UTS UNIVERSITY OF TECHNOLOGY SYDNEY

A Visual Comparative Study on Multivariate Data Analysis

by Yu Dong

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Computer Science

under the supervision of Dr. Christy Jie Liang and Prof. Yi Chen

University of Technology Sydney
Faculty of Engineering and Information Technology

March 2023

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yu Dong, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy of Computer Science, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 09/07/2023

DEDICATION

To my parents, my lover, my family, and my four passed relatives. . . .

ACKNOWLEDGMENTS

Completing this Ph.D. thesis has been a long and challenging journey, and I owe my sincere gratitude to the many people who have supported and encouraged me along the way.

First and foremost, I am deeply grateful to my supervisor group, principal supervisor Dr. Christy Jie Liang, co-supervisor A/Prof. Mao Lin Huang, and external supervisor Prof. Yi Chen, for their invaluable guidance, patience, and intellectual insights throughout my research. Their expertise and commitment to excellence have been instrumental in shaping my ideas and pushing me to new heights. In particular, Christy has been incredibly kind and supportive, helping me not only in my academic pursuits but also in adapting to life in a foreign country. I am truly grateful for the wonderful teacher-student and friend relationship that we have developed.

I would also like to express my sincere appreciation to Prof. Longbing Cao, A/Prof. Quang Vinh Nguyen, Prof. Jie Hua, Prof. Xiaoru Yuan, Prof. Daniel Catchpoole, Prof. Ian Oppermann, Alex Fauth, and all domain experts for each publication, for their invaluable feedback, suggestions, and assistance, which have helped me to refine and improve my research at every stage.

I am also deeply grateful to have had the opportunity to teach the course Data Visualisation and Visual Analytics twice, which has been a constant source of inspiration and motivation. I have learned so much from the course and from each of my students, and it has been a pleasure to share my knowledge and experience with them.

To my parents, lover, and family, I cannot express enough gratitude for their unwavering support, love, and encouragement throughout this journey. They have celebrated my achievements, lifted me up when I needed it most, and provided a constant source of motivation and inspiration.

Finally, I would like to acknowledge the financial support I received from the UTS-CSC and UTS International Research Scholarship (UTS IRS) scholarships, as well as multiple top-up scholarships conducted by my principal supervisor, Christy. Without this support, the completion of this thesis would not have been possible.

Thank you all for your support, encouragement, and kindness. I am deeply grateful and will carry these experiences with me as I continue on to the next stage of my career.

List of Publications

Related to the Thesis :

1. **Yu Dong**, Alex Fauth, Maolin Huang, Yi Chen, and Christy Jie Liang. "Pansytree: Merging multiple hierarchies." *In 2020 IEEE Pacific visualization symposium (PacificVis)*, pp. 131-135. IEEE, 2020.

2. **Yu Dong**, Christy Jie Liang, Yi Chen, and Jie Hua. "A Visual Modeling Method for Spatiotemporal and Multidimensional Features in Epidemiological Analysis: Applied COVID-19 Aggregated Datasets." *Computational Visual Media*. Accept.

3. **Yu Dong**, Ian Oppermann, Christy Jie Liang, Xiaoru Yuan, and Nguyen Quang Vinh. "User-centered visual explorer of in-process comparison in spatiotemporal space." *Journal of Visualization* (2022): 1-19.

4. **Yu Dong**, Christy Jie Liang, LongBing Cao, and Daniel Catchpoole. "ClinicLens: Visual Analytics for Exploring and Optimizing the Testing Capacity of Clinics given Uncertainty." *arXiv preprint*. arXiv:2303.13558 (2023)

Others :

5. Caixia Wu, Yi Chen, **Yu Dong**, Fangfang Zhou, Ying Zhao, and Christy Jie Liang. "VizOPTICS: Getting insights into OPTICS via interactive visual analysis." *Computers and Electrical Engineering* 107 (2023): 108624

6. Yi Chen, Yandi Guo, Qiuxu Fan, Qinghui Zhang, and **Yu Dong**. "Health-Aware Food Recommendation Based on Knowledge Graph and Multi-Task Learning." *Foods* 12, no. 10 (2023): 2079.

7. Yi Chen, Xiaoran Sun, Wenqiang Wei, **Yu Dong**, and Christy Jie Liang. "A Prediction and Visual Analysis Method for Graduation Destination of Undergraduates Based on LambdaMART Model." *International Journal of Information and Communication Technology Education (IJICTE)* 18, no. 2 (2022): 1-19.

8. Guozheng Li, Yu Zhang, **Yu Dong**, Christy Jie Liang, Jinson Zhang, Jinsong Wang, Michael J. McGuffin, and Xiaoru Yuan. "Barcodetree: Scalable comparison of multiple hierarchies." *IEEE transactions on visualization and computer graphics* 26, no. 1 (2019): 1022-1032

Abstract

Comparative analysis plays a crucial role in real-world data analysis, especially when dealing with multivariate data. Efficient comparison of hierarchies and node attributes, spatiotemporal data, and even data uncertainties in multivariate data still remains a challenge and gap in certain application contexts. This thesis provides a comprehensive visual comparative analysis of multivariate data, consisting of four individual research approaches that explore innovative visual solutions to improve the understanding and comparison of multivariate data.

The first approach focuses on balancing the comparison of hierarchies and node attributes. It introduces a novel visualization technique, *PansyTree*, which utilizes a tree metaphor and node merging to depict merged nodes in the structure. This enables the merging of three datasets into a single tree, facilitating easier exploration and comparison of structures, nodes, and node attributes.

The second approach, *+msRNAer*, presents a portrait-based visual modeling method for time series and multidimensional feature comparison in epidemiology. It presents time series and multidimensional features in a reduced-dimensional space, highlighting similarities and differences among designed portraits. This approach has been tested on COVID-19-related datasets and has proven effective in identifying location-based patterns and relationships between COVID-19 cases and risk factors in census data.

The third approach is the User-centered Visual Explorer (*UcVE*), which offers customizable views for exploring and comparing spatiotemporal and multidimensional features. *UcVE* reduces the cognitive load of users by allowing them to visualize, save, and track their exploration results. With an easy-to-use interface, *UcVE* enables users to switch between views and explore data at various levels of detail, making the analysis of complex spatiotemporal data more accessible and intuitive.

The fourth approach, *ClinicLens*, is an interactive visual analytics system aimed at exploring, comparing, and optimizing the testing capacities of healthcare clinics in the presence of multivariate uncertainties. *ClinicLens* leverages a combination of collaborative visual views and AI algorithms to support domain experts in making informed decisions and adjustments regarding testing capacities and COVID-19 situations.

In conclusion, these four approaches offer diverse perspectives for improving the understanding and comparison of complex multivariate data and have demonstrated their usefulness and effectiveness in real-world applications.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Background

The advent of big data has brought about significant increases in data volume and diversity of multivariate data types, including hierarchical structures, spatiotemporal features, and multidimensional structured and unstructured data features, presenting numerous challenges for data analysis. As such, it is essential to effectively balance the representation of various types of features in abstract datasets and to uncover and understand connections between these features through visual analytics. Despite this, traditional visualization and visual analytical methods are insufficient for dealing with complex implicit features and applications, posing persistent challenges in specific analysis tasks. Effective visual analytics can greatly aid in the interpretation and analysis of multivariate data, and there is a growing need for visual analysis systems that support flexible exploration and provide explicit comparisons of implicit features during the analysis process [191].

Returning to applied scenarios, the growing emphasis on the need for informatics and analytics in public health over the past 30 years has led to an increasing amount of investment cost in information systems [38]. Visual analytics in bioinformatics raised significant roles in big data analysis tasks in public health, especially in epidemiological analysis as connections tighten up with people. However, many research challenges and gaps still need to be filled, although many new tools and algorithms have already been developed to aid experts in analyzing and visualizing the complex data used in epidemiological analysis [43]. As Lauren et al. [38] highlighted in a survey, complex epidemiological exploration and comparison require novel visualization tools, and most of the existing visualizations applied in analysis tasks suffer from limited adoption. More and more standard visualizations in epidemiology evolve into visual analytics that aims to combine multivariate data such as analytical tasks of considerable complexity, dynamism, and uncertainty rather than simple features. [15, 17].

Taking the current global pandemic of COVID-19 as an example, Australian federal, state, and local community public health officials synthesize highly disparate data to facilitate timely communication with the public and inform decisions regarding policies to protect public health. Since the first outbreak, visualization has been crucial in addressing risks to the public's health. The phenomenon and its global effects have been captured in an astounding number of visual representations [6, 215]. Given that both the volume and complexity of data on infectious diseases are increasing, comparative analysis of epidemiology data combining multivariate data, such as hierarchical structures [105], spatiotemporal features [164], or other complex distinct features[209], continues to be a fundamental challenge.

1.2 Problem Statement and Gaps

The field of data visualization continues to grapple with the challenge of effectively presenting multiple data types for comparison, despite the proliferation of techniques developed for multivariate data analysis. Research problems and gaps in the visual comparison of multivariate data are mainly aggregated into two aspects: multivariate-oriented and scenario-oriented. In this subsection, we first introduce the general research problem in multivariate data, followed by specific data types of hierarchical and spatiotemporal features as examples. Finally, we emphasize introducing the problem and gaps from epidemiology to COVID-19-related scenarios as target applications.

1.2.1 Problems and Gaps in Multivariate Data Visualization

In recent years, the amount of data being generated and collected has increased exponentially. With this increase in data volume and varied types, there is a need for effective ways to analyze and interpret it. Multivariate data visualizations are powerful tools for exploring and understanding complex datasets [57, 149]. However, there are still gaps in the research on this topic that need to be addressed, particularly in comparison tasks among different data types. Gintautas et al. [63] provided academic chapters entitled

"Multidimensional Data Visualization - Methods and Applications" that categorized conventional and traditional visualization methods for multivariate data. However, they also mentioned the posed difficulties and challenges that researchers faced consistently. These difficulties and challenges have corresponded in another prior survey paper [39] which illustrated them from four perspectives:

1. **Effecitve Mapping.** Finding a suitable mapping method or layout for selected dimensions from multivariate data is not a simple task, because High-dimensional multivariate data can be difficult to map into a 2D visual form. Thus, graphical attributes should be designed carefully to make them easy to understand. The visual representation should allow for integrated analysis of different attributes while allowing users to judge each dimension separately and independently.

2. **Selected Dimensionality.** Multivariate data is often large in size and high in dimensionality, resulting in a dense structure. As a result, it can be challenging to present such data in a single visual display, making it difficult for users to explore the data space intuitively and interactively, as well as distinguish individual dimensions. Additionally, the ordering of dimensions may have a significant impact on the expressiveness of visualization [101]. Different arrangements can lead to different conclusions, but currently there is no established ordering principle.

3. **Design Tradeoffs.** Visualization is an effective way to provide users with a qualitative overview of large and complex datasets, allowing them to identify structure, patterns, trends, and relationships more efficiently [76]. However, due to the high dimensionality of multivariate data, we often need to sacrifice the ability to show detailed

information for each attribute [213] as we have fewer graphic attributes for encoding, for instance, balancing design space and data scope displayed. Therefore, when visualizing multivariate data, there is always a tradeoff between the amount of information, simplicity, and accuracy.

4. **Effectiveness Assessment.** The evaluation of visualization and visual analytic methods has been a subject of debate due to their subjective nature. Although some visual methods have been validated through quantitative analysis, such as time complexity and comparisons with analogous techniques, the majority of methods still rely on user studies and case studies for evaluation. Consequently, objectively assessing the effectiveness of an information visualization technique poses the fourth significant challenge.

In sum, to explore the research problems and gaps, we set one key research question thoroughly in this thesis:

How to dynamically and effectively analyze complex multivariate data?

We further expand to list this key question-based with the following detailed subquestions with their main allocated discussion parts in this thesis:

1. What types of conventional and traditional data are utilized in multivariate data analysis? (Chapter 1 Introduction, Section 1.1 Background)

2. What are the current limitations and challenges in exploring and comparing multidimensional data using these conventional and traditional data types? (Chapter 1 Introduction, Section 1.2 Problem Statement and Gaps)

3. How can explorations and comparisons between different data types be effectively represented in multivariate data visualizations? (Chapter 2 Literature Review, Section

2.1 Methodology-based Related Works)

4. What visualizations or visual analytic methods can be designed and presented for these types of data in multivariate analysis for exploration and comparison purposes? (Chapter 2 Literature Review, Section 2.2 Multivariate-based Related Works)

5. How can the usefulness and effectiveness of these presented methods be convincingly evaluated? (Different Approach-based Evaluations Designed in Sections 3.4, 4.5, 4.6, 5.6, and 6.6)

1.2.2 Problems and Gaps in Visualizing Hierarchical Features

Hierarchical features are considered conventional and key features in multivariate data analysis. They represent a structured organization of data elements, typically arranged in a hierarchical or nested manner. However, hierarchical features are often accompanied by other multivariate data features that pose significant challenges due to their complex and implicit relationships. The intricate dependencies between different levels of the hierarchy can make it difficult to effectively analyze and interpret the data. To uncover the underlying patterns and relationships in multivariate data with hierarchical features, advanced methods and techniques are required, along with expertise in data analysis and domain-specific knowledge.

Despite the widespread use of tree-based visualizations, such as trees and treemaps, to depict hierarchical relationships, these methods have been criticized for their inability to effectively compare multiple hierarchical structures, each of which may possess multidimensional attributes. Alternatives, such as node-link or space-filling methods, also have limitations in effectively displaying complex hierarchical relationships within data. Consequently, there is a pressing need for innovative visualization methods designed specifically for hierarchical data comparison, with the capability to interactively present intricate relationships in a clear and concise manner. Such methods should balance both clearer hierarchies display and node multidimensional attributes to further facilitate exploratory analysis and provide a visually compelling representation of the data.

1.2.3 Problems and Gaps in Visualizing Spatiotemporal Features

Spatiotemporal features play a crucial role in understanding the relationships between data points in both spatial and temporal dimensions. They capture the spatial and temporal characteristics of the data, providing valuable insights into patterns, trends, and interactions over time and space. However, analyzing spatiotemporal data poses its own challenges due to the complexities of spatial dependencies, temporal dynamics, and the integration of both aspects. Effectively harnessing the information contained within spatiotemporal features requires specialized analytical approaches and domain knowledge.

Specifically, based on the challenges and research gaps in spatiotemporal data mining (STDM) highlighted in [80], five factors must be addressed when exploring spatiotemporal space.

1. Complex and implicit spatiotemporal object relationships.

2. The need for interdisciplinary effort and integration of various heterogeneous datasets and multiple data mining algorithms.

3. The problem of discretization of the spatiotemporal region due to scale and zoning effects on data mining results.

4. Heterogeneity and dynamicity of the data characteristics.

5. The need for further efforts in STDM for data representations, advanced modeling, visualization, and comprehensiveness.

The worldwide COVID-19 pandemic has caused severe strain on healthcare systems and had detrimental impacts on society and the economy all over the world [36]. To date, growing attention has been paid to visualization techniques, such as visual dashboards [98, 183, 206], CT diagnoses [96, 119], and genomics modeling [126, 127]. They assist domain experts in comprehending, analyzing, and modeling COVID-19 [36]. Visual analytics assists in containing and controlling this global crisis from multiple perspectives [238].

