled as outputs. In contrast, systems science and systems thinking emphasize the importance of loops:

“What really differentiates this kind of thinking from ordinary linear cause/effect reasoning is that none of these concepts can be regarded as more primary than the other. A change can be initiated everywhere in an event circle and after a certain time be read off as either cause or effect elsewhere in a system.” [214]

In this perspective, an ‘input’ is simply a part of the system selected for a policy intervention. A change in the input may trigger self-regulating mechanisms in the system, which eventually affect the input itself. Such mechanisms are well illustrated for complex conditions such as obesity (Figure 7.3) through the Foresight Obesity Map, or the landmark *Thinking in Circles About Obesity* [103]. When considering whether a functionality is required, it should not only be an important feature, but one that participants need assistance with. The volume *Structure of Decision: the cognitive maps of political elites* examined how participants thought of systems, and whether they were aware of existing loops. Throughout the book, findings are consistent: participants did not show that they were thinking of loops

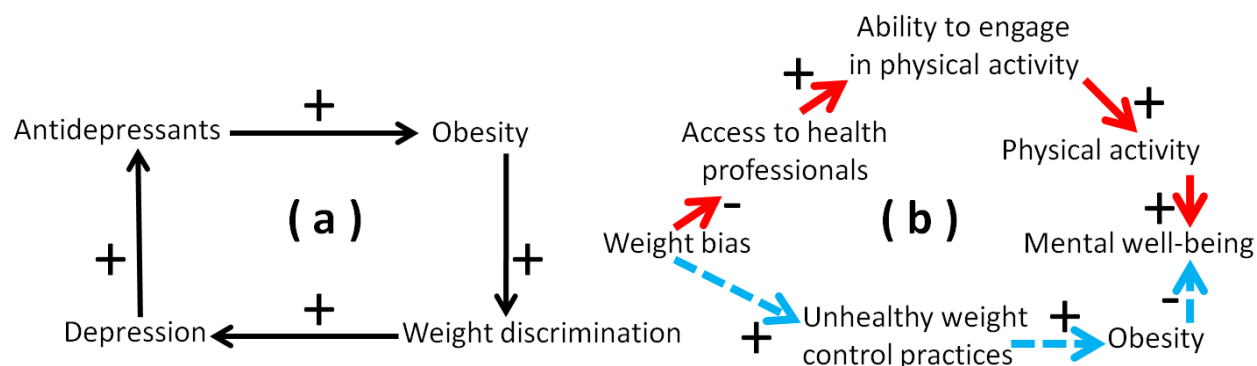


Figure 7.3: Example of a loop in obesity (a), where some classes of antidepressants cause weight gain [88]. Example of disjoint paths from weight bias to mental well-being (b), going through physical activity (solid red arrows) or eating behavior (dashed blue arrows).

when discussing complex systems. Ross saw it as peculiar that “those who set policy think only acyclically, especially since the cyclical nature of causal chains in the real world has been amply demonstrated” [199]. Examinations showed that the odd lack of loops or feedbacks in the networks was not due to lacking expertise, voluntarily simplifying the structure, of being focused on the near-term. Rather, the suggestion was to look for a cognitive explanation [12]: individuals unconsciously reduce complexity [237]. Therefore, loops have been shown to be crucial in understanding a complex system, and users need support to navigate them.

Functionality #1: Participants need to easily find the loops such that they can think of policy ‘inputs’ within the context of a broader system.

Similarly to loops, other patterns in the map are relevant to policies and difficult to manage. A policy intervention may target some factor(s) and measure its impact on other(s). In between, there can be disjoint paths carrying the intervention (Figure 7.3). For example, the intervention may fail significantly in one path and mask the relative success of another path. Understanding how the intervention permeates through the system is thus important

for its evaluation. Disjoint paths face the same issue as loops: individuals reduce complexity and ignore multiple paths [237].

Functionality #2: Participants need to easily find disjoint paths, in order to monitor how interventions unfold between a starting point and the outcomes.

The design and evaluation of an intervention may have taken into account what gets directly affected by the policy, the outcomes, and everything in between (through disjoint paths). However, a long-established hallmark of systems thinking is to understand “the rippling effects created by choices” [45], even if the effects do not contribute to the outcome. For instance, the ecological model of health promotion posits that (emphasis added) [64]:

“individual, familial, communal, national, international, and global health is highly intertwined and interdependent. Negative perturbations in any of the functional units may have untold negative *rippling effects*”

Note that the goal should not be to see *all* rippling effects. In some maps, an intervention may impact a massive number of factors, and it would be overwhelming rather than instructive to see them all. Rather, participants should be able to access a *filtered* set of rippling effects, depending on their needs. There are many possible needs: participants may want to ensure that a given set of factors are not affected by the policy (preserving a status-quo), sort the factors that are affected into categories (for cross-sectorial coordination), or perform an exploratory search with a limited depth (to see what would be affected through paths of up to x factors).

Functionality #3: Participants need to find and filter the rippling effects of interventions.

As part of our previous usability study, we also asked policymakers to provide their overall thoughts on a software supporting the use of maps for policymaking. Our thematic analysis revealed that participants did not see the role of software as limited to exploring or planning. They also suggested that it could serve as a tool for evidence synthesis, which is important for justifying policies. In the words of a participant: “if I could just go to one place, it would have all the information, that’s sort of my dream” [86]. In a comprehensive map, information cannot consist solely of the names of factors and their interrelationships. Given the multi-sectoral nature of public policies, the names of factors may have different (or no) meanings for different stakeholders. This is exemplified by our map for obesity and well-being [63], which included physiological (e.g., adipocytes, inflammation), legal (e.g., restrictive covenants), and behavioral factors (e.g., disordered eating, eating disorder). A map can thus serve as a common tool in which the meaning of each term is clarified, in addition to a synthesis of research regarding this term.

Functionality #4: Participants need to access and update the definitions and evidence supporting each factor.

Finally, when developing a new software for a given audience, we should be mindful of the existing workflow. That is, the software should integrate with existing tools rather than assuming that participants will let go of all of their past work and migrate *en masse*. Research on Information and Communication Technology (ICT) to support policies has shown that many categories of tools exist, and within each, many tools are available [125]. The three categories with which the proposed software would directly integrate are visualization (which serve for information provision), argumentation (which support structured deliberation), and simulation tools (which address ‘what-if’ questions through computations). There are also different forms of integration. For example, scientific workflow systems can connect

applications in pipelines where they automatically exchange information. The emphasis here is not on the automation but on supporting human decision-making processes, thus a new software should at least be able to take in data created via other tools.

Functionality #5: Participants need to manually exchange information between the new software and existing visualization, argumentation, and simulation tools.

7.2.2 Functionalities supported by existing software

There exists plethora of software for visualization, argumentation, or modeling. As a comprehensive analysis of these software and their reviews would be the subject of a dedicated review, we instead focus on a subset of these software drawing from a recent comparative analysis [125]. In Table 7.1, we summarized whether these software support the five functionalities defined in the previous section.

Out of ten software, we found that loops were only supported by the software **Vensim**, and similarly only the software **Gephi** provided (limited) support to finding paths. In this case, it allowed to find the *shorter* paths, instead of *all* paths that lead to a selected outcome. The other functionalities were supported by more software. Rippling effects were supported in two software, with **Vensim** allowing to see rippling effects on all the system (i.e., without filtering) and **Commetrix** offering an advanced level of customization including the depth. Note that to qualify as rippling effects, we looked for a depth greater than 1: simply clicking on a node and see what it *directly* affects was not counted as supporting *rippling* effects.

Many software offered the possibility of storing definition and evidence, but varied tremendously in how convenient they make it for users to access or update the data.

Table 7.1: Current ICT software and the five key functionalities

Software	Loops	Paths	Rippling effects	Definitions and evidence	Formats
Gephi	No (at-tempted a plugin but does not work).	Yes, shortest paths only.	No	Embedded in the data (spreadsheet)	Imports: 11 formats including GraphML, Pajek, UCINET, and CSV (edge list or matrix). Exports: 6 formats including CSV and GraphML.
Visone	No	No	No	Embedded in the data (spreadsheet)	Imports: GraphML, CSV/TXT (adjacency matrix), UCINET, Pajek, edge list, Siena. Exports: GraphML.
Commetrix	No	No	Yes, with advanced filtering	Yes	Imports: Excel, XML, database (mysql). Exports: CSV (node and link files).
Gpfinder	No	No	No	No	No imports. No exports.
Health Infoscapes	No	No	No	Yes, available when clicking on a factor	Developed for one dataset only. No exports.
Google Public Data Explorer	No	No	No	No	Imports: public data from Google or one's dataset. No exports.
MentalModeler	No	No	No	Yes, as side notes	Imports: own format (MMP). Exports: MMP, CSV.
iThink	No	No	No	Users can add a text area to the model.	Imports: own formats (STMX, ITMX), system dynamics (XMILE).
Vensim	Yes	No	Yes, without filtering.	No	Imports: own formats (MDL, VMF, VPM).
CLASP's Policy Analysis Modeling Systems (PAMS)	No	No	No	Yes, in Excel cells.	Designed for Excel workbooks.

In **Gephi** and **Visone**, data about factors and their relationships is imported in the form of a spreadsheet, so users ‘could’ go through the internal data storage and manually create/edit columns for meta-data on definitions and supporting evidence. In **Health Infoscape** and **MentalModeler**, the access is much more immediate: clicking on a factor suffices to see the information as either a pop-up window (in **Health Infoscape**) or a side panel with notes (in **MentalModeler**). Finally, all software but two allowed to import and export files. However, the intention is not to merely to have a large collection of formats, but to promote interoperability between software such that practitioners can add capabilities to an existing workflow. Findings on formats are thus nuanced: three software operate with their own formats (**iThink**, **Vensim**, **MentalModeler**), two are meant for graphs yet they each use over ten different formats (**Gephi** and **Visone**), and even the same file CSV extension is used to store very different graph data (one list for **Gephi** but two lists in **Commetrix**, and a matrix in **Visone**). This paints some of the difficulties that practitioners have in navigating this software ecosystem, with its many formats, and even different meanings for what could appear to be the same format.

In summary, no software supports 4 or all 5 of the key functionalities for systems thinking in policymaking. Only **Commetrix** fully supports 3 functionalities, while **Vensim** and **Gephi** partially support 3. As the need for software supporting all functionalities is currently unmet, the next section details the design of our solution.

We note that the software surveyed here also have their own strengths, which are not necessarily captured by the five functionalities on which we focus. While a software may not easily connect to others through file input/output, some provide additional (often less intuitive) means to facilitate a workflow. **Visone** possesses a console to use the language **R**, which offers extended capabilities to connect with other software. **Gephi** is designed as an extensible software (i.e., uses a plug-in architecture), with over 21 plugins to import and export. A visualization software also tends to provide extensive support for different ways to

render information. In the case of a map, the position of the map elements on the screen is determined by a *layout*, and software often support a variety of layouts (with over 15 layouts in `Gephi` and over 10 in `Visone`). Even though the modeling software studied here all deal with some extension of the concept of graphs, none supports layouts: the user is entirely responsible for deciding on the position of all elements. Their strength is instead to support practitioners in quantitatively evaluating what-if scenarios.

7.3 Proposed software: `ActionableSystems`

7.3.1 Overview

The design of our proposed software took place over a three year period. Starting in 2015, our joint work with the Provincial Health Services Authority (PHSA) of British Columbia produced a very comprehensive policy map [63], which was difficult to analyze with existing software (Table 7.1). Through extensive discussions with members of the PHSA and other researchers, we developed a software tailored to the PHSA map. In 2016, we pilot-tested the software with several policymakers [86]. Our usability sessions resulted in over 30 recommendations on how to improve the user experience. In addition, the analysis of semi-structured interviews revealed that policymakers saw more potential uses for the software than it was initially designed for. Based on these results, we clarified the key functionalities that policymakers need to navigate policy maps (section 7.2), and we created the new `ActionableSystems` software focused on these functionalities. Our software is written in the Java programming language and is open-source. It is hosted on the third-party repository Open Science Framework at <https://osf.io/7ztwu/>, where programmers can re-purpose the code, while users can download and run the software.

7.3.2 Design Principles

Three principles underlie the design of `ActionableSystems`. First, the software should be *simple*. The expertise of our intended users resides in policymaking, or in specific domains impacted by a policy. We do not assume that users are experts with specific computer techniques, such as network visualization. Consequently, the software needs (i) to use a language that is free of technical jargon, (ii) contextualize what its functionalities mean in a policymaking context, and (iii) include training. Figure 7.4 illustrates these principles. The left panel uses simple terms (e.g., ‘See’ instead of ‘Interactive Visualization’) and operations emphasize what they are for rather than how they work (e.g., the button “Can’t see well? Click here to reorganize!” would be labeled as applying a network layout in many software). Examples are provided in three forms: long walkthrough tutorials (accessed via the ‘Tutorials’ button), legends (Figure 7.4, bottom left), and tool-tips (e.g., hovering over a policy domain provides examples of what it includes).

Our second principle is to emphasize *consistency*. This is a key principle in design, and it contributes to making a software intuitive to use. It requires that similar elements are seen the same way, and that similar controls function the same way. For example, simple elements such as buttons in the same category should have the same sizes and fonts: Figure 7.4 shows visual consistency for the main buttons in the left panel, sub-categories in the top panel, and tool-specific buttons in the bottom panel. One added difficulty in maintaining visual consistency in our software is that the same data can be viewed in different ways, each emphasizing a different aspect: the policy network can be seen at a high level (Figure 7.4), or through specific cycles and disjoint paths. All these views maintain visual consistency on some aspects (e.g., the thickness of a causal relationship shows its strength) but differ on others (e.g., relationships are arranged in a circle when showing a cycle). Consistency in

control is relatively easier to maintain, as interactions within a same category such as ‘See’ all trigger the same effect (e.g., a double-click always packs or unpacks a policy domain into its individual factors).

Finally, our third principle is *relevance*. Our software is designed for policymaking, hence its functions must be relevant in this context and the relevance must be clearly conveyed either through short explanations or longer tutorials. A consequence is to avoid the temptation to implement functions just because we can, as may happen during software development cycles that gradually lose track of their intended audience. For instance, while a myriad of network measures exist, the ‘Measure’ button provides access to few measures but emphasizes their meaning in a policy context. Similarly, two networks can be compared on many possible features, but the ‘Compare’ button contrasts two policy networks based on features such as the presence and types of feedback loops.

7.3.3 Implementation of the five functionalities

A video demonstration of the five functionalities in `ActionableSystems` is provided at <https://youtu.be/OdKJW8tNDcM>. The first three key functionalities are accessed via the top panel (Figure 7.4), after clicking on the ‘See’ button. The analysis of cycles (functionality 1) gives access to a list of all cycles, and each one is displayed by arranging the content as a circle (Figure 7.5-a). When finding paths (functionality 2), we use a pop-up window whose design implements recommendations from our previous usability study [86]. Users select the end and start node, either through a drop-down menu or by typing a few letters and using auto-complete. End-nodes that are greyed out cannot be reached from the selected starting node, whereas black end-nodes can be reached from at least one path. When one or more paths exist, they are graphically organized so the user can see the different paths (Figure 7.5-c).

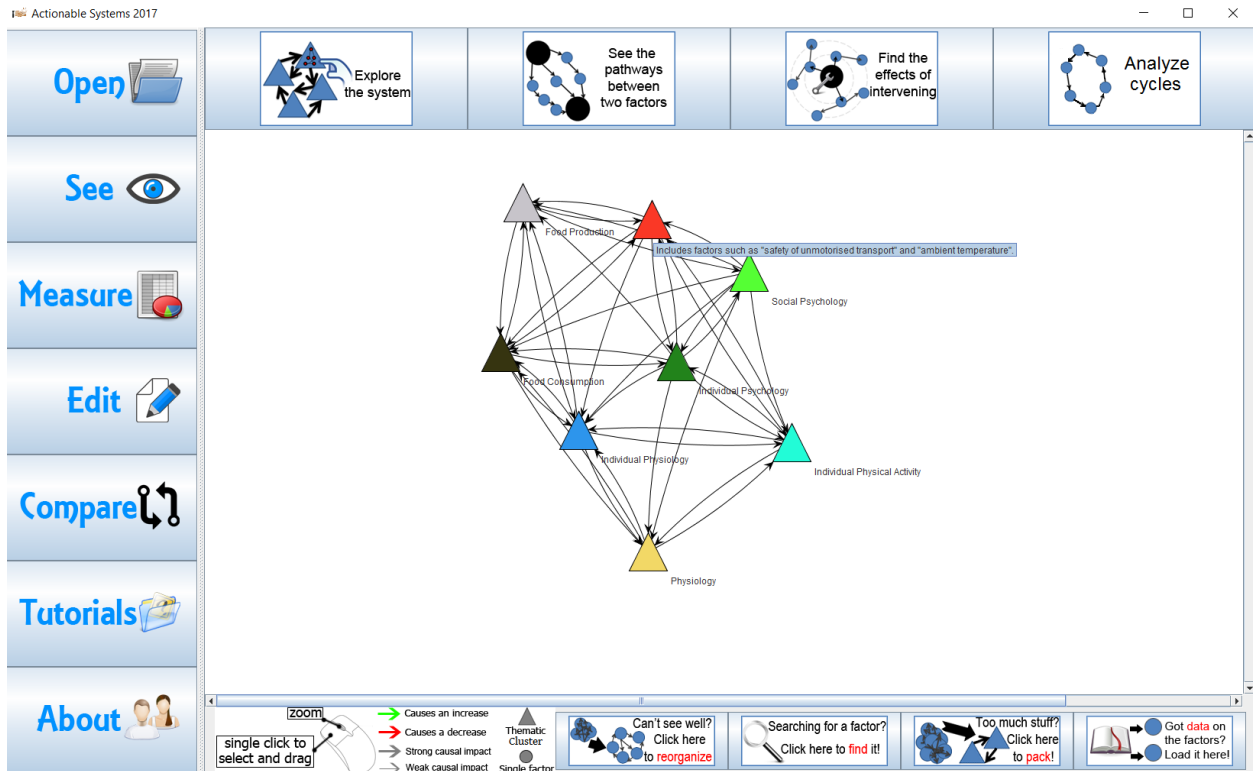


Figure 7.4: Rendering of the System Dynamics model from Verigin *et al.* [238] within ActionableSystems as a hierarchical network. Each policy domain is shown as a thematic cluster (triangle). Hovering over a cluster shows examples of its content, while clicking on it will unpack the factors that it contains. Tools to find rippling effects, paths, and cycles are at the top.

To find rippling effects (functionality 3), users choose the factor on which to intervene, and how far they want to be screening for rippling effects. The result is organized in concentric circles to emphasize ‘rippling’ effects (Figure 7.5-b). Participants can import definitions and evidence *en masse* via the ‘data’ button (Figure 7.4, bottom right). The evidence can be viewed, edited or created (functionality 4) by clicking on a factor and opening an editor within a pop-up window.

While previous software opened files created by ‘similar’ software (Table 7.1), our software works across domains. Specifically, it opens maps created by visualization, argumentation, and simulation tools (functionality 5). To connect with network visualization software (e.g.,

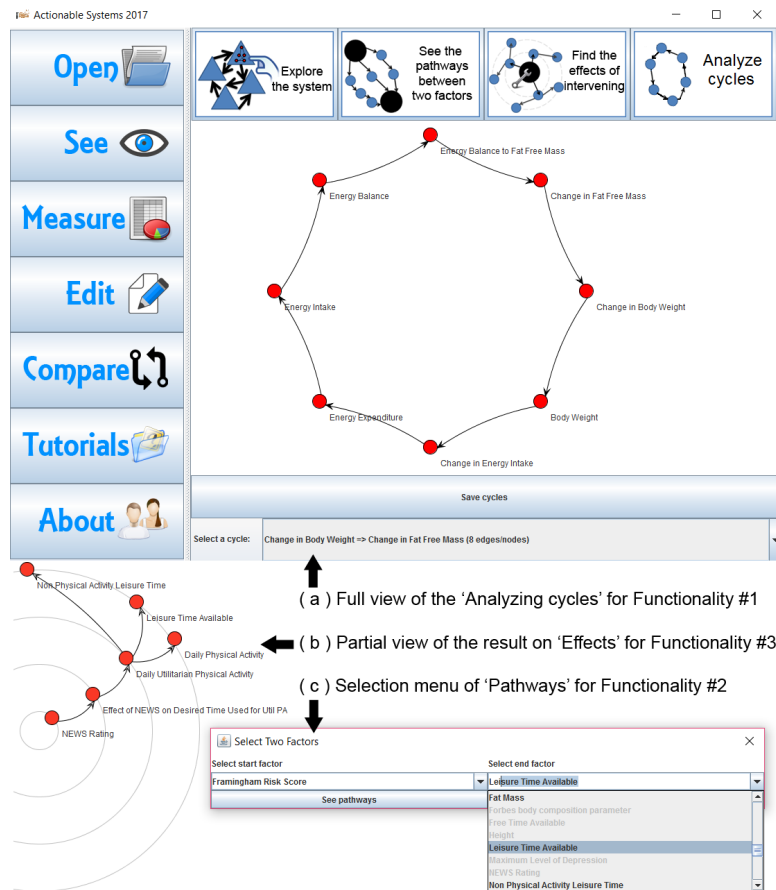


Figure 7.5: Key functionalities for the System Dynamics model from Verigin *et al.* [238]: (a) a cycle with 8 nodes, (b) rippling effects of intervening on the NEWS rating up to depth 3, and (c) searching for paths starting at the Framingham Risk Score.

Gephi and Visone), we use the GraphML format which is defined as a common (XML-based) format for exchanging graph data in a visualization context [30]. We connect with argumentation software (Cmap and Coggle) by reading files in their own formats, and similarly we access data from simulation software such as MentalModeler by reading its own format. As the idea of integrating in a workflow means that we can get data both in *and out*, all results generated by users can be saved. For instance, the whole list of cycles can be exported with the “Save cycles” button (Figure 7.5 bottom), and the same applies to pathways, or results obtained via the ‘Measure’ or ‘Edit’ button.

7.4 Demonstrations

In 2013, the Provincial Health Services Authority of British Columbia authored a discussion paper on the inter-relationships among obesity, overweight, weight bias and mental well-being [183]. The paper narrated the evidence, but did not visually represent it. As a follow-up, we created a map of obesity and well-being [63]. Using `ActionableSystems`, we can now ask key policy questions from the map.

First, we can investigate in which ways weight bias affects mental well-being, in the context of obesity. This is an essential question, whose answers runs through the pages of the previous discussion paper but were not previously available in a simple, synthesized form. Using the tool for disjoint paths in `ActionableSystems`, we immediately know that there are 6 disjoint paths (Figure 7.6-a). They vary in length from a direct impact of weight bias on mental well-being (length 1) to a path going towards depression and its effect on physical health (length 5). The PHSA map is annotated, as its causal connections have a strength and a type (either a causal increase or decrease). This information allows us to more precisely understand the type of paths running from weight stigma to mental well-being. The composition rule for causal effects (often used in System Dynamics) can intuitively be understood as a multiplication: if A increases ($\times 1$) B, and B decreases C ($\times -1$), then A decreases C ($1 \times -1 = -1$). More formally, in a causal path, an odd number of causal decreases leads to the path representing an overall decrease. In Figure 7.6-a, we observe that all of the paths have an odd number of red edges (i.e. causal decrease), that is, weight bias decreases mental well-being in six different ways.

Second, identifying potential policy levers requires a deep understanding of what already drives the dynamics of the system. Loops are important drivers, either to balance (odd number of causal decreases) or to amplify dynamics. Finegood has long suggested that

“new methods are likely required to assist stakeholders in [...] creating new feedback loops as a means to shifting the dominance away from [the loops that] currently give rise to obesity” [70]. **ActionableSystems** provides support in this regard, by allowing policymakers to list all loops and their types. Figure 7.6-b shows a balancing loop, where heart diseases reduce exercises, which in turn increases the likeliness for heart diseases. Policymakers need to counter-act this undesirable loop, but cannot simply ‘remove’ it because most of it involves physiological causes (which are outside their control). They can thus add to the system by supporting exercise-based cardiac rehabilitation, which has been proven to reduce cardiovascular mortality in an updated 2016 Cochrane systematic review [7]. This will ‘take away’ from the balancing loop and promote exercise. Conversely, Figure 7.6-c shows a reinforcing loop in which individuals eat less healthily. The ‘Tragedy of the Commons’ in System Dynamics suggests that, if a harmful loop needs a resource and this resource cannot be directly modified, then a less harmful loop could be created in order to tap into that same resource and deplete it. Rather than telling individuals to just eat less, an approach can be to promote the consumption of healthy foods (e.g., high in fibers) which should deplete ‘appetite’ as the resource and consequently have individuals eat less unhealthy foods.