1.2.4 Problems and Gaps from Epidemiology to COVID-19-related Scenarios

Epidemiology-related. Epidemiology, the study of the distribution and determinants of health and disease in populations, has become increasingly important in the context of global health challenges such as infectious disease outbreaks and chronic conditions. Despite the advancement of analytical methods and technologies, there remain significant research gaps and challenges in the field of epidemiological analysis. Some of the key issues summarized from a systematic survey [38] include the points:

1. Incomplete and inaccurate data, which can limit the ability to accurately charac-

terize disease patterns and risk factors.

2. Complex and interrelated determinants of health and disease, including both individual-level factors (e.g., lifestyle, genetics) and population-level factors (e.g., socioeconomic status, environmental exposures).

3. The need for integrative analysis methods that can effectively account for the multilevel and dynamic nature of epidemiological data.

4. The challenge of generalizing findings from observational studies to broader populations, given the potential for selection and measurement bias or uncertainties.

5. The need for more sophisticated and flexible data visualization methods to effectively communicate the results of epidemiological analyses to stakeholders.

These research gaps highlight the ongoing need for innovative analytical methods and approaches in the field of epidemiology to support the understanding and control of health and disease in populations.

COVID-19-related. The ongoing COVID-19 pandemic has had a profound impact globally, making it a critical area of study in epidemiology. Information visualization and visual analytics have been crucial in exploring COVID-19 information, with a large number of visual dashboards implemented to support real-time and correlation analysis tasks [237]. However, despite these efforts, there are still significant challenges in effectively analyzing COVID-19 data.

Existing visual dashboards aggregate specific COVID-19-related information and provide multiple views to display non-interactive visualizations and limited data types, such as current case counts across countries [59, 157] and broad trends of cases within countries [153, 200]. Despite the success of these visual dashboards, the sheer volume of COVID-19 data and the need for in-depth research into its data types and uncertainties highlight the limitations of information visualization for complex analysis tasks [36, 166]. To better manage COVID-19 and future infectious disease outbreaks, it is necessary to combine more metadata, such as spatiotemporal and multidimensional features, to support more advanced analysis [53]. Recent efforts have been made to introduce advanced visual analytics methods [110, 228, 233] and AI4VIS [145, 159, 175] to address these limitations. However, dynamic and multi-aspect visual analyses of COVID-19 data remain insufficient. A systematic survey of existing literature [26] highlights key research gaps and problems in this area, including:

1. The need for integrative and multidisciplinary approaches that effectively consider both the medical and social-economic aspects of the pandemic.

2. The challenge of dealing with incomplete, inconsistent, and rapidly evolving data, which can impact the accuracy and reliability of visualizations.

3. The need for more sophisticated and flexible visual encodings to effectively communicate the results of COVID-19 analyses to stakeholders.

4. The challenge of effectively representing the spatial and temporal dimensions of the pandemic, given the dynamic and evolving nature of COVID-19 data.

5. The need for more interactive and user-centered visualization tools to support real-time decision-making and exploration tasks.

These research gaps highlight the ongoing need for innovative and effective information visualization and visual analytics techniques to support the understanding and control of the COVID-19 pandemic.

1.3 Research Aims and Objectives

This thesis aims to conduct a comprehensive visual comparative study on multivariate data analysis. The study will investigate visual comparison approaches for typical multivariate data, such as hierarchical structures and spatiotemporal features, in comparison to other multivariate data. Additionally, the study will identify and summarize the challenges and gaps in current analysis tasks.

To achieve the objectives, the study proposes novel and original approaches based on visual comparison of multivariate data. These approaches aim to individually address the limitations of traditional visual analytics methods in handling complex and diverse multivariate data. The proposed approaches will support the flexible exploration and comparison of implicit features in the data and will be validated on various data and applied scenarios.

To substantiate the research inquiry, the study will conduct a comprehensive literature review of the existing research on multivariate data analysis, with a particular focus on the conventional and traditional data types utilized. Additionally, the study will explore the current limitations and challenges of visualizing multidimensional data using these data types and investigate how relationships and comparisons between different data types can be effectively represented.

To address the identified research gaps, the study will propose novel visualizations or visual analytic methods that can be designed and presented for these data types in multivariate analysis for exploration and comparison purposes. The study will also evaluate the usefulness and effectiveness of these methods convincingly through the use of empirical studies and experiments.

It is essential to note that the effectiveness of general visual analytics methods largely depends on the specific application context. Thus, the objectives of this study are summarized concisely in bullet points to ensure their alignment with the intended application scenario:

1. A visualization technique needs to be presented facing the challenge of comparing more than two hierarchical structures. Multiple hierarchical structures and the multidimensional attribute values of each node should be able to have balanced comparisons using the method.

2. A visual modeling method needs to be proposed for location-based time series and multidimensional features in epidemiological analysis tasks, which should support comprehensive exploration and interactions.

3. A visual analytic method further breaks down geo-barriers in order to investigate implicit relationships among location-based units on maps, while a based visual analytic system runs for comparing multiple visualization units chosen by the fulfilled interactions in collaborative views.

4. To enable interactive exploration of the COVID-19 pandemic based on actual needs, a visual analytics approach with a focus on a dataset related to COVID-

19 is anticipated to be developed. This strategy aims to improve comprehension of multivariate data such as the connection between clinic testing capacities, confirmed cases, and COVID-19 testing. The methodology also enables comparison and improvement of clinic testing capacities.

Overall, this study aims to contribute to the field of multivariate data analysis by providing novel visualizations and visual analytic methods that can be used to effectively explore and compare different data types in multivariate analysis.

1.4 Significance and Contribution

The significance of this thesis is demonstrated through the presentation of four distinct and novel approaches. Each approach is individually applied to a range of topics and scenarios using either completely unique or partially diverse data sources. The entire visual analytic studies are progressively completed in accordance with research objectives and associated goals: The four sections that make up the pathmap of the entire visual analytical study, as shown in Figure 1.1, provide multiple comparisons of various implicit features across distinct datasets. The first three are meant to be general approaches to common application scenarios, which can be easily applied to other scenarios with slight adjusting, while the last is presented specifically for the clinics' testing capacities within the aggregated COVID-19 dataset.

In summary, the four presented approaches each provide available visualization prototypes or systems, all of which are already open-source, and they adopt unique visual designs and views. The slight differences lie in the fact that *PansyTree* focuses on visualization design, *+msRNAers* represents a visual modeling method, *UcVE* offers comprehension visual exploration, and *ClinicLens* places more emphasis on the application of COVID-19.

Specifically, we first purposed a visualization technique called *PansyTree* for comparing both hierarchical structures and multidimensional features as node attributes by merging three hierarchies; secondly, we further introduced a visual modeling method for general epidemiological analysis, namely *+msRNAers*, for assembling location-based portraits with implicit features of time series and multidimensional features; furthermore, we presented a user-centered visual explorer (*UcVE*) for progressive comparing multiple visualization units with implicit features in spatiotemporal space; finally, we implemented *ClinicLens*, a novel visual analytics system, to allow domain experts to forecast, compare, and optimize the multivariate features that may affect clinic testing capacities under uncertainties.

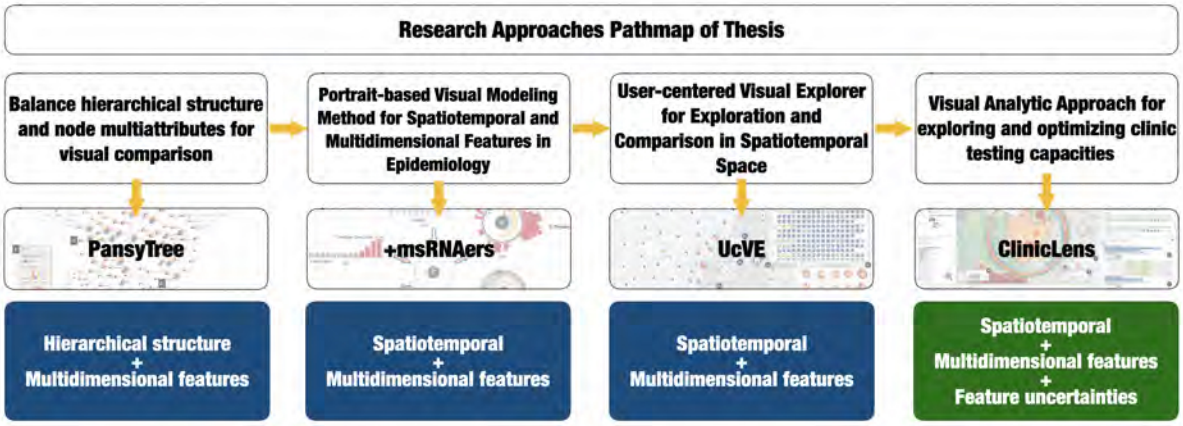

Research Approaches Pathmap of Thesis
Balance hierarchical structure and node multiattributes for visual comparison
Portrait-based Visual Modeling Method for Spatiotemporal and Multidimensional Features in Epidemiology
User-centered Visual Explorer for Exploration and Comparison in Spatiotemporal Space
Visual Analytic Approach for exploring and optimizing clinic testing capacities
PansyTree
+msRNAers
UcVE
ClinicLens
Hierarchical structure + Multidimensional features
Spatiotemporal + Multidimensional features
Spatiotemporal + Multidimensional features
Spatiotemporal + Multidimensional features + Feature uncertainties

Figure 1.1: The whole pathmap of this thesis comprises four included approaches, where *PansyTree*, *+msRNAers*, and *UcVE* can be considered as general approaches that can be easily applied to suitable data types with slight adjusting, while *ClinicLens* is specifically designed for the clinics' testing capacities within the aggregated COVID-19 dataset.

The contributions of *PansyTree* are summarized as:

1. PansyTree, a visualization technique that combines three hierarchies to represent the hierarchical structure and node attributes of three shallow and steady trees.

2. Introduce the pansy and the design of its blossoming phase for better-comparing differences among hierarchical structures and multidimensional features.

3. Visual cues and interactions in the prototype system help users explore merged hierarchical dataset patterns easily.

The contributions of *+msRNAers* are summarized as:

1. Design visual modeling method for location-based time series and multidimensional features. Inspired by viral anatomy, we developed a visual modeling method, *+msRNAers*, interacting with collaborative views for spatiotemporal and multidimensional features in the majority of epidemiological analysis tasks.

2. Integrate platform for storage spatiotemporal features in COVID-19 aggregated dataset. we create an integrated platform for all LGAs based on multiple datasets, including census data, LGA geographical data, COVID-19 cases, and events data extracted from government websites.

3. Implement a visual prototype for analyzing relationships between COVID-19 cases in spatiotemporal and census key factors in multidimensional features. This prototype aims to support the government in its investigation and comparison of implicit community factors and in identifying useful implications for the possible

patterns of the firmly established community characters against the vulnerability faced by COVID-19.

The contributions of *UcVE* are summarized as:

1. Design-oriented: A visual metaphor of unit visualization with the customizable aggregated view to expand the visual representation scalability. The UcVE maps unit visualization with the encoded abstraction of spatiotemporal information in three statuses: single, auto-clustered, and custom-merged, which allows for easy tracking and details on dynamic demand for each individual variation, making exploration easier.

2. Human-oriented: User-centered progressive exploration approach with saving and tracking in-process visualization results to ease user cognition workload. Users can record each interacted result on a map as a storage sequence for further iterative callback and exploration to deal with spatiotemporal region discretizations; ranking and target tracking can also be done interactively between different storage sequences.

3. Comparison-oriented: Comparative visualization with user-selected multiple visualization units concurrently, breaking to geo-barriers to explore implicit relationships among units. Combined with other visualization views and interactions, a comparison matrix view enables a detailed comparison of spatiotemporal attribute values among multiple visualization units in a scalable manner.

The contributions of *ClinicLens* are summarized as:

1. *ClinicLens* can assess clinic testing capacities under uncertainties by involving domain expert opinion and formulating feature modeling methods within the Back-end Engine.

2. A visual analytic system assembles the Back-end Engine with the Front-end Visualization, enabling domain experts to interactively explore the COVID-19 situation and optimize the testing capacities of clinics.

3. Three real-life case studies with expert interviews cross-evaluate the usefulness and effectiveness of *ClinicLens*.

1.5 Overview of thesis structure

The remainder of this study is organized as follows: The literature review is listed in Chapter 2. We begin with *PansyTree* in Chapter 3. In Chapter 4, we elaborate on the visual design for *+msRNAers*. Chapter 5 describes the workflow of *UcVE*. Chapter 6 is structured for a detailed approach to *ClinicLens*. We summarize the whole visual analytical study of the thesis in Chapter 7. Chapter 8 includes all four related approaches' codes published on GitHub.

CHAPTER 2

LITERATURE REVIEW

2.1 Methodology-based Related Works

This section provides methodology-based related works that contain common approaches in visual comparison and how AI techniques benefited visual analytics for further decision-making.

2.1.1 Visual Comparison

Tasks involving data analysis and visualization frequently involve comparison [70]. Finding trends in a collection of social networks [61, 231], comparing two CT scans [102], searching for similarities between several temporal trends [147], or searching for patterns in a collection of genetic sequences [163] are some examples of what it might entail in specific comparing tasks. Comparison analysis is frequently not limited to one data type but calls for an understanding of the implicit relations between multivariate data, regardless of the data type or domain [25]. Since they involve both the problems of the multiple data features and their potential relationships, such comparisons are frequently challenging.

A broad survey [70] offers a conceptual framework to help with developing responses to scenarios involving visual comparison. The framework comprises a set of four considerations that aid in understanding comparison tasks, their difficulties, and potential solutions. We can further identify the most significant considerations in visual comparison are comparative elements and layout designs. As summarized, they conclude elements may be affected by: 1. the number of items being compared; 2. the size or complexity of the individual items; 3. the size or complexity of the relationships. Thus, how to balance the number of items, the multivariate data in individual items, and reveal implicit relationships among multivariate data is a long-term issue [160]. Numerous comparative visualization approaches only support comparisons between two items (e.g. File Comparison [142], peer-peer relations [135], time series comparison [11]), which fulfill comparison in both detailed single elements and other detailed information such as multidimensional attributes, trends, and relationships; others focus on the expanded number of items for comparison (e.g. fusion and concatenation side by side [42, 89], hundreds of vertical sequences [115], etc), which are emphasized the amount of compared items but lose some detailed information. As a result, balancing the numbers and the details must be taken into account based on the analysis tasks.