Finally, maps are *representations* of systems. Maps may depart from the real system depending on how they were created, and understanding these discrepancies is important when policymakers base their analysis off maps. **ActionableSystems** provides a summary of a map, and can also contrast it with another map. In Figure 7.7, we see that the Foresight Map has a roughly equal share of reinforcing and balancing loops. In contrast, the PHSA Obesity and Well-Being map almost only has reinforcing loops despite being a view about a very similar problem. Policymakers will thus be presented with very few balancing loops when it comes to obesity, which under-estimates readily available solutions that policies may reinforce.

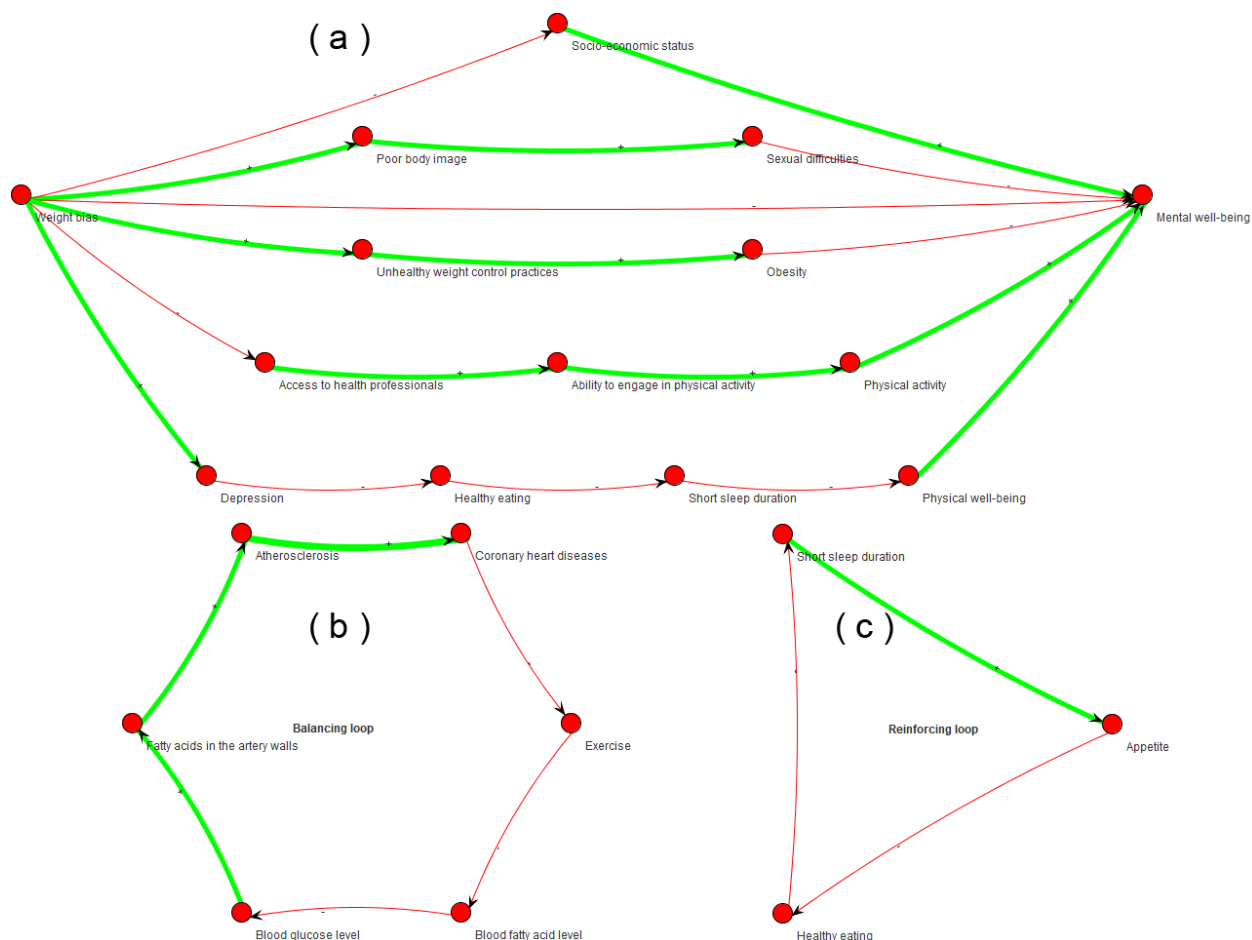


Figure 7.6: Visualizations of the PHSA obesity and well-being map [63] showing pathways from weight stigma to mental well-being (a), and two examples of balancing loops (b-c).

Your main file is: **Foresight Map (edges).csv**

You're currently comparing with: **Drasic et al (edges).csv**

[Want to compare with another file? Click here!](#)

Loops		
Total number:	245	30
Of which reinforcing:	138	28
Of which balancing:	107	2
Average length:	7.595918367346939	5.666666666666667

Edges/links		
Total number:	294	177
Average per node:	2.7222222222222223	1.8061224489795917
Isolated nodes (no edge):	0	0

Figure 7.7: Comparison of the Foresight Map (left) with the PHSA obesity and well-being map [63], using metrics such as the number and types of loops.

7.5 Discussion and conclusion

We designed a new software solution to support policymakers in navigating complex systems, and demonstrated its possibilities on systems used in obesity research. Our software integrates with visualization, argumentation, and simulation tools. However, there are several other types of important tools in policymaking [125]. Integrating with opinion mining tools is an important next step for public health informatics. Indeed, policymakers need to know what constituents support in a policy [87]. If this was readily accessible in a network form, they would be able to more easily find policies that can positively impact the dynamics of the system *and* are endorsed by constituents. This integration would not significantly alter the design of our software (e.g., the level of endorsement can be visually shown through edge patterns). However, other types of integration call for new designs. In particular, eParticipation tools require an online, distributed design. In contrast, our software and all the other solutions reviewed are designed for a single user. The next frontier in public health informatics is to develop tools that allow multiple users to navigate complex systems, and possibly in an asynchronous manner (i.e. when not all users are interacting with the software at the same time). This will require new designs and usability studies, but the effort also comes with the promise of a more inclusive approach to policy-making and a more comprehensive exploration of complex systems.

CHAPTER 8

CONCLUSIONS AND PERSPECTIVES

8.1 Introduction

The new millennium initiated the creation of a new perspective on public policies as a complex system. The process of formulating a policy is non-linear and time-consuming, due to a few important aspects [121]. Firstly, the development of new regulations involves many stages, linked together into a “policy cycle”, as shown at Figure 8.1. Each stage itself can be a composite process, requiring cooperation of various authorities and institutions. Secondly, the important actors change across the various phases of policy-making, which might result in fragmentation, since actors are often not aware of the decisions made by the other actors. Politicians, members of parliament, executive branches, courts, and interest groups may be involved in these formulations. Finally, often contradictory proposals are made, and the impact of a proposal is difficult to determine as data can be missing, models may inadequately capture the complexity of a policy, and interpreting the results of policy models may be challenging.

A new notion for policymaking requires new methodologies, which would enhance efficiency, increase transparency, and improve the acceptance level. Public health informatics, and specifically Information and Communication Technologies (ICTs) used for policymaking, have produced a variety of tools, including [125]:

- *Visualization tools*, which provide better understanding of data and give users a broader context, especially when representing data in graphical form.

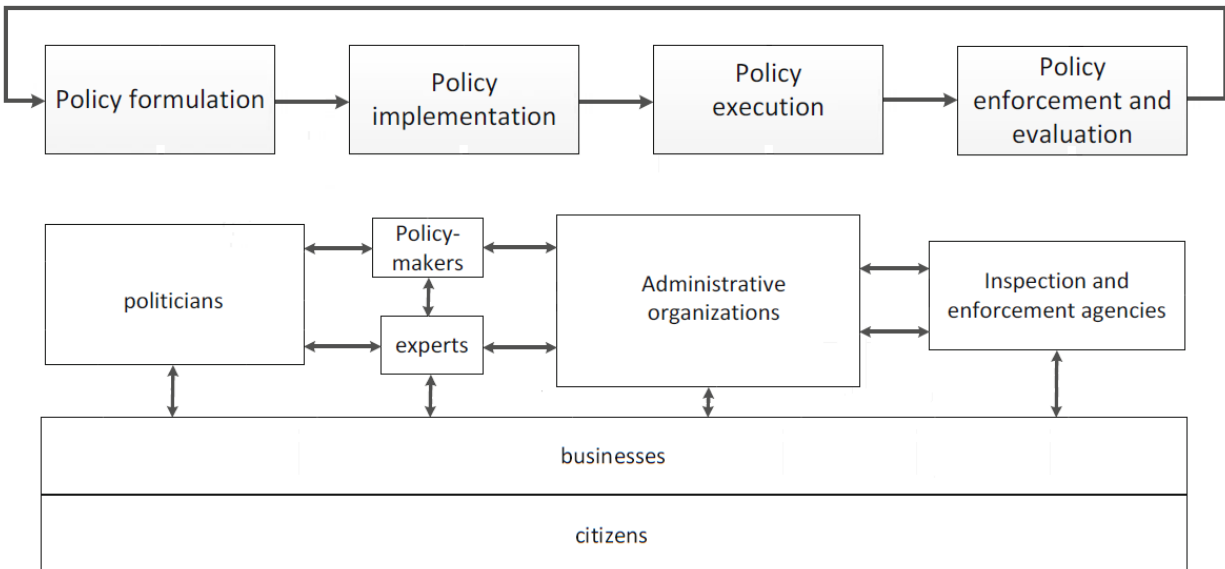


Figure 8.1: Overview of policy cycle and stakeholders. Adapted from [121]

- *Argumentation tools*, which organizes complex argumentations and debates through means including network visualizations.
- *Simulation tools*, in which real-world phenomena are abstracted so that can be better analyzed, and decisions can be formulated.

8.2 Achievements

8.2.1 Contributions to computer science and simulation research

The design and evaluation of public policies can be seen as complex systems because of non-linear interactions between factors and many actors involved, as presented at Figure 8.1. There are many popular Modeling & Simulation methods used for simulating complex problems (e.g., agent based modeling, cellular automata, genetic algorithms)[231]. This thesis

sought to further the use of M&S for complex systems, ranging from the design of models for policy settings to the interpretation of their output through novel environments. Two particular achievements are notable in this regard:

1. We have developed a novel model including food exposure for children across levels of deprivation, combining several English databases to achieve a detailed large-scale network simulation. This is a major improvement over previous studies, which used networks only at the scale of one city or region [178, 177] whereas we have brought it up to the national scale.
2. We provided and evaluated new interactive ways to visualize data, for example as it is produced by discrete simulations. The innovation of the method proposed in this thesis lies in organizing the data with replications and multiple time steps into a compact representation. As a complement to this new environment, we developed several tools connected with the main visualization, such as filtering or details on demand. Our empirical evaluation of this new environment showed that participants could perform important modeling tasks in a confident and accurate manner, while taking no longer than they would have spent on the previously available but much less accurate tools.

8.2.2 Contributions to public health and obesity research

This thesis proposed brand-new systems, which addressed the multidisciplinary character of policymaking by drawing from an array of fields including social science, economics, statistics, and computer science. By analyzing and predicting complex environments, our work has a potential to provide researchers and policymakers with tools to design and evaluate interventions taking place in composite social systems. The important role of computational

tools is considered to be a promising direction for the future of public policies. As Janssen and Wimmer stated:

“... the explosive growth in data, computational power, and social media creates new opportunities for innovating the processes and solutions of ICT-based policy-making and research. To take advantage of these developments in the digital world, new approaches, concepts, instruments, and methods are needed, which are able to deal with societal and computational complexity.” [121]

The main contributions of this thesis to public health and obesity were to enable systems thinking using computational tools. These tools can serve both as specialist platforms for the experts and as teaching systems to enhance citizen’s engagement:

1. The aforementioned model of food exposure for children across levels of deprivation was used to understand the factors contributing to the change of food premises locations. While some association was found between the presence of schools and fast-food outlets, we did not find a causal relationship. That is, fast-food outlets did not purposely target schools to a visible extent, suggesting that more significant factors were involved in the choice of a location. As there is a wide range of zoning regulations, and each regulation may also take a wide range of parameter values, our model can serve to quantify the expected impacts of these regulations together with their interactions. The availability of such a model is very timely, given that many such policies have been already implemented in a few English neighborhoods [46, 48, 47, 246], while others are in the process of implementing them.
2. To create a complete data science pipeline for policymaking, we went from data analysis (network mining) to predictive analytics (discrete simulations) and data visualization. Each step has a potential to significantly improve policymakers’ understanding of social

phenomenon such as obesity and help them to design more effective interventions. The network analysis centered on the relationship between fast-food premises and schools in England supports policymakers in understanding how combined planning measures around schools affect the England food landscape across different levels of deprivation. Finally, a new tool, which transforms models into more ‘actionable’ tools through visualization, guides policymakers in navigating complex systems.

8.3 Limitations

A great challenge for researchers and scientists is the availability of data. Many datasets that are free of charge contain incomplete data, and the ones with complete records may be expensive. However, data is essential: in our simulation model, it serves to initialize the variables at the beginning of the simulation, as well as to calibrate the model. Although our study is the most comprehensive in terms of using data (Points of Interest, Lower Layer Super Open Area Boundaries, Boundary-Line, Ordnance Survey Open Roads, Indices of Multiple Deprivation 2015), we did not include data on the profile of the outlets or their customer base. This limits our ability to precisely capture exposure, and thus utilization, of fast-food outlets. Other datasets exist such as Dun and Bradstreet, which can complement the datasets used in our study, but they require fees.

While availability of data is paramount, one should not ignore the many issues that arise in the process of data cleaning and wrangling. The individual-level information is very attractive for researchers in computational social science, but such datasets are very hard to obtain. A common solution is to use aggregated data, which does not give the same detailed perspective on individuals behavior but is easier to find (e.g. because there are less risks to participants when sharing highly aggregated data across studies). We used the well-

established Deprivation Index, which combines many factors [73]. However, each component of the index may have a separate effect on the relation between individuals and fast-food outlets. Disentangling these factors would thus allow later studies to capture specific subsets of the population.

Another limitation that is common to modeling work is that models are necessarily a simplification of reality: they never include all the factors that one may think of. Models capture only the most important aspects of a problem, and thus narrow the object of the study. In our case, we considered that the spatial distribution of fast-food outlets only impacts school-year students' exposure to calorie-dense foods, which greatly contributes to obesity among children. In practice, other factors such as distinction of a different types of school or patterns of urban growth in the area might also play a role. However, the selection of factors used in a model is often limited by available data, as explained before. If several factors are included in a model but there is no data to support them, then additional assumptions have to be formulated about these factors' values, which ultimately raises the uncertainty of simulation outcomes.

The utilization process of the tools proposed in this thesis also faces few challenges and critiques. In 2016 the Food Foundation prepared a report containing recommendations and guidelines to help prioritize public policies. Two of them, listed among those poorest implemented, emphasize the need for an effective change of food environment (Table 8.1). However, the priorities of policy-makers and public policies' developers may differ from the Food Foundation, or vary across constituencies. A key argument that transcends local preferences is the matter of *cost*. Our simulations can output the efficacy but do not provide calculations in terms of cost, which is half of the equation given that *cost-effectiveness* analysis is necessary to conclude as to which interventions should actually be prioritized.

Our proposed CA visualization also faces several limitations. First, our tool can deal with many time steps and many replications, but not a large number of cells. Cluttering occurs

Table 8.1: Ranking of the best and worst implemented public policies in England [225]

<i>Highest scores (good implementation of policies)</i>	<i>Lowest scores (poor implementation of policies)</i>
Monitoring of overweight, obesity	Platforms between civil society and government
Monitoring of NCD risk factors	Subsidies in favor of healthier food
Labelling with regard to nutrient declarations	Investment management and non-food policy development that takes account of public health nutrition
Access to information and key government documents relating to the food environment	Planning policies that favor healthier foods
Dietary guidelines established	Systems based approach to improving food environments
School food standards	Advertising in child settings
Population intake targets established	Coordination mechanisms across different government departments
Labelling with regard to FOP	Workplace food provision
Monitoring of nutrition status	Advertising through non-broadcast media
Food composition standards established	Comprehensive implementation plan to improve food environments

when visualizing a large CA. Our participants reported a significant increase in difficulty when working with a CA of 50 by 50 cells, instead of 25 by 25. Additional research is needed to identify, implement and evaluate the use of complementary scaling methods for spatial data aggregation. Second, problems can occur when working with a great number of segments per cell. In our first version of the environment we provided three options for division of a single cells: 4, 8, or 16 segments per cell. After consulting with experts in the field, the division on eight segments was chosen as providing the most readable version of the visualization. However, the right number of divisions may vary depending on the phenomenon model, or the visualization task, and such interactions were not explored.

Finally, our new approach to visualization focused on two tasks: detecting hypervariability and trends from time series. Many other tasks are routinely performed on cellular

automata, particularly when it comes to checking the output for the debugging process. We are thus only able to conclude about the usefulness of our proposed approach for specific tasks. In future versions of the environment, development of additional tools might be required to support these additional characteristics for CA tasks such as finding a specific feature as in geographic systems [240] or exploring the distribution of a distinct state over time and over space as in fire spreading studies [5].

8.4 Future work

This thesis emphasized the importance and outlined the benefits resulting from applying system thinking approach to policymaking. Special interest focused on developing new ICTs to support public health experts in their uneasy task of constructing effective interventions to tackle obesity. As presented in this work, computer models and simulations creates new opportunities to even better influence societies. Therefore, many other directions of collaboration between computer science experts and public health authorities should be further explored. We highlight three potential avenues for future work. In section 8.4.1, we proposed possible improvements to our visualization environment, which include both adapting to specific users' needs and expanding our method to other simulation techniques. In section 8.4.2, we outline how GPS data incorporated to our model can help to collect additional information about people's walking paths and estimate their exposure to unhealthy food. Finally, the potential application of our network model to test policies both separately and altogether is explored in section 8.4.3.

8.4.1 A new visualization approach to multiple run and multiple replicated data

We proposed a new visualization technique to complex, hypervariable data based on glyphs applied to each cell of a Cellular Automata grid. To improve our implementation and expand its applications beyond CA, we have identified three specific objectives for future work. First, the use of *weights* can help to differentiate important states within the visualization. Our aggregation method showed the state that occurred the most. However, it is not necessarily the most important state to see. Customizable weighting would allow to under- and over-weight certain states for display. For example, ‘death’ in a cell may be recorded just once, but it is of the utmost importance. Figure 5.3 in chapter 5 showed a CA for modeling HIV spreading with four possible states [188]. In this context ‘dead’ state is impossible to track and distinguished from other, more common cells. Consequently, users may want to assign different weights to the states based on context. Future work should examine the extent to which it is possible to automatically capture a context to set the weights, rather than requiring users to set them all manually.

Second, there is a growing interest in using visualizations to analyze the output of simulation models. In this work, we proposed a new aggregation technique to visualize the data generated by a two-dimensional cellular automaton with square cells. This technique has a lot of potential for other types of modelling approaches. For example, it can be applied to spatial simulations that did not necessarily use a grid-based system. To do so, the simulation’s geographical space can be discretized (see Figure 1.3). Similarly, it can be used for hybrid simulations of which one component can be translated to a CA [93]. Using multiple models can be particularly helpful for complex problems, since it helps to deal with a plethora of important characteristics (e.g., agent characteristics, feedback and accumulation

effects, spatial and network effects) which could be hard to capture using a single simulation model.

Third, we envision the creation of a large visualization databank, in which each dataset is visualized using different aggregation methods and number of segments. A large collection of visualizations, created using different aggregation techniques and number of divisions within a cell, is an essential step on a way to understand how visualization parameters (i.e. aggregation method or number of segments) affect task performance. Modelers' contribution is also necessary, to evaluate each display by assigning a numeric score to every visualization, based on how informative it is for a given task. The challenge would be to select tasks most relevant to modeling and suitably varied to capture the perceptual notions involved. In this thesis, we focused on two tasks: detecting hypervariability, and trends from time series, but other possibilities exist, such as identifying cells whose final state is the initial one or localizing a spread. However, assembling a comprehensive dataset with cooperation with modelers to judge a large number of visualization is a strenuous and hard work.

8.4.2 Tracking peoples' walking paths

Two important barriers to accurately modeling dynamics in the food environment include a lack of data to capture how individuals navigate the food environment, and a lack of methods to integrate such data in models. In our work, we focused on a detailed dataset containing the road network system across the entire England. This network-based model can be extended, and transformed into a full Agent-Based Model (ABM). For example, we could incorporate various datasets into our model to adjust the exposure function based on the agents' behavior, such as information about children's walking paths, or how long individuals stay in a fast-food outlet. A particularly promising possibility is to use data on

individual mobility patterns captured using Global Positioning Systems (GPS). However, it was previously emphasized that such datasets need to have both a sufficient time window and over a hundred individuals [91]. These requirements are essential, in order to calibrate and validate a system with reasonable confidence margins. While small number of participants (less than 100) within dataset might not be sufficient to infer behavioral patterns, extensive datasets, containing records collected over a single day are hard to process.

8.4.3 Modelling public policies

This thesis presented a large scale geospatial simulation using a network model to capture the food landscape across England. As this landscape is targeted by myriad different regulations, that can each take a range of values, our model can serve in future work as a tool to quantify the expected impacts of these regulations together with their interactions. For example, the most typical measure of creating a buffer around schools has seen a variety of distances in the UK: 65 meters (to prevent "ice cream trading"), 300 meters (by Glasgow City Council), or 400 meters [76]. Even the most common value of 400 meters tends to be justified by heavily simplified calculations: assuming that children can walk 10 minutes, that would make 800 meters as the crow flies, so having some physical barriers along the way would make it half as much, hence 400 meters [76]. Our simulations might serve to quantify the shift in fast-food exposure that would be expected for each value. Capturing these interactions will be an essential step to bring a system science approach into local regulations, as previous research on obesity and other chronic conditions has shown that these interactions can be particularly informative about the system as a whole [249, 88]. The last challenge would be to provide policymakers with sets of synergetic scenarios depending on the intended local priorities regarding changes in food exposure. However, part of the

reason for the variety in regulations witnessed so far is that each local authority has its own characteristics and priorities. Consequently, rather than providing a 'one-size-fits-all' recommendation, we believe that local councils seek to prioritize different aspects and thus further work should provide a set of synergistic scenarios that could achieve their own targets. The primary objective will be to identify the metrics that local authorities currently pay close attention to, and the range of values that are deemed desirable. This will allow us to generate different combinations and relate them straightforwardly to these metrics such that local authorities can find approaches that deliver on their own objectives. Possible metrics may include a reduction in fast-food outlets around schools, a balance of takeaways across levels of deprivation, or a reduction of fast-food outlets in areas with higher levels of childhood obesity.

REFERENCES

- [1] "2020Health". Careless eating costs lives, 2014.
- [2] A. Abbasi, L. Hossain, and L. Leydesdorff. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6(3):403–412, 2012.
- [3] A. Aggarwal, A. J. Cook, J. Jiao, R. A. Seguin, A. V. Moudon, P. M. Hurvitz, and A. Drewnowski. Access to supermarkets and fruit and vegetable consumption. *American Journal of Public Health*, 104(5):917–923, 2014.
- [4] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing time-oriented dataa systematic view. *Computers & Graphics*, 31(3):401–409, 2007.
- [5] A. Alexandridis, D. Vakalis, C. Siettos, and G. Bafas. A cellular automata model for forest fire spread prediction: The case of the wildfire that swept through spetses island in 1990. *Applied Mathematics and Computation*, 204(1):191–201, 2008.
- [6] J. Amcoff. The geographical redistribution of retail outlets in sweden 1998-2008. *The international review of retail, distribution and consumer research*, 26(3), 2016.
- [7] L. Anderson et al. Exercise-based cardiac rehabilitation for coronary heart disease. *Journal of the American College of Cardiology*, 67(1):1 – 12, 2016.
- [8] G. Andrienko and N. Andrienko. A general framework for using aggregation in visual exploration of movement data. *The Cartographic Journal*, 47(1):22–40, 2010.
- [9] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak, and D. Keim. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum*, 29(3):913–922, 2010.