The visual layouts of comparison also need to be addressed. The survey [71] also contains the concerns of common layout designs used in the visual comparison which are mainly divided into three aspects: juxtapose, superpose, and explicit encoding. Contrasting with superposition, juxtaposition arranges objects in different views next to one another in order while superposition arranges objects in the same location [193]. The relationships are the main focus of explicit encoding, which frequently combines juxtaposed, superimposed, or transformed layouts. The juxtaposition fits well in the limited number of items comparison and can depict very detailed attributes but lose space in some contexts. The superposition saves space but may constrain the other information. The explicit encoding also incurs additional learning costs due to its transforming visual design but can absorb the benefits of both juxtaposition and superposition. Moreover, the chosen visual comparison layout may be determined by the multivariate data, for instance, some variation or element-based comparisons usually used juxtaposition to show the trends with details, as Mauve [47] and Sequence Surveyor [12] in the genomes visualization field; time-series data is often applied with superposition such as due to comparison in same view will easily get insights of the differences [138, 141]; the explicit encoding may remap and redesign view for further comparison, for example, improved parallel coordinates [91, 107] for comparing multivariate data among axes. Nowadays, visual comparisons are no longer limited to a single view but have evolved into a visual analytics system with multiple views that work together to respond to comparing queries. The visual design layouts are essential for comparison tasks because we must select the right layouts for various usage scenarios.

To sum up, there is a wealth of literature in the field of visualization that addresses the challenge of comparing multivariate data that has already been summarized. From the conventional use of star glyphs, [66] and parallel coordinates [94] to more advanced methods [85, 221] and finally to recent innovations such as dimension reduction techniques [65, 67] and interactive visualizations [83], even with visual analytic methods, researchers have developed a variety of research for effectively visualizing and comparing multivariate data. These methods have been applied in diverse domains, from biology and engineering to finance and the social sciences. By offering different perspectives and trade-offs, these techniques provide valuable tools for exploring and understanding complex relationships within multivariate data, enabling data analysts and domain experts to gain new insights and knowledge.

2.1.2 AI-empowered Visual Analytics Aided Decision-making

Many fields have extensively benefited from AI-empowered techniques and applications [37]. Among the available tools and techniques, advancements in AI-empowered techniques hold the potential to propel visual analytics to new heights, enabling the development of sophisticated applications for data analysis. A survey reveals that AI-empowered visual analytics facilitates the progressive exploration of intermediate results and computational processes, undergoing continuous convergence iterations towards informed decision-making [64].

Through feature engineering with diverse data types, researchers have applied machine learning predictions and reasoning to a wide range of specific analysis tasks and decision-making problems. For example, Stolper et al. [197] devised a progressive visual analytics system that presents intermediate outcomes derived from a sequence-mining algorithm applied to medical treatment events, providing clinicians with valuable insights. *PlanningVis* [198] integrates an automatic planning algorithm with interactive visual explorations to efficiently optimise daily production planning in the manufacturing industry. The tool also offers support for quick responses to unanticipated incidents. *Tac-Simur* [212], an AI-driven visual analytics platform, simulates the processes surrounding table tennis competitions, assisting coaches in establishing competition strategies. *PassVizor* [223] provides tools for in-depth analysis of passing dynamics in football games, while *CohortVA* [236] helps historians identify the social structures and mobilities of historical figures by analysing group behaviour.

To evaluate the performance of machine learning models in visual analytics, researchers typically employ measurements by comparing their indicators among models, e.g., $RMSE$, $MAPE$, and R^2, to further choose the appropriate model for their expected tasks [185]. Xu et al. [225] used multiple models in *mTSeer* to perform 3E forecasting on multivariate time series data, where 3E stands for exploration, explanation, and evaluation. On the *PromotionLens* platform [235], Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP) emerged as winners in the comparison for providing a visual exploration of strategies to promote e-Commerce commodities. In this vein, *RISeer* [41] visually conducts interregional inspections and comparisons for urban economic development using RF and XGBoost, while *LEGION* [48] enables users to compare and choose regression models that were created by feature engineering or by fine-tuning their hyperparameters.

2.2 Multivariate-based Related Works

Both multivariate data and high-dimensional data involve visualizing datasets with multiple variables, but they differ in terms of the number of dimensions represented [220]. After collecting vast surveys on multivariate data [39, 104, 139, 148] and high-dimensional data [129, 179, 182] visualization, the similarities and differences between them are summarized below.

The similarities between multivariate data and high-dimensional data visualizations:

Representation of multiple variables: Both multivariate and high-dimensional data visualizations aim to represent multiple variables simultaneously. They allow analysts to explore relationships, patterns, and trends within complex datasets.

Exploration of relationships: Both types of visualizations provide tools for exploring relationships between variables. They help users identify correlations, clusters, outliers, and other important patterns in the data.

Visual encoding techniques: Visual encodings such as color, size, shape, position, and texture can be used in both multivariate and high-dimensional data visualizations to represent different variables and their values.

The differences between multivariate data and high-dimensional data visualizations:

Dimensionality: Multivariate data visualization typically deals with datasets where the number of variables (dimensions) is relatively low (e.g., 2-5 variables). On the other hand, high-dimensional data visualization focuses on datasets with a large number of variables (dimensions), often exceeding the ability to visualize them directly.

Techniques used: Multivariate data visualization commonly employs techniques like scatter plots, parallel coordinates, heatmaps, and small multiples to represent and compare variables. High-dimensional data visualization requires specialized techniques such as dimensionality reduction methods (e.g., t-SNE, PCA) or projection methods (e.g., parallel coordinates in higher dimensions) to reduce the dimensionality and visualize the data effectively.

Data complexity: High-dimensional data visualization faces challenges due to the curse of dimensionality, as the number of variables increases, making it difficult to visualize the data comprehensively. Multivariate data visualization, on the other hand, is relatively simpler to handle due to the lower number of variables.

Interactivity and exploration: Given the complexity of high-dimensional data, interactive visualization tools are often essential to navigate and explore the data effectively. Multivariate data visualization also benefits from interactivity but to a lesser extent.

In summary, both multivariate and high-dimensional data visualizations aim to represent multiple variables, explore relationships, and uncover patterns. However, high-dimensional data visualization requires specialized techniques to address the challenges posed by the higher number of dimensions, while multivariate data visualization is more straightforward due to the lower dimensionality.

In this section, we retrieve relevant work on hierarchical structures and spatiotemporal features, two typical data types used in visualization and visual analytics. Other multivariate data combined with both of them maintain challenges for visual design that need to be addressed further.

2.2.1 Tree Visualization for Comparison

The connection approach and the enclosure approach are the two most basic streams of tree visualization. They are both effective approaches for the visualization of hierarchies, and which one we should use depends primarily on the properties of the data in a particular application domain [87].

In the last two decades, many good tree visualization methods have appeared that combine both the connection and the enclosure concepts into one so that they can take advantage of both techniques. The treevis.net [188] concludes with over 300 methods for tree visualization, such as Cone-Tree [178], Tree-Maps [99], Spaced-Optimized Tree [146], Treemapbar [90], Angular Treemaps [118] and SFMDVis [92]. Among them, some ideas are closely related to plants in nature. For instance, Botanical Visualization [106] shows the layered relationships between trees, branches, and fruits that have emerged using 3D technology; Interring [229] is designed as a flower-like metaphor to display hierarchical information within an individual icon. Moreover, researchers further improve tree visualization techniques with visual cues [171] and interaction optimizations [46] to enhance the implicit insight of users. Zheng et al. [240] investigate the extension of four hierarchical univariate concepts the sunburst chart, the icicle plot, the circular treemap, and the bubble treemap to the multivariate domain, which provides relatively comprehensive visual solutions for multivariate data in single tree visualization.

Comparison is not a single task in tree visualization[71]. Instead, it requires the user to perform a series of interactions on a group of objectives to complete a set of explorations, such as finding similarities, performing difference analysis, etc. Li et al. [117] conducted an updated survey on tree comparison. Specifically, tree comparison can be divided into the comparison of the hierarchical structures and the node attribute values. Among them, most techniques are limited to comparing two trees at a time [42, 58, 78, 108]. Although there remain visualization methods for more than two trees [115, 239], they are still considerably fewer and have limitations in the comparison of both node attribute values and hierarchical structures.

Merging tree visualization for comparison is one of the common methods in tree comparison which refers to the use of algorithms to calculate the correspondence between the nodes of the comparison tree. According to the relationships, these nodes are merged together and a tree structure including all merging nodes is constructed. Within each merged node, attributes can be compared through visual coding [117]. Therefore, the core of the merge tree visualization is to directly compare the merged nodes' attributes in detail, and the structural difference is compared by their presence or absence. Existing merge methods are mainly divided into two types. One is to merge the juxtaposed hierarchical structure to achieve the effect of aggregation as a whole, and the difference is reflected by the fluctuation of the relationship [31, 75, 78]; the other is to merge the nodes in advance through algorithms to build the effect of aggregation [187, 219, 227]. The former is closer to the conventional node-link methods, so it is easier to recognize the attributes and hierarchical information.

2.2.2 Visual Exploration and Comparison of Spatiotemporal Features

A summary of geospatial content by Alex et al.[232] indicates that there were 94 of 220 papers in recent IEEE VIS publications used geospatial data. A variety of visual analytics methods and tools have been developed to visually make sense of geospatial data[16]. Among them are the analysis of the origin-destination[120, 242], the vehicle trajectory[10, 122, 186], the vehicle flow analysis[216], and etc[50, 79, 103, 112, 144].

The spatial-temporal features of geospatial datasets complicate matters of analysis[217]. In contrast to simple geographic information path or sampling point analysis[169, 170], time-varying attributes in especially large-scale dataset[52] bring a high level of uncertainty in the dissemination and change of geographic information, and different data attributes introduce a large number of cascading relationships, making it difficult to analyze patterns underlying datasets using traditional visual methods. Time curves[21] provide a visual method to illustrate and reveal informative patterns in a range of different spatiotemporal datasets. Brehmer et al.[30] surveyed several timelines and designed a hybrid timeline representation that combines different timeline representations in a three-dimensional space.

As a result, in addition to a series of spatiotemporal visualization encodes on the map[211], the researchers also purposed the KDE[192] (kernel density estimation), which can help with the problem of uneven distribution density. Hurter et al.[93] provided a KDE-based visual clustering approach to depict clutters in complex graph drawings.

Since then, various clustering methods have been applied to spatiotemporal data to depict clustering by different parameters, which can effectively reduce visual overlapping issues[14, 125, 131]. DDLVis[113] provides a visual analytic system with an applied peak-based kernel density estimation method to produce the data distribution for the spatiotemporal data. Compass[54] is presented for analyzing the dynamic causality in urban time series. In addition to clustering algorithms and other AI algorithms[15, 97, 243] to assist analysis, spatiotemporal data analysis also necessitates human-computer interaction, such as Andrienko et al. combined multiple map screenshots with SOM method for analyzing multiple map views; Lee et al. present a visual analytic system[109] that used Short-Term Memory model to forecasting, as well as support users to inspect each traffic congestion caused in multiple views. A bus network-based visual analytic system [216] supports visually comparing each route parameter in spatiotemporal aspects.

Most visual analysis systems based on spatiotemporal data today can help people reduce overlapping and dimensionality using AI algorithms[32, 50, 79, 114], as well as record and compare the results of multiple analyses by taking screenshots or displaying parameters side by side. However, few of them can support the flexible expansion of the interaction between each storage, like tracking and comparing spatiotemporal features among different storages, which is reducing the capabilities of each storage significantly. As a consequence, the support system needs to include not only algorithms for parsing spatiotemporal data but also fulfill some exploration strategies for gathering as much other related multivariate data from each user interaction as possible for later recall and comparison.

2.3 Scenario-based Related Works

Since the majority of our approaches utilized COVID-19-related datasets as their foundation, we introduced visualization in epidemiology analysis and visual analytics in COVID-19 in this section. The section's objective is to clearly direct our subsequent research work through specific scopes by categorizing typical epidemiology visualization applications as well as the growth of visualization and visual analysis within the context of COVID-19.

2.3.1 Visualization in Epidemiology Analysis

As early as 2020, a systemic review of COVID-19 epidemiology [162] proved high-spreading speed when the epidemic broke out. Since then, an increasing number of visual representations were presented to aid COVID-19 analysis. A novel study [181] on computational modeling used visualization-centric and algorithm-assisted for epidemiological modeling that has proved visualization plays a critical role in epidemiological analysis in 2022. Another ongoing collaboration [62] between epidemiological modelers and visualization researchers summarized the concurrent challenges and solutions. They listed common visualization charts such as heat maps [165, 184] and timelines [161], and further composite graphics with small multiple views which could support epidemiological modeling. Wei et al. [214] surveyed and categorized geographic visual display techniques in epidemiology research into two categories of Traditional Cartography and Geo-visualization.

From distinct angles of data analysis, certain risk features [51, 124] may influence the analysis in epidemiology and make the analysis tasks more challenging. As Chui et al. [45] added human factors (age, gender, etc.) into the study of infectious diseases in the paper, further analysis of visualization in epidemiology is beneficial, and it improves the precision of algorithms like modeling and prediction [180]. According to our classification, there are two perspectives for visualization in epidemiological analysis depending on the type of data used: spatiotemporal-based and multi-dimensions-based.

Pandemics are geographical in nature, and constitute spatiotemporal phenomena across large ranges of scales [137]. Improved geographic visualization plays an important role in pandemic research, that offers an environment to represent multivariate data by cartographic means, based on its geographical information effectively and attractively [73, 189, 202], and it is one of the top ten keywords of IEEE VIS [95] (top conference in visualization field); also 16% of existing related visualization works adapt maps [100].

Multidimensional-based data analysis makes epidemiological analysis possible to connect to other factors. A Singapore epidemiology of eye diseases research is a population-based study where 8,697 adults of Malay, Indian, and Chinese ethnicity [44]. Steinger et al. [196] used generalized linear models to investigate the influence of key epidemiological factors on potato virus infection risk. A tool has been deployed to demonstrate the impact of social distancing strategies during the H1N1 (swine flu) outbreak by Maciejewski et al. [130]. Trajkova et al. analyze relevant Twitter data and discuss facilitating data interpretation via visualization to avoid the spread of misconceptions and confusion on social media [203]. Also, multidimensional data visualization such as Parallel coordinates is commonly employed for visualizing multidimensional geometry [82, 94, 134]. They could apply visualization research on multidimensional attributes during the pandemic, to promote the understanding of how data entries compare to each other.

2.3.2 Visual Approaches to COVID-19

Since the COVID-19 pandemic's initial outbreak, an astounding number of visual representations or models have been created to reflect the virus's global spread and the effects it has had on various nations and regions [23, 56, 123, 184, 204]. A survey [238] conducted 668 COVID-19 data visualizations to map the landscape of existing visual works. Another novel research [237] focused on investigating complex interplay based on COVID-19 datasets between design goals, tools and technologies, data information, emerging crisis contexts, and public engagement by a qualitative interview study among dashboard creators. It could be summarized as two types of data sources in common COVID-19 research: directly-linked data such as infected cases, recovery, and mortality rates [1, 218], and indirectly-linked data, which contains community information such as [174], financial impacts [9, 199], etc., which are not linked to the pandemic directly as objective factors.

In common COVID-19 visualization research, basic techniques include traditional line charts, bar charts, maps, etc. Kahn's report reveals that 38% of related works apply line/area charts, and bar charts take a 29% share [100]. We have collected 48 existing related research works to date in Australia; 10 of them deliver a similar dashboard view, and others either offer graphs or are still ongoing works; the University of Melbourne conducts an online tool that gives a 10-day forecast [205]. Seven projects import data from aspects such as financial, and local government area (LGA) details other than only the pandemic case details. Most works apply traditional bar, stacked bar, line, map, etc. visualization methods.