- [10] S. Athanassopoulos, C. Kaklamanis, G. Kalfoutzos, and E. Papaioannou. Cellular automata: simulations using matlab. In *Proceedings of the Sixth International Conference on Digital Society (ICDS)*, pages 63–68, Valencia, Spain, 2012. Conference: 6th International Conference on Digital Society (ICDS).
- [11] S. Austin, S. Melly, B. Sanches, A. Patel, S. Buka, and S. Gortmaker. Clustering of fast-food restaurants around schools: A novel application of spatial statistics to the study of food environments. *American Journal of Public Health*, 95(9):1575–1581, 2005. [Online; accessed 22-September-2016].
- [12] R. Axelrod. *Structure of decision: the cognitive maps of political elites*, chapter Results, pages 221–250. Princeton University Press, 1974.
- [13] D. Bahr, R. Browning, H. Wyatt, and J. Hill. Exploiting social networks to mitigate the obesity epidemic. *Obesity*, 17(4):723–728, 2009.
- [14] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: an explanation of $1/f$ noise. *Physical Review Letters*, 59(4):381–384, 1987.
- [15] F. Ball, D. Sirl, and P. Trapman. Analysis of a stochastic sir epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53–73, 2010.
- [16] M. Barthélemy. Betweenness centrality in large complex networks. *The European Physical Journal B*, 38(2):163–168, Mar 2004.
- [17] M. Barthelemy. Spatial networks. *Physics Reports*, 499:1–101, 2011.
- [18] M. Barthelemy, P. Bordin, H. Berestycki, and M. Griboaudi. Self-organization versus top-down planning in the evolution of a city. *Scientific Reports*, 3:2153, 2013.
- [19] M. Batty. Accessibility: in search of a unified theory, 2009.

- [20] F. Beck, M. Burch, and D. Weiskopf. A matrix-based visual comparison of time series sports data. In *Proceedings of the Conference on Vision, Modeling and Visualization, VMV '16*, pages 53–60, Goslar Germany, Germany, 2016. Eurographics Association.
- [21] H. Bester. Quality uncertainty mitigates product differentiation. *The RAND Journal of Economics*, pages 828–844, 1998.
- [22] M. Bezbradica, M. Crane, and H. J. Ruskin. Applications of high performance algorithms to large scale cellular automata frameworks used in pharmaceutical modelling. *Journal of Cellular Automata*, 11(1):21–45, 2016.
- [23] C. Bhat, S. Handy, K. Kockelman, H. Mahmassani, Q. Chen, and L. Weston. Urban accessibility index: literature review. *Center of Transportation Research, University of Texas at Austin, Springfield*, 2000.
- [24] T. C. Blanchard and T. L. Matthews. Retail concentration, food deserts, and food-disadvantaged communities in rural america. *Remaking the North America Food System*, pages 201–15, 2007.
- [25] J. P. Block, S. K. Condon, K. Kleinman, J. Mullen, S. Linakis, S. Rifas-Shiman, and M. W. Gillman. Consumers estimation of calorie content at fast food restaurants: cross sectional observational study. *BMJ*, 346:f2907, 2013.
- [26] J. P. Block, R. A. Scribner, and K. B. DeSalvo. Fast food, race/ethnicity, and income: a geographic analysis. *American journal of preventive medicine*, 27(3):211–217, 2004.
- [27] K. M. Booth, M. M. Pinkston, and W. S. C. Poston. Obesity and the built environment. *Journal of the American Dietetic Association*, 105(5):110–117, 2005.
- [28] U. D. Bordoloi, D. L. Kao, and H.-W. Shen. Visualization techniques for spatial probability density function data. *Data Science Journal*, 3:153–162, 2004.

- [29] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramée, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. *Eurographics State of the Art Reports*, pages 39–63, 2013.
- [30] U. Brandes, M. Eiglsperger, J. Lerner, and C. Pich. *Graph Markup Language (GraphML)*, pages 517–541. CRC Press: Boca Raton, 2013.
- [31] J. Buhl, J. Gautrais, N. Reeves, R. Sole, R. Valverde, P. Kuntz, and G. Theraulaz. Topological patterns in street networks of self-organized urban settlements. *European Physical Journal B*, 49:513–522, 2006.
- [32] T. Burgoine, S. Alvanides, and A. A. Lake. Assessing the obesogenic environment of north east england. *Health & place*, 17(3):738–747, 2011. [Online; accessed 22-September-2016].
- [33] T. Burgoine, N. G. Forouhi, S. J. Griffin, N. J. Wareham, and P. Monsivais. Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in cambridgeshire, uk: population based, cross sectional study. *BMJ*, 348:g1464, 2014.
- [34] T. Burgoine, A. A. Lake, E. Stamp, S. Alvanides, J. C. Mathers, and A. J. Adamson. Changing foodscapes 1980–2000, using the ash30 study. *Appetite*, 53(2):157–165, 2009.
- [35] A. M. Buyx. Personal responsibility for health as a rationing criterion: why we dont like it and why maybe we should. *Journal of medical ethics*, 34(12):871–874, 2008.
- [36] J. W. Byrom, D. J. Bennison, T. Hernández, and P. D. Hooper. The use of geographical data and information in retail locational planning. *Journal of Targeting, Measurement and Analysis for Marketing*, 9(3):219–229, 2001.
- [37] A. M. Cali and S. Caprio. Obesity in children and adolescents. *The Journal of Clinical Endocrinology & Metabolism*, 93(11_supplement_1):s31–s36, 2008.
- [38] A. Cardillo, S. Scellato, V. Latora, and S. Porta. Structural properties of planar graphs of urban street patterns. *Physical Review E*, 73:066107, 2006.

- [39] C. E. Caspi, G. Sorensen, S. Subramanian, and I. Kawachi. The local food environment and diet: a systematic review. *Health & place*, 18(5):1172–1187, 2012.
- [40] N. Cavill and H. Rutter. Healthy people, healthy places briefing: obesity and the environment: regulating the growth of fast food outlets. *Public Health England*, 2013.
- [41] Z. Chalabi and T. Lorenc. Using agent-based models to inform evaluation of complex interventions: examples from the built environment. *Preventive medicine*, 57(5):434–435, 2013.
- [42] J. Charlene. Hot food takeaway supplementary planning document draft for consultation. http://www.solihull.gov.uk/Portals/0/Planning/LDF/Draft_HFT_SPD.pdf, 2014. Accessed: March 13, 2017.
- [43] H. Charreire, R. Casey, P. Salze, C. Simon, B. Chaix, A. Banos, D. Badariotti, C. Weber, and J.-M. Oppert. Measuring the food environment using geographical information systems: a methodological review. *Public health nutrition*, 13(11):1773–1785, 2010.
- [44] L. H. Commission. Better health for london, 2014.
- [45] K. G. Congdon and D. C. Congdon. A systemic approach to art education and mental health. *Visual Arts Research*, 12(12):73–79, 1986.
- [46] Council – City of Bradford. The hot food takeaway supplementary planning document. <https://www.bradford.gov.uk/media/3039/hotfoodtakeawaysupplementaryplanningdocument.pdf>, 2014. Accessed: March 13, 2017.
- [47] Council – Halton Borough and Square, Gloucester. Hot food takeaway supplementary planning document. <http://moderngov.halton.gov.uk/documents/s23068/29%2003%2012%20EB%20Adoption%20Hot%20Food%20SPD%20final%20V1.doc.pdf>, 2012. Accessed: March 13, 2017.

- [48] Council – London Borough of Barking & Dagenham. Saturation point - addressing the health impact of hot food takeaways. <https://www.lbbd.gov.uk/wp-content/uploads/2014/10/Saturation-Point-SPD-Addressing-the-Health-Impacts-of-Hot-Food-Takeaway.pdf>, 2010. Accessed: March 13, 2017.
- [49] I. Council. Location and concentration of uses supplementary planning document. https://www.islington.gov.uk/~/_/media/sharepoint-lists/public-records/planningandbuildingcontrol/publicity/publicconsultation/20162017/20160505locationandconcentrationofusesspdadoptedapril2016, 2016. Accessed: March 13, 2017.
- [50] W. B. Council. Hot food takeaways supplementary planning document. https://www.warrington.gov.uk/download/downloads/id/8680/hot_food_takeaway_spd_april_2014.pdf, 2014. Accessed: March 13, 2017.
- [51] A. M. Craigie, J. N. Matthews, A. J. Rugg-Gunn, A. A. Lake, J. C. Mathers, and A. J. Adamson. Raised adolescent body mass index predicts the development of adiposity and a central distribution of body fat in adulthood: a longitudinal study. *Obesity facts*, 2(3):150–156, 2009.
- [52] R. Crutzen and P. Giabbanelli. Using classifiers to identify binge drinkers based on drinking motives. *Substance use & misuse*, 49(1–2):110–15, 2014.
- [53] R. Crutzen, P. Giabbanelli, A. Jander, L. Mercken, and H. de Vries. Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: building classifiers on adolescent-parent paired data. *BMC public health*, 15(1):1, 2015.
- [54] C. Cui and Z. Han. Spatial patterns of retail stores using pois data in zhengzhou, china. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on*, pages 88–92. IEEE, 2015.

- [55] S. C. Cummins, L. McKay, and S. MacIntyre. Mcdonalds restaurants and neighborhood deprivation in scotland and england. *American journal of preventive medicine*, 29(4):308–310, 2005. [Online; accessed 22-September-2016].
- [56] V. Dabbaghian, P. Jackson, V. Spicer, and K. Wuschke. A cellular automata model on residential migration in response to neighborhood social dynamics. *Mathematical and Computer Modelling*, 52(9):1752–1762, 2010.
- [57] C. d’Aspremont, J. J. Gabszewicz, and J.-F. Thisse. On hotelling’s” stability in competition”. *Econometrica: Journal of the Econometric Society*, pages 1145–1150, 1979.
- [58] F. A. Davidson. Mathematical modelling of mycelia: a question of scale. *Fungal Biology Reviews*, 21(1):30 – 41, 2007.
- [59] P. L. Day and J. Pearce. Obesity-promoting food environments and the spatial clustering of food outlets around schools. *American journal of preventive medicine*, 40(2):113–121, 2011. [Online; accessed 22-September-2016].
- [60] I. Demir, C. Dick, and R. Westermann. Multi-charts for comparative 3d ensemble visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2694–2703, 2014.
- [61] Department of Health. Children poverty at local authorities in england. <http://www.endchildpoverty.org.uk/poverty-in-your-area/>, 2014. Accessed: March 13, 2017.
- [62] A. Deutsch and S. Dormann. *Cellular automaton modeling of biological pattern formation: characterization, applications, and analysis*. Springer Science & Business Media, 2007.
- [63] L. Drasic and P. J. Giabbanelli. Exploring the interactions between physical well-being, and obesity. *Canadian Journal of Diabetes*, 39:S12–S13, 2015.
- [64] D. L. Dustin, K. S. Bricker, and K. A. Schwab. People and nature: Toward an ecological model of health promotion. *Leisure Sciences*, 32(1):3–14, 2009.

- [65] B. C. Eaton and R. G. Lipsey. Freedom of entry and the existence of pure profit. *The Economic Journal*, 88(351):455–469, 1978.
- [66] A. Ellaway, L. Macdonald, K. Lamb, L. Thornton, P. Day, and J. Pearce. Do obesity-promoting food environments cluster around socially disadvantaged schools in glasgow, scotland? *Health & place*, 18(6):1335–1340, 2012.
- [67] L. H. Encinas, S. H. White, A. M. del Rey, and G. R. Sanchez. Modelling forest fire spread using hexagonal cellular automata. *Applied Mathematical Modelling*, 31(6):1213 – 1227, 2007.
- [68] D. Eppstein and J. Wang. Fast approximation of centrality. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '01*, pages 228–229, 2001.
- [69] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann. Time-hierarchical clustering and visualization of weather forecast ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):831–840, 2017.
- [70] D. Finegood. The complex systems science of obesity. In J. Cawley, editor, *The oxford handbook of the social science of obesity*, pages 208–236. Oxford University Press, 2011.
- [71] D. T. Finegood. The importance of systems thinking to address obesity. In *Obesity Treatment and Prevention: New Directions*, volume 73, pages 123–137. Karger Publishers, 2012.
- [72] E. A. Finkelstein, H. Thompson, and W. Dietz. Obesity and severe obesity forecasts through 2030.
- [73] D. for Communities and L. Government. English indices of deprivation 2015: Statistical release - main findings. <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>, 2015. Accessed: July 12, 2017.
- [74] C. for Diet and A. Research. Neighbourhood food environments, diet and health: research policy meeting. <http://www.cedar.iph.cam.ac.uk/healthy-place-2014/programme/>, 2014. Accessed: July 12, 2017.

- [75] N. I. for Health and C. E. (NICE)". Type 2 diabetes prevention: population and community-level interventions. *Public health guideline*, 35, 2011.
- [76] L. K. Fraser, G. P. Clarke, J. E. Cade, and K. L. Edwards. Fast food and obesity: a spatial analysis in a large united kingdom population of children aged 13–15. *American journal of preventive medicine*, 42(5):e77–e85, 2012. [Online; accessed 22-September-2016].
- [77] L. K. Fraser, K. L. Edwards, J. Cade, and G. P. Clarke. The geography of fast food outlets: a review. *International journal of environmental research and public health*, 7(5):2290–2308, 2010.
- [78] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
- [79] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. *Proceedings of CHI*, pages 3237–2346, 2013.
- [80] M. Fuentes and M. Kuperman. Cellular automata and epidemiological models with spatial dependence. *Physica A: statistical mechanics and its applications*, 267(3–4):471–486, 1999.
- [81] B. A. Furtado, P. A. Sakowski, and M. H. Tóvolli. Modeling complex systems for public policies. *Brasilia: IPEA, Instituto de Pesquisa Economica Aplicada*, 2015.
- [82] S. Galea, A. V. D. Roux, C. P. Simon, and G. A. Kaplan. *Using complex systems approaches in the study of population health and health disparities: seven observations*, pages 297–304. Westphalia Press, 2017.
- [83] T. Ghosh, R. L. Powell, C. D. Elvidge, K. E. Baugh, P. C. Sutton, and S. Anderson. Shedding light on the global distribution of economic activity. *The Open Geography Journal*, 3(1), 2010.

- [84] P. Giabbanelli and J. Adams. Identifying small groups of foods that can predict achievement of key dietary recommendations: data mining of the uk national diet and nutrition survey, 2008-12. *Public Health Nutrition*, 2016.
- [85] P. Giabbanelli and R. Crutzen. An agent-based social network model of binge drinking among dutch adults. *Journal of Artificial Societies and Social Simulation*, 16(2):10, 2013.
- [86] P. Giabbanelli, R. Flarsheim, C. Vesuvala, and L. Drasic. Developing technology to support policymakers in taking a systems science approach to obesity and well-being. *Obesity Reviews*, 17:194–195, 2016.
- [87] P. J. Giabbanelli, J. Adams, and V. S. Pillutla. Feasibility and framing of interventions based on public support: Leveraging text analytics for policymakers. In G. Meiselwitz, editor, *Proc of HCI International 2016*, pages 188–200. Springer, 2016.
- [88] P. J. Giabbanelli, A. Alimadad, V. Dabbaghian, and D. T. Finegood. Modeling the influence of social networks and environment on energy balance and obesity. *Journal of Computational Science*, 3(1):17–27, 2012.
- [89] P. J. Giabbanelli, G. J. Babu, and M. Baniukiewicz. A novel visualization environment to support modelers in analyzing data generated by cellular automata. In V. G. Duffy, editor, *Digital Human Modeling: Applications in Health, Safety, Ergonomics and Risk Management: 7th International Conference, DHM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings*, pages 529–540. Springer International Publishing, Cham, 2016.
- [90] P. J. Giabbanelli and R. Crutzen. Creating groups with similar expected behavioural response in randomized controlled trials: a fuzzy cognitive map approach. *BMC Medical Research Methodology*, 14(1):130, 2014.

- [91] P. J. Giabbanelli and R. Crutzen. Using agent-based models to develop public policy about food behaviours: Future directions and recommendations. *Computational and Mathematical Methods in Medicine*, 2017:5742629, 2017.
- [92] P. J. Giabbanelli et al. Modelling the joint effect of social determinants and peers on obesity among canadian adults. In V. Dabbaghian and V. K. Mago, editors, *Theories and Simulations of Complex Social Systems*, pages 145–160. Springer-Verlag Berlin Heidelberg, 2014.
- [93] P. J. Giabbanelli, S. A. Gray, and P. Aminpour. Combining fuzzy cognitive maps with agent-based modeling: Frameworks and pitfalls of a powerful hybrid modeling approach to understand human-environment interactions. *Environmental Modelling & Software*, 95:320 – 325, 2017.
- [94] P. J. Giabbanelli and P. J. Jackson. Using visual analytics to support the integration of expert knowledge in the design of medical models and simulations. *Procedia Computer Science*, 51:755–764, 2015. Proceedings of the 2015 International Conference On Computational Science (ICCS).
- [95] P. J. Giabbanelli, T. Tornsney-Weir, and V. K. Mago. A fuzzy cognitive map of the psychosocial determinants of obesity. *Applied Soft Systems*, 12:3711–3724, 2012.
- [96] N. Gilbert and K. Troitzsch. *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.
- [97] B. Giles et al. Integrating conventional science and aboriginal perspectives on diabetes using fuzzy cognitive maps. *Soc Sci Med*, 64(3):562–576, 2007.
- [98] K. Glanz, J. F. Sallis, B. E. Saelens, and L. D. Frank. Nutrition environment measures survey in stores (nems-s): Development and evaluation. *American Journal of Preventive Medicine*, 32(4):282 – 289, 2007.
- [99] R. Guimera and L. A. N. Amaral. Modeling the world-wide airport network. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):381–385, 2004.

- [100] D. Haehn, S. Knowles-Barley, M. Roberts, J. Beyer, N. Kasthuri, J. W. Lichtman, and H. Pfister. Design and evaluation of interactive proofreading tools for connectomics. *IEEE transactions on visualization and computer graphics*, 20(12):2466–2475, 2014.
- [101] P. Haggett and R. Chorley. *Network analysis in geography*. Edward Arnold: London, 1969.
- [102] K. D. Hall et al. Quantification of the effect of energy imbalance on bodyweight. *The Lancet*, 378(9793):826 – 837, 2011.
- [103] T. K. Hamid. *Thinking in Circles about obesity*. Springer, 2nd edition, 2009.
- [104] M. C. Hao, M. Marwah, H. Janetzko, U. Dayal, D. A. Keim, D. Patnaik, N. Ramakrishnan, and R. K. Sharma. Visual exploration of frequent patterns in multivariate time series. *Information Visualization*, 11(1):71–83, 2012.
- [105] L. J. Harnack, S. A. French, J. M. Oakes, M. T. Story, R. W. Jeffery, and S. A. Rydell. Effects of calorie labeling and value size pricing on fast food meal choices: results from an experimental trial. *International Journal of Behavioral Nutrition and Physical Activity*, 5(1):63, 2008.
- [106] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.
- [107] "Health and S. C. I. Center". National child measurement program – england, 2013-2014, 2015.
- [108] M. R. Henzinger, P. Klein, S. Rao, and S. Subramanian. Faster shortest-path algorithms for planar graphs. *J. Comput. Syst. Sci.*, 55(1):3–23, 1997.
- [109] T. Hernandez and D. Bennison. The art and science of retail location decisions. *International Journal of Retail & Distribution Management*, 28(8):357–367, 2000.
- [110] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

- [111] N. Hex, C. Bartlett, D. Wright, M. Taylor, and D. Varley. Estimating the current and future costs of type 1 and type 2 diabetes in the uk, including direct health costs and indirect societal and productivity costs. *Diabetic Medicine*, 29(7):855–862, 2012.
- [112] B. Hillier and J. Hanson. *The Social Logic of Space*. Cambridge University Press: Cambridge, 1984.
- [113] T. Höllt, F. M. de Matos Ravanelli, M. Hadwiger, and I. Hoteit. Visual analysis of reservoir simulation ensembles. In *Proceedings of the Workshop on Visualisation in Environmental Sciences*, EnvirVis '16, pages 1–5. Eurographics Association, 2016.
- [114] P. Holme. Congestion and centrality in traffic flow on complex networks. *Advances in Complex Systems*, 6(02):163–176, 2003.
- [115] T. J. Holmes and J. J. Stevens. Spatial distribution of economic activities in north america. *Handbook of regional and urban economics*, 4:2797–2843, 2004.
- [116] J. Homer, B. Milstein, W. Dietz, D. Buchner, and E. Majestic. Obesity population dynamics: exploring historical growth and plausible futures in the us. *24th International System Dynamics Conference*, 2006.
- [117] G. M. Jacquez. Spatial cluster analysis. *The handbook of geographic information science*, 395:416, 2008.
- [118] R. Jain. *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. John Wiley & Sons, 1990.
- [119] M. James and G. Beer. Careless eating costs lives. <http://www.2020health.org/2020health/Publications/Publications-2014/CarelessEatingCostsLives.html>, 2014. Accessed: July 12, 2017.

- [120] H. Janicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1384–1391, 2007.
- [121] M. Janssen and M. A. Wimmer. Introduction to policy-making in the digital age. In *Policy Practice and Digital Science*, pages 1–14. Springer, 2015.
- [122] A. Jennings, A. Welch, A. P. Jones, F. Harrison, G. Bentham, E. M. Van Sluijs, S. J. Griffin, and A. Cassidy. Local food outlets, weight status, and dietary intake: associations in children aged 9–10 years. *American journal of preventive medicine*, 40(4):405–410, 2011.
- [123] P. Jensen, J. Boisson, and H. Larralde. Aggregation of retail stores. *Physica A: Statistical Mechanics and its Applications*, 351(2):551–570, 2005.
- [124] L. M. Johnston, C. L. Matteson, and D. T. Finegood. Systems science and obesity policy: a novel framework for analyzing and rethinking population-level planning. *American journal of public health*, 104(7):1270–1278, 2014.
- [125] E. Kamateri et al. A comparative analysis of tools and technologies for policy making. In M. Janssen, M. A. Wimmer, and A. Deljoo, editors, *Policy Practice and Digital Science: Integrating Complex Systems, Social Simulation and Public Administration in Policy Research*, pages 125–156. Springer International Publishing, 2015.
- [126] I. Karafyllidis and A. Thanailakis. A model for predicting forest fire spreading using cellular automata. *Ecological Modelling*, 99:87–97, 1997.
- [127] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, 2013.
- [128] D. Keim. Designing pixel-oriented visualization techniques: theory and applications. *IEEE Transactions on Visualization and computer graphics*, 6(1):59–78, 2000.

- [129] E. Kenah and J. M. Robins. Network-based analysis of stochastic sir epidemic models with random and proportionate mixing. *Journal of theoretical biology*, 249(4):706–722, 2007.
- [130] M. Kersten-Oertel, S. J.-S. Chen, and D. L. Collins. An evaluation of depth enhancing perceptual cues for vascular volume visualization in neurosurgery. *IEEE transactions on visualization and computer graphics*, 20(3):391–403, 2014.
- [131] G. Kindlmann. Superquadric tensor glyphs. In *Proceedings of the Sixth Joint Eurographics - IEEE TCVG Conference on Visualization, VISSYM'04*, pages 147–154. Eurographics Association, 2004.
- [132] R. Knapen, S. Janssen, O. Roosenschoon, P. Verweij, W. de Winter, M. Uiterwijk, and J.-E. Wien. Evaluating openmi as a model integration platform across disciplines. *Environmental Modelling & Software*, 39:274 – 282, 2013.
- [133] H. Konishi. Concentration of competing retail stores. *Journal of Urban Economics*, 58(3):488–512, 2005.
- [134] Z. Konyha, K. Matkovic, and H. Hauser. Interactive visual analysis in engineering: a survey. *Proceedings of the Spring Conference on Computer Graphics*, pages 31–38, 2009.
- [135] D. Koschützki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski. Centrality indices. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations*, pages 16–61. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
- [136] P. Kothur, C. Witt, M. Sips, N. Marwan, S. Schinkel, and D. Dransch. Visual analytics for correlation-based comparison of time series ensembles. *Computer Graphics Forum*, 34(3):411–420, 2015.
- [137] G. Y. Krasnikov, I. V. Matyushkin, and S. V. Korobov. Visualization of cellular automata in nanotechnology. *Modelling of Artificial Intelligence*, 3(3):98–120, 2014.