In our classifications, visual approaches in COVID-19 datasets are separated into three stages: information visualization, visual modeling, and visual analytics.

From aspect one, raw data collected during the pandemic are applied as inputs to generate graphs. An interactive web-based dashboard [59] to track COVID-19 in real time was first presented by CSSE at Johns Hopkins University. Followed by the WHO created an online global dashboard [156] to show COVID-19 statuses around the globe. Hannah et al. [176] built 207 country profiles with aggregated cases, testings, vaccinations and etc which allow users to explore the detailed statistics on the COVID-19 pandemic. And macro-perspective in multiple related COVID-19 datasets, which contain such infection cases, recovery, and mortality rates with COVID and connected with social factors like geographical[72], social media and journalism[110, 234], human mobility trajectory [228], and other factors[18, 68, 88]. In Australia, the State and Territory governments assist the public with recognizing current statuses by visual dashboards [1, 116, 152, 168]. However, most of these visualization methods focus on displaying data to specific populations[96, 222] or using infectious disease models[6] to analyze and predict data[194], whereas GIS-based geographic information systems can only pre-mark sampling points on the map and cannot interact, limiting the human-computer interaction of extended analysis capabilities and making joint analysis difficult for decision-makers.

From aspect two, its visual modeling imports indirectly-linked data and offers deeper insights by integrating multiple attributes. The prototype interfaces are common implementations from this perspective; they combine computational analysis techniques with interactive visualizations [8, 20, 27, 33, 77, 207, 210, 234], which also emphasize analytical reasoning concerning the pandemic data and other facts that may affect infection cases or get affected by the pandemic by interaction techniques. Carson et al. [111] presented a big data visualization and visual tool for analyzing COVID-19 epidemiological data. Besides, more and more research has included objective factors in the COVID-19 analysis. Wu et al. develop a novel Joint Classification and Segmentation (JCS) system to perform real-time and explainable chest CT diagnosis of COVID-19 infections [222]. Muto et al. import more facts about gender, age, marriage state, poverty, and drinking/smoking habits into a matrix to address Japanese citizens' behavioral changes and preparedness against the outbreak [143]. Jiang et al.[98] presented a knowledge graph modeling method to interactively explore epidemic situations. Yang et al. noticed that based on crowd movement and control measures may have an impact on the epidemic, so they proposed EpiMob[228], a visual analytic modeling method that simulates the changes in human mobility and infection status.

From aspect three, completed visual information is extracted to assist exploration of the COVID-19 pandemic. It takes into account more relevant yet not directly-linked data which includes modeling [22, 69, 86, 172], predicting [18, 40, 110, 180, 224], and other complex explorations [233, 241] with AI modeling [35] for visual analytics. Vast visual analytics methods in COVID-19 have been systematically reviewed in a survey that comprehensively guided challenges, tasks, methods, progress, gaps, and opportunities utilized in approaches to investigating pertinent COVID-19 analysis [36]. Reinert et al. conduct a framework that enables effective and efficient visual exploration through interactive, human-guided analytical environments during the pandemic [173]. Shehzad et al. proposed a decision-making environment [6] for person-to-person contact modeling in the COVID-19 pandemic which was based on their previous works [7, 130] in epidemiology. Bowe et al.'s approach indicate that the pandemic plays out differently across different scales; it is related to the global supply chain, local dynamics, neighborhood mutual aid networks, and personal geographies of mitigation and care [29]. Preim and Lawonn describe visual analytical solutions aiming to provide preventive measures. Prevention advocates behavior and policy changes likely to improve human health [167]. Another research proposes a prediction of pandemic viral attack and how far it is expanding globally by Roy et al. [180]. Guo's system discovers spatial interaction patterns, providing valuable insight into designing more effective pandemic mitigation strategies and supporting visual exploration in time‚Äêcritical situations. An approach by Christopher [84] was conveyed through visual exploration by similarity comparison and predictions. Yu et al. [233] provided a user-centered visual explorer that applied COVID-19 datasets for in-process exploring and comparing spatiotemporal features in portraited-based perspectives. In AI4VIS, Ou et al. [158] anticipated the gas consumption under government intervention during the COVID-19 pandemic by using machine-learning-based models. Yang et al. [230] applied NLP to a visual analytics system, CVAS, identifying key events since the outbreak and the impact of the pandemic on public sentiment. Afzal et al. [6]

created a visual analytics prototype that gives public health professionals the ability to simulate and model the spread of COVID-19 by providing county-level data on the populace, demographics, and hospitalizations. Xu et al. [224] presented a web-based visual analytics tool, EPIsembleVis, for conducting a comparative visual analysis on the consistency of COVID-19 ensemble predictions under model uncertainties. Inspired by these related works, we began our earlier work with subject matter experts, gathered the most pertinent data, and extracted pertinent information for exploring and comparing COVID-19-related multivariate data of various visual analysis tasks in the following chapters.

CHAPTER 3

PANSYTREE: MERGING MULTIPLE HIERARCHIES

3.1 Research Scope of Pansytree

The large volume of data generated in various fields is often stored in hierarchical structures as their natural forms, including organization structure, disciplinary classification, and taxonomy of objects. The increasing complexity of data often prevents humans from understanding and detecting the similarities and differences between multiple datasets. However, most existing visualization methods are designed for a single hierarchical dataset. Inspired by the Pansy flowers with multi-colored petals, we propose PansyTree to visualize, merge, and compare multiple hierarchical datasets. PansyTree uses the centrally rooted tree to convey the merged hierarchical structures and different colored petal sets on each pansy node to represent the respective attribute values for multiple datasets. This new visualization method, supported by interactive filtering and combining functions, allows users to explore and compare one or multiple datasets. We use three datasets of Chinese National College Entrance Examination (CNCEE) results in 2017 and present two use cases to demonstrate its effectiveness.

3.2 PansyTree

The visualization design is inspired by a flower named Pansy. Pansy is a peculiar plant derived by hybridization from several species. A pansy grows different colored petals on one flower. The overall visual design is based on the metaphor of a pansy flower and floral phases to visualize multiple combined datasets on one merged graph. The proposed *PansyTree* has the capability to visualize the respective attribute values of each node among these datasets and also convey the merging hierarchical structure. We introduce the design concept in detail for our pansy flower in the following subsections.

3.2.1 Pansy Design

Each Pansy highlighted petals' colors to distinguish different datasets. Table 3.1 shows how we propose using a visual cue on the flower center to calculate the total number of merged attributes in the current node. The three nodes corresponding to the same node name are merged into one by arranging their respective attributes clockwise, and the height of the petals represents the magnitude of the value, while petals of different colors obey the same located order. Specifically, each petal from 0-120 degrees is represented for red, 120-240 degrees for blue, and 240-360 degrees for yellow by defining the vertically

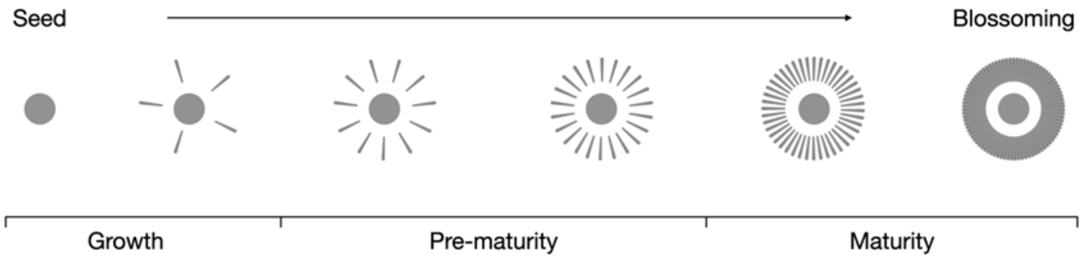


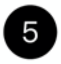
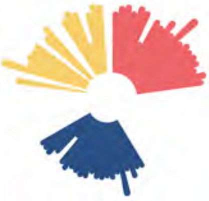
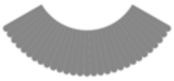
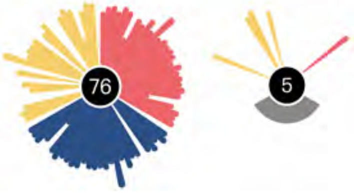
Figure 3.1: A life cycle with different phases in flower blooming.

upward as the starting point and arranging clockwise in turn by each merged node. Similarly, for each particular color, the attribute values are also arranged in clockwise order.

A new issue arises when we use the colored petal height to describe the magnitude of node attributes. That is, when there is no data, we don't know if the attribute value is 0 or if the same node does not appear at all in the merged tree structure. Furthermore, we extend the pansy design with a new definition. We plan to use a petal height equal to 0 to represent the case where the attribute value is 0. At the same time, we introduce "sepal" with gray to distinguish whether there is node data with attribute values of 0 in the merged hierarchy.

Furthermore, each plant in nature has a life cycle. While the flower blooms, its petals grow progressively. Figure 3.1 shows the floral life cycle from seed to full blossom. The life cycle can be divided into three phases: growth, pre-maturity, and maturity. Therefore, we can not only use the petals' height to encode attribute information but also take advantage of this process to describe the growth patterns.

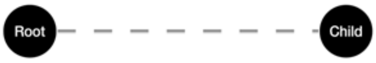
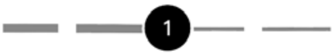
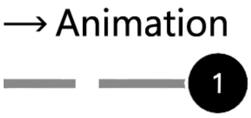
Table 3.1: The description of each Pansy design

Names	Example	Description
Flower Center		A node in the center of Pansy summarizes the petal amount in number.
Petals		Three colored petals in Pansy, represent attribute values by their heights.
Sepals		Grey sepals in the pansy represent no attribute values but occupy hierarchies.
Pansy		The pansy consists of a node element and petals or sepals that represent attribute value through color.

3.2.2 Tree Design

Considering the actual needs of the hierarchical structure are more significant than the location requirements of the same hierarchies in the merged structure, we use force layout to allocate each node's location, thereby forming a topological tree structure as a whole. In order to display the relationships between the hierarchies and the nodes in the force layout, we further define the links between the root and its children as "trunk," while others are "branches." We also set visual cues with animated links called "animated cursors" to show the closeness of relationships, which is defined as the flow speed of the animated cursors between the parent and its child nodes being in a direct ratio with the result of the intersection between this parent and each child divided by the number of the parent node attributes, which are represented in the following table 3.2.

Table 3.2: The description of tree elements design

Names	Example	Description
Trunk		The link between the root node and its child node.
Branch		Reducing the width of links level by level.
Animated cursor		The flow speed of links, called "animated cursors," between parent and child is described as: $V_{Link} \propto \frac{\lVert C_1 - C_2 \rVert}{C_1}$

The equation in the description cell of the animated cursor represents the speed of links, denoted as V_{Link}. It is proportional to the difference in the number of elements in the node sets (father node C_1 and its child node C_2), divided by the number of elements in the father node C_1.

3.2.3 Interaction Design

Although the node-link visualized in the force layout could represent elements' relationships, it is still not the most proper choice for hierarchies [208]. Because user exploration of the process can be tedious, requiring them to follow each link to discern the hierarchies, increased the timing and learning cost. Therefore, we set up some interaction solutions to deal with these possible issues. To solve the crossover, we make the nodes sticky while moving so users can drag to reduce the overlapping of nodes and links. To further help users navigate the hierarchical comparison, we designed three interaction techniques as auxiliary tools, as shown in Figure 3.2.

Structural cue highlighting. Due to the randomness and uncertainty of the force-oriented layout, we are concerned that users may still have difficulties understanding the hierarchical structure when animated links are overlapping or their speed is similar. Accordingly, we set the structural cue to highlight the attributes and relationships of this selected pansy, its ancestor, and its children.

Conditional permutations to filter hierarchies. For the merging tree with hundreds of nodes, it is essential to set different conditional permutation filters for special groups that users usually concern themselves with; otherwise, other merged nodes with other colors will distract them. To facilitate such filters, we add a sketch called Pansy for interactions, which provides several combined visualization outputs by selecting three colors.

Collapse or expand nodes to show details. When the overall hierarchy has a relatively greater depth and width, the screen size will hinder user navigation. We are concerned about users exploring their focused substructure by gradually expanding branches. In other words, we collapse some useless leaf nodes in the initial status. In general, we will collapse or expand the hierarchical structure with a certain depth in the initial state based on the average width of the merged tree to ensure the structure is balanced.

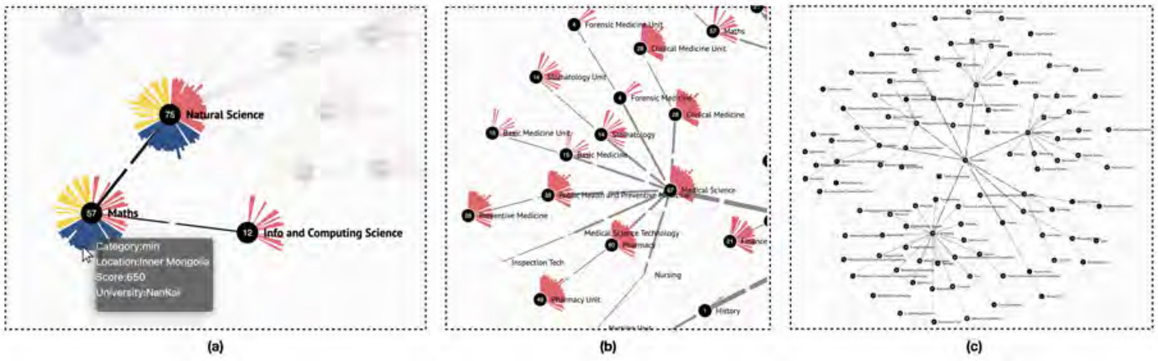


Figure 3.2: Three interaction techniques of *PansyTree* with (a) highlighting the attributes and relationships of hovering pansy node, its ancestor, and its children. (b) filtering pansy nodes by hierarchies. (c) collapse or expand nodes to show node details.

3.3 Data Description

The proposed visualization technique is designed to visualize, combine, and merge multiple hierarchies for comparison that share a similar taxonomy in the same context. It would be particularly useful for those datasets whose structures are balanced.

3.3.1 CNCEE Dataset

This paper demonstrates this technique by using three balanced hierarchical datasets from the 2017 Chinese National College Entrance Examination (CNCEE) results. These three datasets are the entrance scores for different disciplines at three universities: Zhongshan, Nankai, and Wuhan. The structure of these datasets follows a proprietary disciplinary classification stipulated by the Chinese government. In China, the entrance score for each discipline varies between different universities and provinces. The dataset's depth is fewer than 10 levels, and the nodes on each level are fewer than 20 nodes.

3.3.2 Data Reprocessing

After data cleaning, we first sort the attribute values of each discipline in three hierarchies in turn, letting these datasets expose the discipline with the same name. Afterward, we merge these disciplines according to the same names, which further establishes a union tree with their shared structures while keeping their unique portions. The result is a union tree, as well as its attribute values in each node, created from the merging of three different hierarchical datasets.

The merged CNCEE dataset is divided into nine discipline categories, which correspond to engineering, natural science, and other discipline categories. As another example, the Engineering discipline has children's subjects such as Electronics, Materials, Geology, etc. And its children's children, etc, together form a hierarchical structure; on the last level, each leaf node represents the most meticulous unit.

Each university dataset in the merged result is represented by a single color; the red portion represents Zhongshan, the blue portion Nankai, and the yellow portion Wuhan. Each of the discipline node's attributes is recorded as the entrance scores from 29 provinces. To allocate the score data, we used a linearized order based on China's geographical location, from west to east and north to south. In other words, each of the merged nodes combined from three single datasets has three times 29 provinces, which equals 87 petals. For the whole tree structure, this result dataset contains 153 nodes in total, and the deepest level is 5. For clearer descriptions, we define the actual names on each possible level in Table 3.3.

Table 3.3: The description of overview elements

Name	Description	Example
Universities	Root node	Nankai University
Disciplines	Nodes on Level 1	Engineering
Subjects	Nodes on Level 2	Electric information
Majors	Nodes on Level 3 (Optional)	Computer Science
Units	Leaf nodes on Level 4	Software Engineering

3.4 Case Studies

We implemented a prototype interface based on *PansyTree*. As shown in Figure 3.3, there is a main view with merged datasets (D) and a selected panel (E) hovering on the desktop, in which the top shows the number of nodes on each level calculated by each university. On the bottom, this panel provides a draft pansy filter for selecting any combination among the three merging datasets. We demonstrate two case studies that emerged from the exploration of the CNCEE datasets. In the first case, we compare two trunk structures from an overview aspect; the second case reveals hierarchy and attribute differences in detailed branches.

3.4.1 Case I: Branches Comparison

In case I, we compare branches in two discipline trunks in *PansyTree*. After clicking to expand all Pansy nodes, it will be easily recognized as having nine discipline trunks and their branches under each discipline node.

We notice that the maximum number of petals in each node is 87 (29 provinces times 3 universities), so we utilize 0 to 90 petals in each node to describe the whole process and divide it into 6 phases with a 15-petal interval. At the same time, we define which

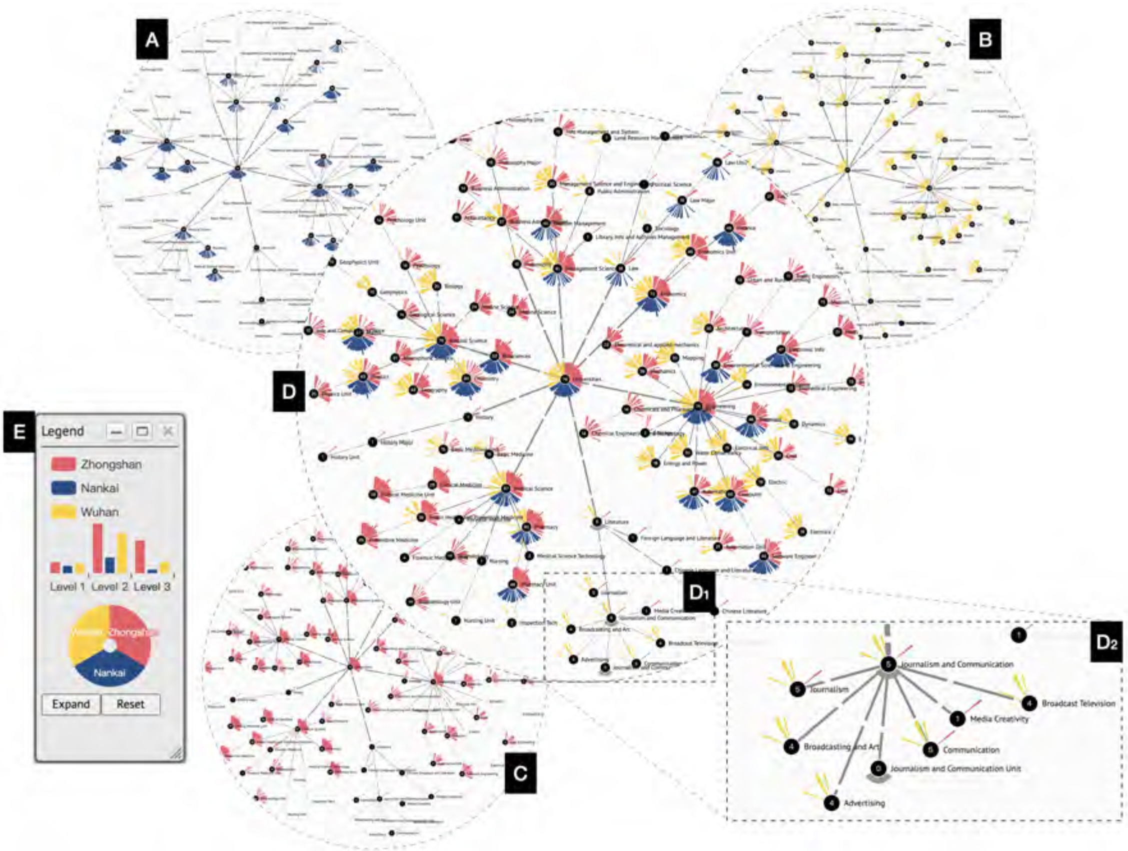


Figure 3.3: *PansyTree* is utilized by merging three datasets from the 2017 CNCEE dataset. A, B, and C visualize the entrance scores for three universities, respectively. B) Zhongshan; and C) Nankai University. The centrally rooted tree represents the hierarchical classification structures of the discipline at these three universities. The flowers on the tree represent the overview of entrance scores for each discipline, whose petals symbolize the scores for different provinces in China. D) Merging into *PansyTree*: merging the three datasets onto one tree; D1) Focus view: The baseline scores to enter universities in the discipline of journalism and communication D2) A closer examination of D1. E) A control panel for filtering or combining multiple datasets.

petals have a number between 0 and 30 as the growth phase, 30 to 60 as pre-maturity, and 60 to 90 as maturity.

To compare their branches, we chose the Engineering and Nature Science disciplines, which have relatively detailed subject classifications. Based on recording the number of petals in each node, we find that the number of petals in each node in the Nature Science discipline, which at growth phase is 11, at pre-maturity is 3, at maturity is 3, summed at 17, has 31 nodes in total, 24 at growth, 5 at pre-maturity, and 2 at maturity in the Engineering discipline.

Comparing the number of nodes and their petals in different phases, we could summarize that the discipline of engineering has more detailed subject classifications and also discern that the subjects related to the engineering discipline are more popular than those related to Nature Science because of the prosperity degree of branches.

3.4.2 Case II: Nodes Comparison

We conduct another case for exploring the node differences in specific branches. We not only focus on attribute value details but also aim to compare hierarchical structure differences among merged datasets.

Users begin by clicking the Journalism and Communication subject node, shown in Figure 3.3 (D2), and the system will expand all of the leaf nodes, which appear as the detailed units under three universities, many of which are characteristic of each university. For instance, the Media Creativity Unit is unique for Zhongshan, as are the Advertising, Broadcast Television, Broadcasting, and Art units for Wuhan.

Users can further their explorations based on the interaction results. It is not difficult to notice that there is no blue color, indicating that there is no attribute value in this branch, but it remains the sepal under the discipline Literature, subject Journalism and Communication, and Journalism and Communication unit. In other words, it clearly means there is no score from any province in Nankai in 2017, but it has the above classifications. There are also nodes named Journalism Unit and Communication Unit that connect both score data and hierarchical structures from Zhongshan and Wuhan. In order to verify our findings, we visited the official websites of three universities and found Zhongshan and Wuhan have divided journalism and communication into two units, as opposed to not being separated in Nankai. This actual case not only illustrates how *PansyTree* could help users compare hierarchical datasets but also reveals hierarchical details and attributes.

3.5 Conclusion of PansyTree

In this chapter, we introduce *PansyTree*, a visualization technique inspired by a nature metaphor that is designed to handle the merging of three shallow and steady hierarchies. The primary objective of *PansyTree* is to provide a comprehensive representation and comparison of both the structures and attribute values of over 100 nodes within these hierarchies. By leveraging the power of *PansyTree*, users can gain valuable insights and make informed decisions based on a holistic understanding of complex hierarchical data.

To demonstrate the effectiveness and practicality of *PansyTree*, we present user cases that showcase its utility with CNCEE datasets. Through these real-world scenarios, we highlight the ability of *PansyTree* to visualize and analyze intricate hierarchical structures and attribute values, thereby aiding users in making data-driven decisions.

From a technical perspective, the usage of *PansyTree* expands the possibilities of visual design and implementation, particularly by enabling the circular layout to accommodate the merging of multiple hierarchies. The design of each Pansy and petal within *PansyTree*, with encoded radius and contour, serves as the technical foundation for the approaches presented in the subsequent three chapters. These technical features provide a solid basis for further advancements and enhancements in visualizing and exploring hierarchical data using *PansyTree* as the underlying framework.

CHAPTER 4

+msRNAers: Spatiotemporal and Multidimensional Visual Modeling Method in Epidemiological Analysis

4.1 Research Scope of +msRNAers

With the assistance of experts, a visual modeling method called *+msRNAers*, was developed for spatiotemporal and multidimensional features in the majority of epidemiological analysis tasks. The *+msRNAers* were also added with collaborated views to assemble a completed visual modeling method for spatiotemporal and multidimensional features. This visual modeling adopts a portrait design inspired by viral anatomy for each community-based location for exploring and comparing the complex relationships between the number of cases in time-varying and objective risk factors that may affect transmissions, in addition to exploring epidemiological data patterns in the fundamental geographic, timeline, and multidimensional visual designs. This creates a vivid understanding of how objective risk factors are interconnected and contributes more broadly to resilient communities, particularly in light of the effects of pandemic transmission.

To validate the usability of *+msRNAers*, we collaborate with Australian Government experts to apply COVID-19 aggregated datasets with processes of:

1. Extracting evidence by each local government area (LGA) from the completed 2 years in 742 days (from January 1st, 2020, to January 11th, 2022) of COVID-19 case data compressed to 106 weeks or 53 fortnights for scalability.

2. Adding a marked timeline with intervention events extracted by NLP.

3. Connecting the most recent census community profiles to each LGA; combining case data with LGA and postal area geo-locations in New South Wales (NSW), Australia.

We then identify expert-supervised risk factors related to LGAs, including demographic indicators (e.g., higher-risk populations), social indicators (e.g., relationships), economic indicators (e.g., rental and mortgage affordability), infrastructure indicators (e.g., housing), and resident travel behavior (e.g., using public transportation).

We adjusted visual designs with multiple views and interactions to facilitate visual exploration in COVID-19-related exploration of interactive community portraits, supported by an interactive control panel with event timelines, a coordinated geographic view, and a multidimensional coordinate with a filtering function. This application not only proves the effectiveness and scalability of *+msRNAers*'s visual modeling but also aims to assist the government in real cases by investigating the pre-existing community factors and discovering practical implications for the potential patterns of the established community characters against the vulnerability facing COVID-19. We offer the applied *+msRNAers* prototype to one user study for iterative improvement and three subject-driven cases of our COVID-19 aggregated datasets, demonstrate how *+msRNAers* works across LGAs and postal areas in NSW in spatiotemporal and multidimensional features of COVID-19 datasets. These case studies provide a selected, high-level picture of community resilience in infrastructure and explore the dimensions of resilience. We evaluated the COVID-19 exploration results and conducted interviews with domain experts to collect their feedback for future research.

4.2 Design Requirements

In 2022, a survey conducted by Jason et al. [62] summarized the challenges, solutions, reflections, and recommendations of visualization for epidemiological modeling. The authors categorized the supporting visual modeling in epidemiological analysis into three stages based on different time scales:

The initial stage involves the quick application of candidate templates with visualization tools to establish problems. Data is transferred to preset combinations of views for simple comparison.

In the short-mid term stage, ongoing research provides a redesign of the visualization prototype for iteratively redefining the problem, exploring potential patterns, and providing users insight into complex tasks.

In the long-term stage, a more stable visual system is developed for widespread application to common usage scenarios, as demonstrated through multiple cases in epidemiological analysis.

This summary underscores the motivation for our proposed visual modeling method in epidemiology to fulfill multiple objectives. The visual modeling method for epidemiological analysis should not only enable quick responses to basic information trends of a pandemic, such as the infection cases in time-varying trends but also facilitate the discovery of knowledge concerning combined analysis tasks, such as identifying the factors that affect infection cases. Finally, the visual modeling method must be applied to a completed visual system and validated using real-world cases.

In consultation with experts from the epidemiology and health domains, we outlined the following user requirements: Experts required information to assess the profile of communities in terms of their resilience to virus attacks. Combined with the inspection of the number of infected cases, a few key factors might affect and help understand the impacts of different pandemic phases concurrently. They must assess the impact of the government interventions and measure the resulting pandemic situations in order to investigate both the community profile and the infection cases concurrently. They needed to investigate the link between community factors and the number of infection cases caused by outbreaks, intervention events, and responses. Usability studies conducted with early existing prototypes identified a variety of requirements. The desired features, which we distilled into three progressive categories of peer-to-peer design requirements for our approach from aspects of visual design, visual analytics, and modeling prototype application during several design iterations with subject matter expertise.

R1. Provide comparative visualization portrait of the numerical distribution of consecutively transmitted cases for each community: To perform effective epidemiological analysis, it is essential to describe consecutive infection cases in terms of timelines and locations. This involves investigating the geographic and temporal trends of the epidemic's spread, along with qualitative analyses utilizing spatiotemporal features. By providing comparative visualization portraits of these aspects for each community, we could better understand the patterns and trends of transmission.

R2. Offer visual exploration for analyzing transmission patterns with spatiotemporal and multidimensional features among each community: Epidemic outbreaks are often related to various objective factors in specific locations, such as socioeconomic or cultural factors. Therefore, it is crucial to support multidimensional feature exploration, including humanities and finance, among other fields, in addition to visualizing the spreading situation with spatiotemporal features. By combining both perspectives in interactive portraits of location-based risk factors and infection cases over time, we can provide a more comprehensive understanding of transmission patterns.

R3. Verify the effectiveness of visual modeling through a prototype system using actual epidemiological cases: To demonstrate the effectiveness of our visual modeling approach, the prototype system should offer multiple visual views with robust interactive functions, such as collaborative filtering and comparison. By applying this system to real user studies and case studies in epidemiological tasks, we can show how our approach can help researchers gain insights and make informed decisions in the field.

4.3 Visual Modeling

To address the design requirements, this section presents our design rationale through visualization, design considerations, and guidelines for the components.