- [138] R. E. Krider and D. S. Putler. Which birds of a feather flock together? clustering and avoidance patterns of similar retail outlets. *Geographical Analysis*, 45(2):123–149, 2013.
- [139] S. Kumanyika, R. Jeffery, A. Morabia, C. Ritenbaugh, and V. Antipatis. Obesity prevention: the case for action. *International journal of obesity*, 26(3):425, 2002.
- [140] K. Kurzhals, M. Hlawatsch, F. Heimerl, M. Burch, T. Ertl, and D. Weiskopf. Gaze stripes: Image-based visualization of eye tracking data. *IEEE transactions on visualization and computer graphics*, 22(1):1005–1014, 2016.
- [141] A. Lake and T. Townshend. Obesogenic environments: exploring the built and food environments. *The Journal of the Royal society for the Promotion of Health*, 126(6):262–267, 2006.
- [142] T. Lang and V. Schoen. Food, the uk and the eu: Brexit or bremain? Working paper, Food Research Collaboration, UK, March 2016. © This working paper is copyright of the authors.
- [143] K. Lawonn, E. Trostmann, B. Preim, and K. Hildebrandt. Visualization and extraction of carvings for heritage conservation. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):801–810, 2017.
- [144] D. Levy et al. Simulation models of obesity: a review of the literature and implications for research and policy. *Obesity reviews*, 12:378–394, 2010.
- [145] A. J. Lind and S. Bruckner. Comparing cross-sections and 3d renderings for surface matching tasks using physical ground truths. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):781–790, 2017.
- [146] Local Government Association. Tackling obesity: local government’s new public health role. http://www.local.gov.uk/c/document_library/get_file?uuid=dc226049-df94-487e-be70-96bdcb4a9115&groupId=10180, 2012. Accessed: March 13, 2017.

- [147] Local Government Association. Changing behaviours in public health: to nudge or to shove? <http://www.local.gov.uk/documents/10180/11463/Changing+behaviours+in+public+health+-+to+nudge+or+to+shove/5ae3b9c8-e476-495b-89b4-401d70e1e2aa>, 2013. Accessed: March 13, 2017.
- [148] D. López, L. Oehlberg, C. Doger, and T. Isenberg. Towards an understanding of mobile touch navigation in a stereoscopic viewing environment for 3d data exploration. *IEEE transactions on visualization and computer graphics*, 22(5):1616–1629, 2016.
- [149] J. Lopez-Zetina, H. Lee, and R. Friis. The link between obesity and the built environment. evidence from an ecological analysis of obesity and vehicle miles of travel in california. *Health & place*, 12(4):656–664, 2006.
- [150] K. L. Ma. Machine learning to boost the next generation of visualization technology. *IEEE Computer Graphics and Applications*, 27(5):6–9, 2007.
- [151] S. Macintyre, L. McKay, S. Cummins, and C. Burns. Out-of-home food outlets and area deprivation: case study in glasgow, uk. *International Journal of Behavioral Nutrition and Physical Activity*, 2(1):16, 2005.
- [152] V. K. Mago, L. Bakker, E. I. Papageorgiou, A. Alimadad, P. Borwein, and V. Dabbaghian. Fuzzy cognitive maps and cellular automata: An evolutionary approach for social systems modelling. *Applied Soft Computing*, 12(12):3771–3784, 2012.
- [153] E. Maguire, T. Burgoine, and P. Monsivais. Area deprivation and the food environment over time: a repeated cross-sectional study on takeaway outlet density and supermarket presence in norfolk, uk, 1990-2008. *Health & Place*, 33:142–147, 2015. [Online; accessed 22-September-2016].
- [154] E. R. Maguire and P. Monsivais. Socio-economic dietary inequalities in uk adults: an updated picture of key food groups and nutrients from national surveillance data. *British Journal of Nutrition*, 113(01):181–189, 2015.

- [155] K. Matkovic, M. Jelovic, J. Juric, Z. Konyha, and D. Gracanin. Interactive visual analysis and exploration of injection systems simulations. In *VIS 05. IEEE Visualization, 2005.*, pages 391–398, 2005.
- [156] R. A. McKinnon, J. Reedy, M. A. Morrissette, L. A. Lytle, and A. L. Yaroch. Measures of the food environment: a compilation of the literature, 1990–2007. *American journal of preventive medicine*, 36(4):S124–S133, 2009.
- [157] K. R. McLeroy, D. Bibeau, A. Steckler, and K. Glanz. An ecological perspective on health promotion programs. *Health Education & Behavior*, 15(4):351–377, 1988.
- [158] M. Minkler. Personal responsibility for health? a review of the arguments and the evidence at centurys end. *Health Education & Behavior*, 26(1):121–141, 1999.
- [159] M. Mokhtari, E. Boivin, and D. Laurendeau. *Virtual, Augmented and Mixed Reality. Systems and Applications: 5th International Conference, VAMR 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II*, chapter Making Sense of Large Datasets in the Context of Complex Situation Understanding, pages 251–260. Springer-Verlag Berlin Heidelberg, 2013.
- [160] P. Monsivais, A. Aggarwal, and A. Drewnowski. Are socio-economic disparities in diet quality explained by diet cost? *Journal of epidemiology and community health*, pages jech–2010, 2010.
- [161] M. Murphy, H. Badland, M. J. Koohsari, T. Astell-Burt, G. Trapp, K. Villanueva, S. Mavoa, M. Davern, and B. Giles-Corti. Indicators of a health-promoting local food environment: a conceptual framework to inform urban planning policy and practice. *Health Promotion Journal of Australia*, 2016.
- [162] U. H. C. Network”. Information booklet, 2011.
- [163] M. Noble, G. Wright, C. Dibben, G. Smith, M. McLennan, C. Anttila, H. Barnes, C. Mokhtar, S. Noble, and D. Avenell. The english indices of deprivation 2004: Report to the office of the deputy prime minister, 2014.

- [164] T. Nocke, M. Flechsig, and U. Bohm. Visual exploration and evaluation of climate-related simulation data. In *2007 Winter Simulation Conference*, pages 703–711, 2007.
- [165] M. Novotny and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900, 2006.
- [166] S. I. O’Donnell, S. L. Hoerr, J. A. Mendoza, and E. T. Goh. Nutrient quality of fast food kids meals. *The American journal of clinical nutrition*, 88(5):1388–1395, 2008.
- [167] D. of Health. *Healthy lives, healthy people: Our strategy for public health in England*, volume 7985. The Stationery Office, 2010.
- [168] A. of Royal Medical Colleges. Measuring up: The medical professions prescription for the nations obesity crisis, 2013.
- [169] P. O’Leary, J. Ahrens, S. Jourdain, S. Wittenburg, D. H. Rogers, and M. Petersen. Cinema image-based in situ analysis and visualization of mpas-ocean simulations. *Parallel Computing*, 55:43 – 48, 2016. Visualization and Data Analytics for Scientific Discovery.
- [170] I. Omer and R. Goldblatt. Spatial patterns of retail activity and street network structure in new and traditional israeli cities. *Urban Geography*, 37(4):629–649, 2016.
- [171] Organisation for Economic Co-operation and Development. Global science forum. applications of complexity science for public policy: new tools for finding unanticipated consequences and unrealized opportunities. <http://www.oecd.org/science/sci-tech/43891980.pdf>, 2009. Accessed: March 28, 2017.
- [172] M. A. Papas, A. J. Alberg, R. Ewing, K. J. Helzlsouer, T. L. Gary, and A. C. Klassen. The built environment and obesity. *Epidemiologic reviews*, 2007.
- [173] R. E. Park. The city: Suggestions for the investigation of human behavior in the city environment. *American journal of sociology*, 20(5):577–612, 1915.

- [174] J. Pearce, T. Blakely, K. Witten, and P. Bartie. Neighborhood deprivation and access to fast-food retailing: a national study. *American journal of preventive medicine*, 32(5):375–382, 2007.
- [175] G. A. Picone, D. B. Ridley, and P. A. Zandbergen. Distance decreases with differentiation: Strategic agglomeration by retailers. *International Journal of Industrial Organization*, 27(3):463–473, 2009.
- [176] M. Polèse and É. Champagne. Location matters: comparing the distribution of economic activity in the canadian and mexican urban systems. *International Regional Science Review*, 22(1):102–132, 1999.
- [177] S. Porta, V. Latora, F. Wang, S. Rueda, E. Strano, S. Scellato, A. Cardillo, E. Belli, F. Cardenas, B. Cormenzana, et al. Street centrality and the location of economic activities in barcelona. *Urban Studies*, 49(7):1471–1488, 2012.
- [178] S. Porta, E. Strano, V. Iacoviello, R. Messori, V. Latora, A. Cardillo, F. Wang, and S. Scellato. Street centrality and densities of retail and services in bologna, italy. *Environment and Planning B: Planning and design*, 36(3):450–465, 2009.
- [179] S. Porta, A. Venerandi, M. Zanella, and O. Romice. The form of gentrification: Common morphological patterns in five gentrified areas of london, uk. Workingpaper, University of Strathclyde Glasgow, 11 2014.
- [180] S. Pratt, P. Giabbanelli, P. Jackson, and V. Mago. Rebel with many causes: A computational model of insurgency. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*, pages 90–95, 2012.
- [181] M. J. Proudfoot. City retail structure. *Economic Geography*, 13(4):425–428, 1937.
- [182] S. R. Proulx, D. E. Promislow, and P. C. Phillips. Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, 20(6):345–353, 2005.

- [183] Provincial Health Services Authority. Technical report, from weight to well-being: Time for a shift in paradigms? http://www.phsa.ca/population-public-health-site/Documents/W2WBTechnicalReport_20130208FINAL.pdf, 2013. Accessed: March 22, 2017.
- [184] Public Health England. National child measurement programme 2011/12 to 2013/14. prevalence of obesity by area of child residence, electoral ward (2014) and middle super output area (2011). <http://www.noo.org.uk/visualisation/>, 2014. Accessed: March 13, 2017.
- [185] Public Health England. National child measurement programme 2013-2014. <http://content.digital.nhs.uk/article/2021/Website-Search?productid=16565&q=National+Child+Measurement+Programme+2013-2014&sort=Relevance&size=100&page=1&area=both#top>, 2015. Accessed: March 13, 2017.
- [186] Public Health England. ["https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2"](https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2), March 2017. Accessed: April 12, 2017.
- [187] Public Health England Obesity Knowledge & Intelligence team. Making the case for tackling obesity- why invest? https://www.noo.org.uk/slide_sets, 2015. Accessed: March 13, 2017.
- [188] E. Rana, P. J. Giabbanelli, N. H. Balabhadrapathruni, X. Li, and V. K. Mago. Exploring the relationship between adherence to treatment and viral load through a new discrete simulation model of hiv infectivity. In *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM PADS '15, pages 145–156, New York, NY, USA, 2015. ACM.
- [189] K. Reda, A. E. Johnson, M. E. Papka, and J. Leigh. Effects of display size and resolution on user behavior and insight acquisition in visual exploration. In *Proceedings of the 33rd Annual*