4.3.1 Design Metaphor

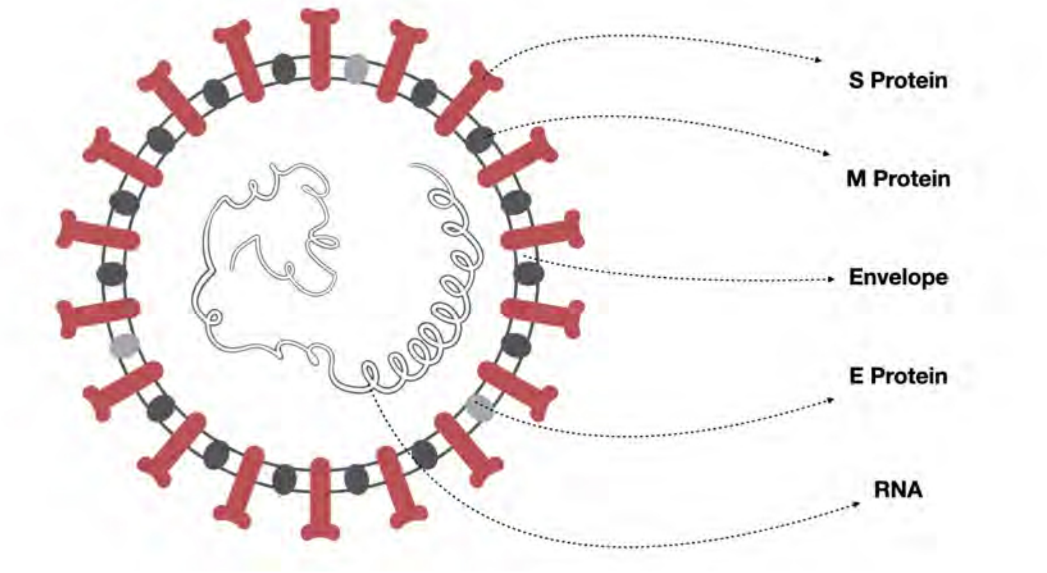


Figure 4.1: The cross-sectional simulation of SARS-related coronavirus with its main components.

Inspired by viral anatomy, our visual design primarily adopts the 2D genome structure of SARS-related coronavirus particles. Coronaviruses, named for their "crown-like" shape observed in the electron microscope, have particles packed with shells. This illustration became familiar to the public, as shown in Figure 4.1, and is widely introduced by the news and media. The coronavirus particles are organized with +*sRNA* (positive single-stranded RNA) polymers packed inside, further surrounded by outer inserted proteins. These outer proteins derive from the cells in which the virus is last assembled but are modified to contain specific viral proteins, including the Spike (S), Membrane (M), and Envelope (E) Proteins. The S Protein allows viruses to enter and infect other cells. After the virus enters the host cell, the genome was transcribed, and replication takes place involving coordinated processes of RNA synthesis. Positive Multiple Strands of RNA Encoder, abbreviated as "*+msRNAers*", is the name of our proposed visual modeling method. Inspired by this simulation and combined with our previous research [60, 233], we map different parts of the virus into multiple meanings and design *+msRNAers* to apply the aggregated datasets of its inner and outer parts.

4.3.2 Visual Design

We depict a novel visualization that leverages the biological components of the coronavirus as metaphors to represent and compare communities' epidemiological characteristics. Specifically, we define the particle symbol for each community, which we call a "Portrait", and designed the outer Proteins and inside RNAs to encode information related to cases (such as actual cases and cases per 10k/100k population) and transmission trends and cases of each community's unique risk factors, respectively. Figure 4.2 shows the detailed design for all components. Notably, we varied the RNA from positive single strand to positive multiple strands to capture multiple key risk variables related to community characteristics. This design choice provides multidimensional insight into the epidemiological patterns of each community and aims to fulfill R1.

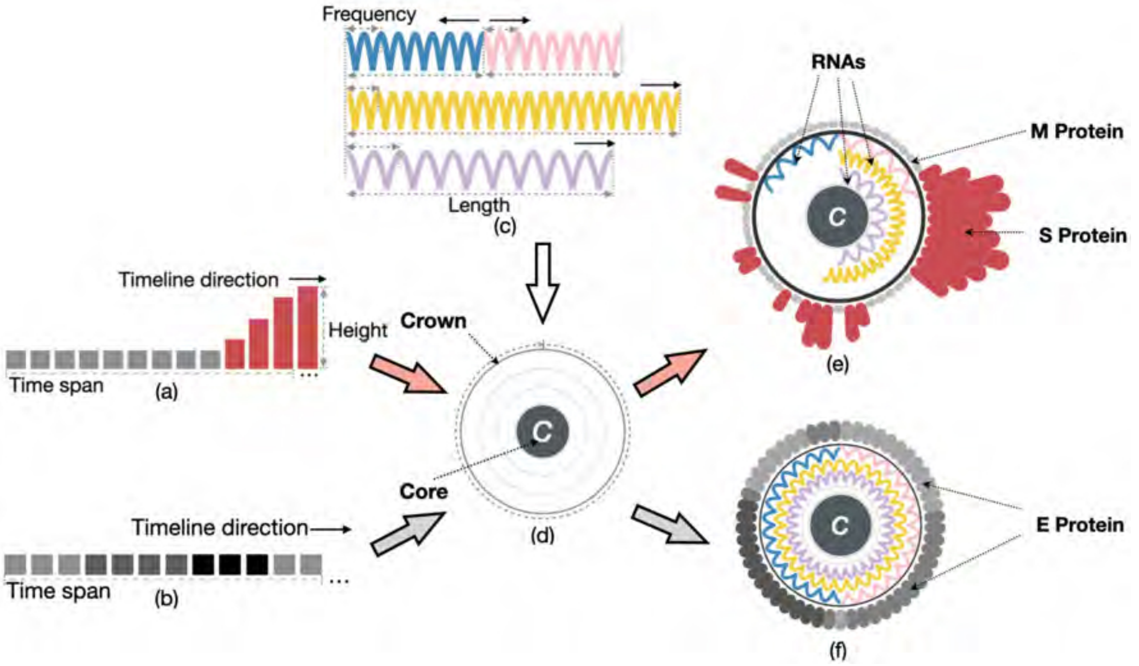


Figure 4.2: Two types of visual portraits' implementation processes, which consist of (e) Portrait with S and M Proteins, and (f) Portrait with E Proteins. Added (c) 4 strands of RNAs into both Portrait (e) and (f), the red arrow path represents the process of (a) combining bars on the timeline and encoding as S and M Proteins; the light gray arrow path represents the portrait with (b) E Protein in grayscale as different phases of the timeline.

As domain experts suggested, the number of infection cases caused by the pandemic is related to multiple potential factors among communities, e.g., population and its percentage of higher-risk groups, including the aged, lower-income, and lone-person groups. As a result, we could emphasize community portraits based on the characteristics of the vulnerable population in high-risk areas. Hence, we assemble all visual elements as the portrait for each community, as shown in Figure 4.2. This single aggregated crown-like portrait describes both the community's existing characteristics and also how the community reflects viral infection case number information. The visual portrait consists of three components: the crown, outer designs with Proteins, and inner designs with RNA. The circle of a crown represents one unit of the whole timeline. On the outside, we use multi-segments as time spans, with the S Protein representing case number information, the M Protein representing zero cases, and multiple RNAs representing the highlights of selected factors in higher risk groups inside the crown. The E Protein, unlike the S and M Proteins embedded on the crown, can represent significant intervention events rather than case numbers.

Crown Design: We define the central core with radius R_c of the Crown to label the community C, as shown in Figure 4.2(d). The circular loop surrounding the Crown easily depicts the entire timeline of any transmission taking place. The circle is evenly divided into continuous segments matching time spans for S or M Proteins, clockwise from the top point to the looped end. Further, we use E Proteins to denote the categorized intervention events within the timeline, colored in grayscale.

Outer Design: The outer-growing S Proteins are densely packed together. For each community over each time span, we encode distinct bars with rounded corners, with their heights representing the case numbers shown in Figure 4.2(a). We use the smallest-sized M Proteins to indicate zero cases in a certain time span. The S or M Protein heights of time span x are calculated as follows:

$$H_{timespan_x} = \begin{cases} h + R_c' * (a + \ln f(x)) * b & , f(x) \in N^+ \\ h & , f(x) = 0 \end{cases} \tag{4.1}$$

We define $f(x)$ is the function of recording infection cases corresponding to this time span. A base height h appears when there are 0 cases in one time span, i.e. $f(x) = 0$; a growth height is proportional to the number of non-zero infection cases in a time span x, where $f(x)$ must be positive integer, defined as $f(x) \in N^+$. Both a and b are customized parameters, and R'_c refers to the initial radius of the Crown.

Inner Design: Although viruses normally only carry one strand of RNA, it is not ideal to employ only one strand to represent several important elements. We suggest splitting the RNA strand into numerous strands to show these parameters, which conserves space and reduces unnecessary requirements. Thus, we propose to distribute the four RNA strands across three channels rather than joining them head-to-tail. Each element's maximum value is set to occupy half of a channel when multiple data elements of the same type must be represented, such as visualizing the numbers of males and females (exclusive of transgender people). These two can then be combined and allocated to a single channel. Other whole channels may be assigned to data of a single type. Figure 4.2(c) demonstrates the potential of our visual design to share one channel with two RNAs. We implement three channels of RNA that grow from the middle point of the circle. Combining two related factors in one shared channel accommodates all four RNAs placed in three oriented channels within the Crown.

To fulfill the visual design, we employ a visual metaphor of spiral genomes with four cosine wave-shaped RNAs with the same amplitude in the three channels. As shown in Figure 4.2, the length of RNA is encoded by the exact values (e) of its categories.

To denote the length of RNA L_{ij} in the arc, we define

$$L_{ij} = \left[R_c + (m - 0.5) * \frac{R'_c - R_c}{3} \right] * \left(\frac{N_{ij}}{max\{N_i\}} * \frac{2\pi}{n} + \frac{\gamma}{n} \right) \tag{4.2}$$

i.e., the arc angle θ_{ij} corresponding to the arc length L_{ij} is

$$\theta_{ij} = (\frac{N_{ij}}{max\{N_i\}} * \frac{2\pi}{n} + \frac{\gamma}{n}) * \frac{2\pi}{2\pi + \gamma} \tag{4.3}$$

Equation (3) employs a rescaling that ensures the maximum will not exceed the current RNAs' located channel lengths. N_i represents the set of data from the independent variable RNA category i, and $max\{N_i\}$ denotes the maximum value in N_i. N_{ij} refers to the value of data from an independent variable community j in an independent variable RNA category i. Parameter $m \in \{1,2,3\}$ allocates the exact location in the channel of the current RNA category i. Parameter γ is the minimum arc angle in RNAs, maintaining RNA even if the value of N_{ij} is very small. Parameter $n \in \{1,2\}$ indicates whether the maximum length of RNAs occupies a full or half channel. We define Θ_{ij} as the maximum arc angle with $\Theta_{ij} = \frac{2\pi}{n}$ when $N_{ij} = \max\{N_i\}$. Therefore, any $\theta_{ij} \in [0, \Theta_{ij}]$.

We further define the cosine wave function F_{ij} for drawing each RNA as

$$F_{ij} = \frac{R'_c - R_c}{3} * \left[|\cos(\theta_{ij} * \frac{N_{ij}}{N_j})| + (m-1) \right] \tag{4.4}$$

Equation (4) is simplified by the designed function $F_{ij} = \frac{R'_c - R_c}{3} * |\cos(\theta_{ij} * \frac{N_{ij}}{N_j})| + (m - 1) * \frac{R'_c - R_c}{3}$. This equation is assembled from two parts: one draws the cosine shape with an absolute value function, while the other is used for radial translation. The parameter in equation (4) applies $\frac{R'_c - R_c}{3}$ as amplitude after the absolute value is calculated, N_j as the sum value of data from community j, and the frequency in cosine waves $\frac{N_{ij}}{N_j}$ represents the ratio of the current value from community j in RNA category i divided by the sum of community j. The above RNA design allows for the comparison of different communities within the same RNA category from two distinct perspectives. The first involves comparing current categorical values among multiple communities based on RNA length, while the second relates to comparing the ratio of current values with the total value in the same community, as determined by RNA frequencies.

As an illustration, the RNA for selected factor $income$ from community $Sydney$ should be mapped when allocated in the second channel: $F_{income,Sydney} = \frac{R'_c - R_c}{3} + \frac{R'_c - R_c}{3} * |\cos(\theta_{income,Sydney} * \frac{N_{income,Sydney}}{N_{Sydney}})|$.

Filter Trigger Design: All the portraits are needed to motivate a filter trigger. We reset the Crown and created a sample portrait with E Proteins for filtering purposes, which can be used interactively to attach events to the timeline. On the outer circle, the grays indicate different events by timeline. In the inner Crown, three full-circled values of RNAs, as shown in Figure 4.2(f), are the indicators for selected factor groups. During exploration, we intend to interact with the visualization by interacting with all these elements in the Control Panel.

Design for All Colors: Two sets of color scales for Control Panel and Portrait View are used in our visual design. To raise awareness of the threat, we encode bright red for the S Protein design, encoding the infection case numbers, and light grey for the M Protein, indicating there were no cases this week. Inside the Crown of the portrait, inner color scales are used to representatively paint community portraits on RNAs, which include azure blue, mint pink, gold yellow, and pale purple. In the Control Panel, a pre-defined grayscale is designed for the E Protein to differentiate the types of intervention events along the timeline, initially with normal gray, silver gray, and dark gray. A darker shade of gray indicates a higher level of restriction for the events. For interaction, charcoal gray is used in both the Control Panel and portrait design for selection interaction.

4.4 Data Preparation and Prototype Application

We further developed a web-based visual prototype based on *+msRNAers*, which aims to assist in investigating the pre-existing community factors and discovering practical implications for potential patterns of established community characters against the vulnerability faced by epidemiological analysis.

Although *+msRNAers* should be suitable for most epidemiological analyses with community factors, we introduced the aggregated COVID-19 datasets in NSW as hot topics from epidemiological analysis to better present the design ideas in the previous section. We then applied *+msRNAers* to the prototype implementation by each view and introduced the interaction designs in this application.

4.4.1 Data Sources

We investigated whether there are any patterns or potential relationships between NSW COVID-19 cases and demographics, geo-information, infrastructure information, or other factors in the census fields, as clarified by the requirement of analyzing COVID-19 situations. Multiple datasets were aggregated to demonstrate the effectiveness of our approach, which includes COVID-19 case data with event timelines and selected columns of NSW census data.

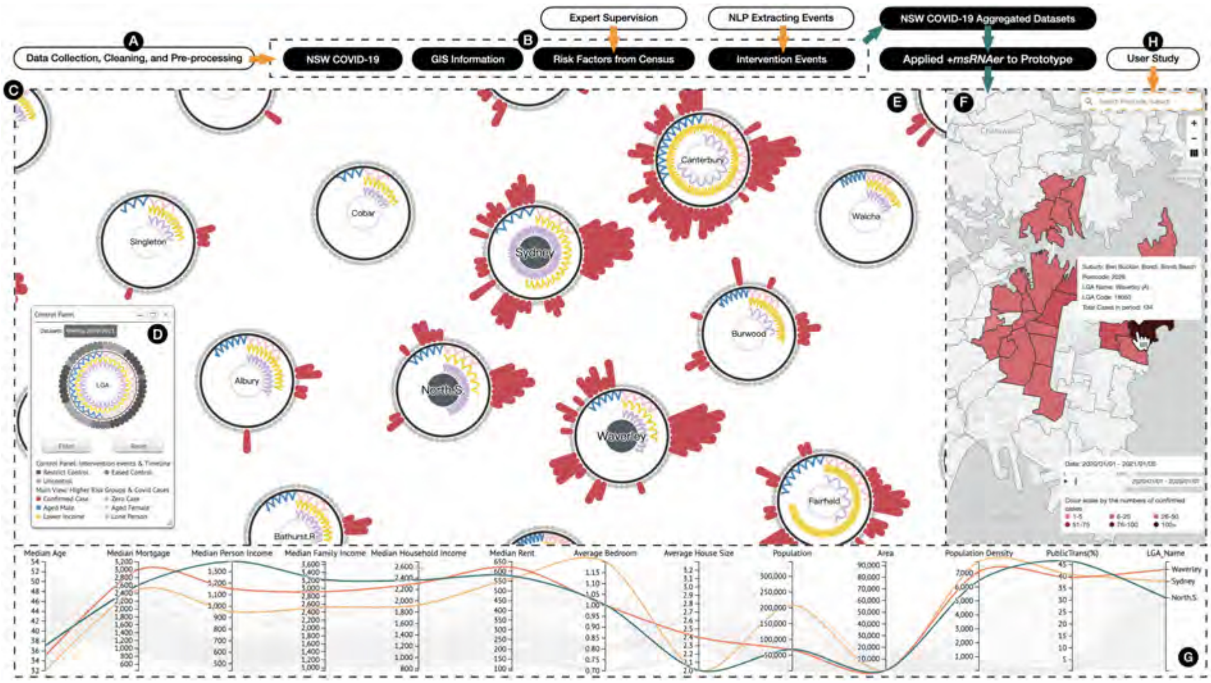


Figure 4.3: The entire workflow of the visual modeling method +*msRNAers* applied to NSW COVID-19 aggregated datasets, from raw data collection to prototype system which embodies one completed pathway. It consists of two types of arrows, where the tangerine color connects the human-participated actions and the cyan color means the data flow direction. The pathway starts from (A) data collection and processing, (B) dealing with datasets summarized COVID-19 cases, GIS information, LGA-based censuses under experts' supervision, and intervention events extracted by NLP, and assembling the aggregated datasets for application in +*msRNAers*; (C) the prototype interface with (E) Portrait View depicts four high-risk factors and COVID-19 cases corresponding to the selected timeline in (D) Control Panel and interacts with (F) and (G); (F) the GIS View with highlighted polygons of LGAs and postal areas shows their COVID-19 distribution; (G) the MDC View is a high-level overview of other risk factors from the census indicators. This +*msRNAers* prototype is finally completed by a conducted user study by adding a search function (H) in GIS View.

The COVID-19 case data in NSW was collected by the government, and the pandemic data program led by the Data Analytics Centre (DAC) provides digital information to improve the coordination of the government‚Äôs COVID-19 response. The datasets provided applicable information about infection cases based on the location of usual residence since the first infection case; they excluded 189 cases in crew members who tested positive while onboard a ship docked in NSW at the time of diagnosis; and case aspects include confirmed, tested, recovered, and deaths by their notification date, location, age-range, and likely source of infection. Some of them were no longer released due to privacy. Plus, the GIS map data and event information were extracted from media releases [150] and NSW Property Web Service [49] authorized by NSW government websites, respectively.

We decided to identify the high-risk factors that may contribute to COVID-19 infection from the census data, which involves millions of people and households and is conducted by the Australian Bureau of Statistics (ABS) every four years. The data provide a rich snapshot of the nation and inform the government, communities, and businesses. It contains essential concepts, such as populations, rents, mortgages, incomes, religions, languages, and housing. Besides, based on the definition from ABS, a lone person is classified as the only person aged 15 or over who lives in a private dwelling. Also, people in NSW earning more than 50% but less than 80% of the NSW or Sydney median income are described as earning a lower income [4]. The next wave of census data was released in stages since June 2022, but detailed information will not be released until mid-2023. Thus, the latest census data from 2016 was applied in this paper.

4.4.2 Variables Consideration

The COVID-19 case dataset delivers every case recorded by NSW Health and contains multiple attributes concerning cases by notification date and postcode, local health district, LGA, and likely source of infection. Considering the risk of leaking information that could directly identify individuals, only personal age, gender, and location of their usual residence are included in this dataset, which is assessed to measure the risk of identifying an individual and to measure the information gained if it is known that an individual is in the dataset. LGA is an official spatial unit that contains multiple postal areas that represent the whole geographical area, and there are 128 LGAs in NSW in Australia (Bayside Council was formed on September 9th, 2016 after the 2016 census, and relevant LGA data are merged from the City of Botany Bay and Rockdale City Councils). Moreover. As data journalism may affect the COVID-19 pandemic [55], relevant news articles, alerts, and ministerial media releases issued by the NSW government about COVID-19 are attached as events to combine with the notified date of infection cases.

The timeline selection was intercepted during the COVID-19 pandemic, from January 2020 to January 2022. We used Natural Language Processing (NLP) [136] to extract textual information from media data and considered marking the timeline related to keywords in interventions and social restrictions as three types of phases: uncontrolled, eased (e.g., keeping social distance, masks required), and restricted control (e.g., curfew, bubble restriction, lockdown), and the marked timeline was attached to the COVID-19 cases dataset for two data periods: the long period used 53 fortnights of COVID-19 case data as a biweekly time span grouped by 106 weeks of data from January 1st, 2020, to January 11th, 2022; the short period was used 53 fortnights of case data as a weekly time span from January 1st, 2020, to January 5th, 2021. Further imported JSON files combined a single LGA layer for the long period and an LGA layer with postal areas for the short period of COVID-19 case data.

We further consulted domain experts and selected the four top key factors from a suite of factors by their supervision that may cause infection in their communities (LGAs and postal areas) as directed: males and females aged 70+, lower income groups, and living alone groups. Experts extended four additional categories of indicators from censuses that may affect the COVID-19 pandemic: LGA-based demographic indicators, social indicators, economic indicators, infrastructure indicators, and resident travel behavior. Each category contained over 30 indicators and descriptions. Some indicators contained complex hybrid patterns that may influence transmission. For instance, apart from living alone, the indicator of household is defined as different from a family, which refers to at least one person over the age of 15 who lives in the same private dwelling. In this situation, it is necessary to deliberate on the comprehension of demographic, social, and economic indicators. We extracted the data on median age, population, and area size from LGA demographic information and further calculated population density; median rent expense, median mortgage, median personal income, median family income, and median household income from social and economic indicators; average bedrooms per person and average bedroom size per household as dwelling factors from the infrastructure indicator; and public transportation rate for traveling from resident travel behavior.

Finally, we had 96460 rows of COVID-19 case data up until January 11th, 2022, and aggregated both datasets with intervention events and spatiotemporal features as well as simplified objective factors from the NSW census data tables of size 449 MB.

Below lists a detailed statistics table, Table 4.1, summarized the COVID-19 case data and dimensions we established. This dataset were also applied for approaches in

Chapter 5 and 6.

Table 4.1: The overview of COVID-19 case data collection

Dimensions	Data Type
Record Date	Date
LGA Location	String
Postal Area Location	String
COVID-19 Case Amount	Number
COVID-19 Test Amount	Number
LGA-based Median Age	Float
LGA-based Population	Number
LGA-based Area Size	Float
LGA-based Median Rent Expense	Float
LGA-based Median Mortgage	Float
LGA-based Median Personal Income	Float
LGA-based Median Family Income	Float
LGA-based Median Household Income	Float
LGA-based Average Bedrooms per Person	Float
LGA-based Average Bedroom Size per Household	Float
LGA-based Public Transportation Rate	Float

4.4.3 Prototype Implementation

The prototype was implemented to meet requirements that offer extra visual explorations of interactive community portraits. For common epidemiological analysis tasks, we assembled the proposed *+msRNAers* prototype with collaborative visualization views, as shown in Figure 4.3, including the Control Panel (D), Portrait View (E), Geographic View (F), and Multidimensional Coordinates View (G), implemented based on D3.js [28] and Mapbox.js [133] as a base map.

For better illustration, we applied our data preparation to the NSW aggregated COVID-19 dataset. The entire workflow of the applied *+msRNAers* prototype is demonstrated in Figure 4.3. In this subsection, the *+msRNAers* prototype is introduced as being integrated into the pipeline workflow with summarized human-participated actions and data flow directions by steps, which include collecting and dealing with raw datasets, processing specific data with NLP, extracting information under expert supervision to the aggregated COVID-19 datasets, and applying *+msRNAers* to the prototype system with several functions and interactions finalized.

Control Panel and Portrait View. The Portrait View is built containing a Control Panel based on our visual modeling method, with a force-directed layout to avoid overlapping and improve readability. First, we can select the applied datasets and inspection of modes - either actual cases or cases per 10k population - in the Control Panel. Then, we can hover over or click on each time span on the sample portrait to interactively observe the corresponding filter result on each portrait and inspect all color tooltips among each view.

We can further inspect, explore, and compare infection cases and community key factors for each in the Portrait View, with aggregated case numbers by time span on the Crown and overviews of selected risk factors in the inner. We fix the Portrait View as the main view and set each portrait with the first channel with azure blue for older males and mint pink for older females, the second with gold yellow for lower-income groups, and the third inner channel with pale purple for lone people groups.

GIS View. We added extra visual polygonal layers to the Mapbox-drawn landscape to divide different LGAs and postal areas in NSW. When an LGA or postal area is selected, it will be highlighted with a charcoal gray polygonal boundary and will display detailed geographical information and infection cases. Using the GIS View, we can explore geographical information by zooming in on postal areas within the LGA and zooming out for an overview. The GIS View also provides a playable timeline window to enable the inspection of the spreading situations in selected time periods. The visual design of the GIS View utilizes a red gradient scale. The darker the red, the more infection cases are represented in each LGA.

MDC View. We utilized parallel coordinate-based visualization for multidimensional variables. The MDC View aims to show a high-level overview of other factors related to community resilience and allows users to explore the dimensions of resilience. Each LGA polyline is distinguished by coordinates, different colors, and different styles. Polyline uses different dash styles or colors according to the positive proportion of infection cases.

4.4.4 Visual Interactions

We can interact with multiple views, including Portrait, GIS, and MDC Views, triggered by Control Panel. We offer flexible interactions within and among multiple views collaboratively on this visual prototype system for each visual exploration from multiple perspectives, summarized as follows.

By combining the Control Panel and Portrait View, we implemented several interactions that enable efficient navigation through the LGA portraits and visual cues within the core for further exploration and comparison.

Filtering by Time Spans. The interactive Control Panel serves as a starter to filter LGAs in all visual views and by time spans based on intervention events and to address all color legends in the Portrait View. Users can independently inspect the COVID-19 cases on the S or M Protein among each LGA portrait based on the gray-scale mapped on the E Protein, a time period of intervention events, or a time period before and after the restriction events.

Zoom and Pan. An interactive portrait based on SVG supports zooming and panning for exploring overviews or detailed visual elements.

Drag and Reposition. The force-directed layout ensures that each portrait does not overlap. Users can drag and lock any portrait to a new location to allocate any selected LGA for comparison. In addition, we enhance interactions for further functionality by left-clicking to mark the cores and canceling their current location by double-clicking.

Highlight Visual Cues and Switch Contexts. When hovering the mouse over each type of visual cue in the Portrait View, corresponding visual cues and associated data information tooltips will be highlighted for further inspection.

Conjunction with filtering interaction. As shown in Figure 4.4, dynamic highlighting of ranking or case numbers in the core is supported based on the cases or prevalence rate with the average cases per 10k population in the previous filtered timeline added with the hovering S Protein's height. Alternatively, users can hover only on RNA to highlight its ranking among all LGAs.

The GIS and MDC Views collaborate with the Portrait View by highlighting both the selected LGAs and postal areas. Some other interactions are also supported.

Playable Timeline Window. The GIS View provides an interactive timeline window that allows users to investigate the transmission situation over different time spans. It includes several functions, such as auto-play and pause, enabling users to customize their explorations.

Boundaries Highlighted. The GIS View supports clicking to highlight boundaries and reflect LGA portraits in the Portrait View. It also supports connecting with the cores in the LGA selection.

Heatmap Highlighted. The GIS View offers a heatmap layer in LGA or postal areas by timeline filtering in the Control Panel. The colors reflect the number of COVID-19 cases in the selected time period.

Brush on MDC View. In the MDC View, the brush function is used to filter the portrait factors in multiple dimensions and reflect them to other views.

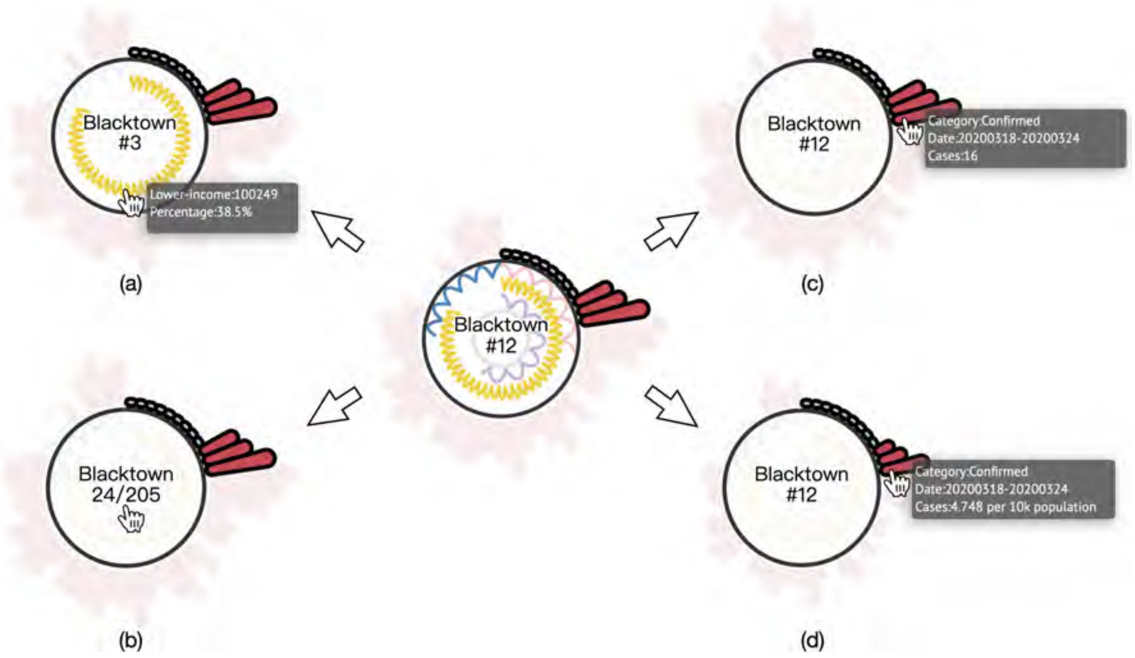


Figure 4.4: Four highlighting strategies are displayed when hovering over the (a) ranking and tooltip of the selected RNA among all LGA portraits, (b) sum of case numbers in the filtered phase with the total cases in the entire period, (c) ranking and tooltip of case numbers summed by hovering over any time span, and (d) time span context is switched to case numbers per 10k population.

4.5 User Study

To ascertain the viability and effectiveness of the *+msRNAers* prototype, we conducted a user study before conducting the case studies. In this section, we present the details of the study setup and analyze the obtained results.