- ACM Conference on Human Factors in Computing Systems, CHI '15*, pages 2759–2768, New York, NY, USA, 2015. ACM.
- [190] D. D. Reidpath, C. Burns, J. Garrard, M. Mahoney, and M. Townsend. An ecological study of the relationship between social and environmental determinants of obesity. *Health & place*, 8(2):141–145, 2002.
- [191] K. Resnicow and R. Vaughan. A chaotic view of behavior change: a quantum leap for health promotion. *International Journal of Behavioral Nutrition and Physical Activity*, 3(1):25, 2006.
- [192] J. Reynolds and S. Wood. Location decision making in retail firms: evolution and challenge. *International Journal of Retail & Distribution Management*, 38(11/12):828–845, 2010.
- [193] A. Ravis and P. Sheeran. Descriptive norms as an additional predictor in the theory of planned behaviour: A meta-analysis. *Current psychology*, 22(2):218–233, 2002.
- [194] S. Rizzi, M. Hereld, J. Insley, P. Malakar, M. E. Papka, T. Uram, and V. Vishwanath. Coupling lammeps and the vl3 framework for co-visualization of atomistic simulations. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1038–1042. IEEE Press, May 2016.
- [195] S. Robinson. *Simulation: the practice of model development and use*. Palgrave Macmillan, 2014.
- [196] T. Ropinski and B. Preim. Taxonomy and usage guidelines for glyph-based medical visualization. *Proc. Simulation and Visualization (SimVis)*, pages 121–138, 2008.
- [197] A. Ross and F. D. Consultancy. Obesity-based policies to restrict hot food takeaways: progress by local planning authorities in england. *Final Draft Consultancy*, 2013.
- [198] C. Ross, C. D. Carothers, M. Mubarak, P. Carns, R. Ross, J. K. Li, and K.-L. Ma. Visual data-analytics of large-scale parallel discrete-event simulations. In *Proceedings of the 7th*

- International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computing Systems*, PMBS '16, pages 87–97. IEEE Press, 2016.
- [199] S. Ross. *Structure of decision: the cognitive maps of political elites*, chapter Complexity and the presidency: gouverneur Morris in the constitutional convention, pages 96–112. Princeton University Press, 1974.
- [200] R. Rothermel. *A Mathematical Model for Predicting Fire Spread in Wildland Fuels*. A Mathematical Model for Predicting Fire Spread in Wildland Fuels. Intermountain Forest & Range Experiment Station, Forest Service, U.S. Department of Agriculture, 1972.
- [201] M. Ruderman. The ecological model in public health. <http://www.jhsph.edu/research/centers-and-institutes/womens-and-childrens-health-policy-center/eco-model/eco-model.html>, 2013. Accessed: March 13, 2017.
- [202] Y. Rui and Y. Ban. Exploring the relationship between street centrality and land use in stockholm. *International Journal of Geographical Information Science*, 28(7):1425–1438, 2014.
- [203] P. E. Rummo, D. K. Guilkey, S. W. Ng, B. M. Popkin, K. R. Evenson, and P. Gordon-Larsen. Beyond supermarkets: Food outlet location selection in four us cities over time. *American Journal of Preventive Medicine*, 2016.
- [204] H. Rutter. The single most important intervention to tackle obesity, 2012.
- [205] H. Rutter, K. Glonti, and J. Lakerveld. The way ahead: where next for research into obesogenic environments? *Obesity reviews*, 17(S1):108–109, 2016.
- [206] A. Sevtsuk. *Path and place: a study of urban geometry and retail activity in Cambridge and Somerville, MA*. PhD thesis, Massachusetts Institute of Technology, 2010.
- [207] D. A. Shoham, R. Hammond, H. Rahmandad, Y. Wang, and P. Hovmand. Modeling social norms and social influence in obesity. *Current epidemiology reports*, 2(1):71–79, 2015. [Online; accessed 22-September-2016].

- [208] Q. Shu, H. Guo, J. Liang, L. Che, J. Liu, and X. Yuan. Ensemblegraph: Interactive visual analysis of spatiotemporal behaviors in ensemble simulation data. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pages 56–63, 2016.
- [209] P. A. Simon, D. Kwan, A. Angelescu, M. Shih, and J. E. Fielding. Proximity of fast food restaurants to schools: do neighborhood income and type of school matter? *Preventive medicine*, 47(3):284–288, 2008.
- [210] S. Sinclair and J. Winkler. The school fringe: What pupils buy and eat from shops surrounding secondary schools. *Nutrition Policy Unit, London Metropolitan University*, 2008.
- [211] S. Sinclair and J. Winkler. The school fringe: from research to action. policy options within schools on the fringe. *Nutrition Policy Unit (London Metropolitan University)*, 2009.
- [212] A. S. Singh, C. Mulder, J. W. Twisk, W. Van Mechelen, and M. J. Chinapaw. Tracking of childhood overweight into adulthood: a systematic review of the literature. *Obesity reviews*, 9(5):474–488, 2008.
- [213] C. Siokou, R. Morgan, and A. Shiell. Group model building: a participatory approach to understanding and acting on systems. *Public Health Research & Practice*, 25(1), 2014.
- [214] L. Skyttner. *General Systems Theory*. World Scientific, 2006.
- [215] D. Smith, S. Cummins, C. Clark, and S. Stansfeld. Does the local food environment around schools affect diet? longitudinal associations in adolescents attending secondary schools in east london. *BMC public health*, 13(1):70, 2013.
- [216] H. Song, B. Lee, B. Kim, and J. Seo. Diffmatrix: Matrix-based interactive visualization for comparing temporal trends. In M. Meyer and T. Weinkauff, editors, *Proceedings of the 2012 Eurographics Conference on Visualization (EuroVis)*, 2012.
- [217] K. Stahl. Therories of urban business location. *Handbook of regional and urban economics*, 2:759–820, 1987.

- [218] F. Stok, D. Ridder, E. Vet, and J. Wit. Don't tell me what i should do, but what others do: The influence of descriptive and injunctive peer norms on fruit consumption in adolescents. *British journal of health psychology*, 19(1):52–64, 2014.
- [219] D. Stokols. Translating social ecological theory into guidelines for community health promotion. *American journal of health promotion*, 10(4):282–298, 1996.
- [220] M. Story, K. M. Kaphingst, and S. French. The role of schools in obesity prevention. *The future of children*, pages 109–142, 2006.
- [221] E. Strano, M. Viana, L. da Fontoura Costa, A. Cardillo, S. Porta, and V. Latora. Urban street networks, a comparative analysis of ten european cities. *Environment and Planning B: Planning and Design*, 40(6):1071–1086, 2013.
- [222] J. Struben, D. Chan, and L. Dube. Policy insights from the nutritional food market transformation model: the case of obesity prevention. *New York Academy of Science Annals*, 1331:57–75, 2014.
- [223] R. Sturm. Disparities in the food environment surrounding us middle and high schools. *Public health*, 122(7):681–690, 2008.
- [224] S. J. E. Taylor, A. Khan, K. L. Morse, A. Tolk, L. Yilmaz, and J. Zander. Grand challenges on the theory of modeling and simulation. In *Proceedings of the Symposium on Theory of Modeling & Simulation - DEVS Integrative M&S Symposium*, DEVS 13, pages 34:1–34:8, San Diego, CA, USA, 2013. Society for Computer Simulation International.
- [225] The Food Foundation. Food environment policy index: Final report. <http://foodfoundation.org.uk/wp-content/uploads/2016/11/ENGLAND-Food-EPI-Report-FINAL1.pdf>, 2016. [Online, accessed March 13, 2017].
- [226] The Food Foundation. Force-fed: Does the food system constrict healthy choices for typical british families? <http://foodfoundation.org.uk/wp-content/uploads/2016/>

- 07/The-Food-Foundation-64pp-A4-Landscape-Brochure-AW-V32.pdf, 2016. Accessed: March 13, 2017.
- [227] A. Tolk. The next generation of modeling & simulation: Integrating big data and deep learning. In *Proceedings of the Conference on Summer Computer Simulation*, SummerSim '15, pages 1–8, San Diego, CA, USA, 2015. Society for Computer Simulation International.
- [228] C. Tominski, P. Schulze-Wollgast, and H. Schumann. 3d information visualization for time dependent data on maps. In *Ninth International Conference on Information Visualisation (IV'05)*, pages 175–181, 2005.
- [229] G. A. Trunfio. Predicting wildfire spreading through a hexagonal cellular automata model. In P. M. A. Sloot, B. Chopard, and A. G. Hoekstra, editors, *6th International Conference on Cellular Automata for Research and Industry (ACRI 2004)*, pages 385–394. Springer Berlin Heidelberg, 2004.
- [230] G. E. Tucker, D. E. Hobley, E. Hutton, N. M. Gasparini, E. Istanbuluoglu, J. M. Adams, and S. S. Nudurupati. Celllab-cts 2015: continuous-time stochastic cellular automaton modeling using landlab. *Geoscientific Model Development*, 9:823–839, 2016.
- [231] A. M. Uhrmacher and D. Weyns. *Multi-Agent systems: Simulation and applications*. CRC press, 2009.
- [232] U. o. C. L. UK Healthy Cities Network. Uk healthy cities network, 2011. [Online; accessed 22-September-2016].
- [233] A. Unger and H. Schumann. Visual support for the understanding of simulation processes. In *2009 IEEE Pacific Visualization Symposium*, pages 57–64, 2009.
- [234] S. Van Segbroeck, F. C. Santos, and J. M. Pacheco. Adaptive contact networks change effective disease infectiousness and dynamics. *PLoS computational biology*, 6(8):e1000895, 2010.

- [235] I. Vandenbroeck, J. Goossens, and M. Clemens. Foresight tackling obesities: Future choices—building the obesity system map. *Government Office for Science, UK Governments Foresight Programme*, 2007.
- [236] A. Venerandi, M. Zanella, O. Romice, and S. Porta. The form of gentrification.
- [237] J. A. Vennix. *Group model building*, chapter Individual and organizational problem construction, pages 9–41. John Wiley and Sons, 1996.
- [238] T. Verigin, P. J. Giabbanelli, and P. I. Davidsen. Supporting a systems approach to healthy weight interventions in british columbia by modeling weight and well-being. In *Proceedings of the 49th Annual Simulation Symposium*, ANSS '16, pages 9:1–9:10, 2016.
- [239] A. Voinov and H. H. Shugart. integronsters, integral and integrated modeling. *Environmental Modelling & Software*, 39:149 – 158, 2013.
- [240] M. Waldner, M. Le Muzic, M. Bernhard, W. Purgathofer, and I. Viola. Attractive flicker-guiding attention in dynamic narrative visualizations. *IEEE transactions on visualization and computer graphics*, 20(12):2456–2465, 2014.
- [241] F. Wang, A. Antipova, and S. Porta. Street centrality and land use intensity in baton rouge, louisiana. *Journal of Transport Geography*, 19(2):285–293, 2011.
- [242] M. O. Ward. Multivariate data glyphs: Principles and practice. In *Handbook of Data Visualization*, pages 179–198. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [243] S. H. White, A. M. del Rey, and G. R. Sanchez. Modeling epidemics using cellular automata. *Applied mathematics and computation*, 186(1):193–202, 2007.
- [244] J. Williams, P. Scarborough, A. Matthews, G. Cowburn, C. Foster, N. Roberts, and M. Rayner. A systematic review of the influence of the retail food environment around schools on obesity-related outcomes. *Obesity reviews*, 15(5):359–374, 2014.

- [245] J. Williams, P. Scarborough, N. Townsend, A. Matthews, T. Burgoine, L. Mumtaz, and M. Rayner. Associations between food outlets around schools and bmi among primary students in england: A cross-classified multi-level analysis. *PloS one*, 10(7):e0132930, 2015.
- [246] T. Willumsen. Hot food takeaway supplementary planning document. <http://www.gateshead.gov.uk/DocumentLibrary/Building/PlanningPolicy/SPD/Hot-Food-Takeaway-SPD-2015.pdf>, 2016. Accessed: March 13, 2017.
- [247] Y. Xiao, S. Orford, and C. J. Webster. Urban configuration, accessibility, and property prices: a case study of cardiff, wales. *Environment and Planning B: Planning and Design*, 43(1):108–129, 2016.
- [248] I. Yamada and J. Thill. Empirical comparisons of planar and network k-functions in traffic accident analysis. In *Proceedings of the 82nd Transportation Research Board Annual Meeting*. Washington DC, 2003.
- [249] D. Zhang, P. J. Giabbanelli, O. Arah, and F. Zimmerman. Impact of different policies on unhealthy dietary behaviors in an urban adult population: an agent-based simulation model. *American Journal of Public Health*, 104(7):1217–1222, 2014.