4.5.1 Participants and Apparatus

In our user study, we endeavored to recruit a diverse group of participants with varied backgrounds and levels of research experience in the field of computer-related disciplines. Ultimately, we successfully recruited 16 volunteers from our campus, comprising an equal representation of 8 male and 8 female individuals. Notably, the age range of our participants was wide, ranging from 19 to 30 years, with a mean age of 24.06 years. Additionally, we verified that 5 of our participants were enrolled college students without any prior research experience, while the remaining 11 were postgraduate candidates with research experience ranging from one to eight years, resulting in a mean of 1.88 years of research experience. Despite their shared interest in visualization, all participants reported being unfamiliar with visualization methodologies.

All user studies were planned to be presented in the campus study pods. Our *+msRNAers* prototype was supported by an Apple MacBook Pro (15in, 2018) equipped with 16 GB of memory, i7 processors, and a Radeon Pro 555X Graphics Card, allowing participants to visualize and interact with the *+msRNAers* prototype clearly and effectively on a 60-inch LED external monitor with 1920 x 1080 resolution.

4.5.2 Tasks

The tasks included in our user study incorporated quantitative and qualitative analysis. The quantitative analysis provided objective numerical data, while the qualitative analysis allowed us to gain subjective insights into participants' interactions that aimed to verify the feasibility and effectiveness of *+msRNAers* prototype.

To enhance clarity in quantitative analysis, we established specific exploration tasks for each view as well as collaborative tasks among the views to test the exploration capabilities of *+msRNAers* prototype. By recording the completion time of each task, we intended to conduct analyses to determine whether *+msRNAers* prototype improved the exploration capabilities for portraits' COVID-19 case trends and risk factors and whether they had positive impacts from a visual metaphor standpoint. Additionally, we planned to gather participants' feedback on their interactions with the prototype as a whole and each view, using a set of six well-designed questions to assess their personal experience. Figure 4.5 lists specific exploration tasks and detailed questionnaires. Besides, we planned to record their feedback in transcripts via interview for qualitative analysis.

4.5.3 Procedure

After setting up the exploration tasks and questionnaires, we rehearsed a tutorial on how to use *+msRNAers* for exploration, repeated the five exploration tasks, and recorded completion times. We found that each exploration task could be completed within four minutes. Although we conducted the design of *+msRNAers*, the randomness of a force-directed layout in Portrait View did not provide significant benefits in terms of saving

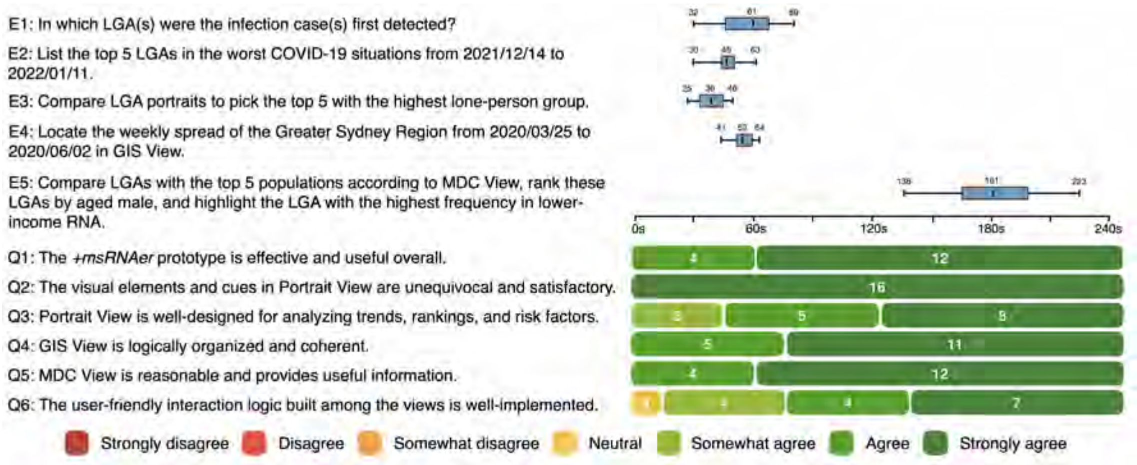


Figure 4.5: The quantitative analysis results of the exploration task list (E1–E5) and the questionnaire (Q1–Q6), where exploration tasks record the completed time in a box plot and the questionnaire counts participants' choices in stacked bars.

our completion time. In other words, every participant could finish each exploration task within four minutes after tutorials. Therefore, we decided to allocate approximately a 60-minute face-to-face session for each participant, consisting of a 10-minute tutorial, approximately 20 minutes of exploration, a fixed 10-minute preset questionnaire, and a 20-minute open-ended interview. Participants were instructed on how to apply *+msRNAers* prototype and were required to complete five specific exploration tasks, with completion times recorded. They were also asked to complete the questionnaire based on their subjective experiences during prototype usage. Additionally, their complementary feedback was also recorded and used for subsequent qualitative analysis.

4.5.4 Results

In this subsection, we discussed the quantitative and qualitative results of our user study.

Quantitative Results. The quantitative results of our user study were reflected in two aspects. The first aspect was task completion time. We observed that the fluctuations in completion time across these five tasks were within reasonable ranges. The median time to complete the first four exploration tasks varied from 36 seconds to 61 seconds, while the last, more complicated task took a median of 181 seconds to complete. This consequence could be easily explained by the fact that the completion time of tasks E2-E4 was slightly reduced as participants became more familiar with the prototype during the first exploration task. Additionally, the first four exploration tasks focused on simple tasks that could be completed within one or two views. In contrast, the last exploration task required participants to engage with all collaborative views and gradually decipher three interactive results. Consequently, this task demanded more time and attention, leading to a higher median completion time. These results were in line with our expectations, as all tasks were completed within four minutes.

According to the results of the questionnaire, the majority of participants expressed satisfaction with the *+msRNAers* prototype. Specifically, 12 out of 16 participants strongly agreed with its performance, while the remaining four were satisfied with it overall. All participants were highly impressed with the prototype's visual design in Portrait View and expressed varying degrees of satisfaction with its analysis functions, including trends, rankings, and risk factors. Two later questions further highlighted the exceptional integration of both GIS View and MDC View in the prototype. The final question evaluated the implementation of interaction logic, with one participant responding neutrally and the other 15 participants expressing varying degrees of agreement that the interaction implementation was user-friendly.

The quantitative results from two aspects indicated that participants rated the feasibility and effectiveness of the applied exploration tasks and prototype designs relatively highly. They also found the visualization and interaction designs to be intuitive and impressive.

Qualitative Results. As an extension of the quantitative results, we further employed qualitative analysis during the last 20-minute open-ended discussion. In particular, we conducted opinion research for individuals who did not provide the most satisfactory options on the questionnaire. To derive the qualitative results, we repeatedly reviewed the interview recording, summarized their critical thinking, and achieved a consensus based on their complementary feedback during the interview.

Their feedback was mostly positive with high marks, with only a few critical comments. These comments can be mainly divided into two aspects:

(1) A few participants felt that the current force-directed layout of Portrait View was a "double-edged sword". On the one hand, it provided dynamicity to avoid overlapping and allowed for easy dragging and relocating, but on the other hand, its randomness may increase the workload of recognizing specific portraits in multiple explorations.

(2) Some participants who marked Neutral or Somewhat Agree in Q6 expected to see more interaction logic among views, such as enabling the inclusion of more risk factors in Portrait View or providing solutions for swapping key risk factors with other factors in MDC for higher levels of comparison purposes.

We felt grateful for their feedback as it served as a pre-evaluation before case studies to domain experts and helped us improve this applied prototype. Regarding the comments on the applied force-directed layout, we believe that its advantages outweigh the disadvantages, but we still consider it necessary to improve. As a result, we implemented the search function in GIS View to filter any LGAs or postal areas, which also strengthened the interaction between GIS View and Portrait View to a certain extent.

Since we are proposing a visual modeling method for epidemiological analysis, any other related applications using *+msRNAers* in prototypes may vary in detail. In this situation, with the prior requirements that we consulted with domain experts, we deemed that locking the highest risk factors in each portrait is acceptable because aged, lone, lower-income groups have been proven by domain experts to be high-risk factors, while other objective factors selected from the census were considered to be indirectly related categories. Thus, we explained the reasoning to the participants and derived their understanding and acceptance.

Overall, our user study showed that the design of *+msRNAers* is creative, and the application of *+msRNAers* with the COVID-19 aggregated dataset was also proven to be feasible and effective, although there were some imperfections in certain details. According to the user study, we not only improved *+msRNAers* with search functions in GIS View but also inspired some interesting cases that were demonstrated in future subject-driven case studies.

4.6 Case Studies

Coastal areas are considered densely populated areas prone to cluster infections for transmission, in contrast to the vast land and sparsely populated areas of NSW, Australia. For other incidents, intermittent spreading cycles of virus variants, corresponding policies or restrictions, densely populated residential areas, areas where older people gather, more impoverished areas, or areas with relatively backward public infrastructure may affect the infection situation among different LGAs. In this section, we finalized our +*msRNAers* prototype in two different data aspects based on the COVID-19 aggregated dataset in NSW, which consists of a weekly case summary of each postal area within an LGA in one year and a bi-weekly case summary of LGAs in two years, to the aggregated dataset to provide three prominent cases: Overview-driven cases, Event-driven cases, and Portrait-driven cases, which are based on highlighted driving aspects to compare and explore the significant connections between COVID-19 issues and detailed factors in each LGA census, even attempting to discover the relationships and potential patterns that were really affecting the COVID-19 pandemic behind each LGA portrait. We also combined our findings with facts and news for verification and analysis.

4.6.1 Overview-driven Cases

We first set up a visual representation for the LGA portraits within the period from January 1st to January 11th, 2022. Connected to the GIS View, it showed the severity distribution of COVID-19 cases in geo-polygons. With a quick glance in Figure 4.6(a), we could distinguish whether the situations are severe or not with roughly two types of

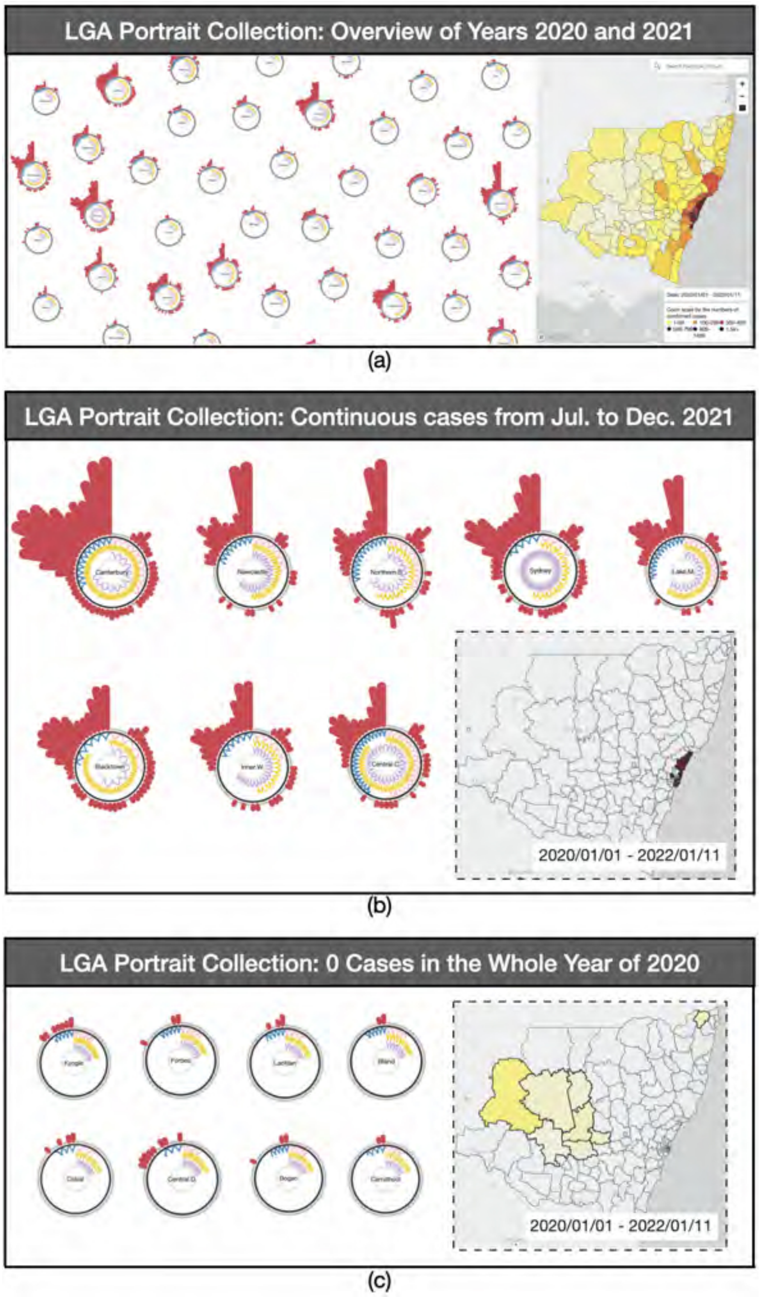

LGA Portrait Collection: Overview of Years 2020 and 2021
(a)
LGA Portrait Collection: Continuous cases from Jul. to Dec. 2021
Canterbury
Newcastle
Northern B
Sydney
Lake M
Blacktown
Inner W
Central C
2020/01/01 - 2022/01/11
(b)
LGA Portrait Collection: 0 Cases in the Whole Year of 2020
Forbes
Lachlan
Bland
Central D
2020/01/01 - 2022/01/11
(c)

Figure 4.6: Three cohorts of LGA Portrait Collections: (a) A partial overview of LGA Portraits and geo-locations for 2020 and 2021. (b) Selected LGA Portrait with cases appearing continuously from July to December 2021. (c) Selected LGA Portraits with no cases for the full year of 2020.

portrait appearances. The results are also correspondingly reflected on the map, showing the contrasting perspectives between coastal and inland areas in case numbers.

When comparing one LGA portrait to another, one can examine variations in the COVID-19 cases spreading situations among LGAs. The S Proteins of each LGA portrait showed the changing trends of COVID-19 cases by height. The only two LGA portrait types that can be distinguished based on appearance are those with COVID-19 cases in both 2020 and 2021 and others with cases only in 2021 but none in 2020.

We randomly sorted a few portraits of each type for analysis, as shown in Figure 4.6(b). The second visual result depicts the trends of COVID-19 cases within an 8-LGA portrait collection. We gained insights demonstrating that they all suffered relatively worse situations and caused two typical waves of cases from mid-year to the end of 2021. Compared to the height of S Proteins in each portrait, they reached the peak of the first wave in September, dropped back to low levels in November, and rose sharply to the highest point in two years. We corroborated the visual result with the actual situation as the NSW government released Delta concerns on July 30th, 2021[151], and the information released on November 28th, 2021, of the first Omicron variant case in NSW[2]. From GIS View, we found they are all located in coastal areas surrounding Sydney City.

The following visual output in Figure 4.6(c) showed all LGAs with portraits had COVID-19 cases during 2021, although they did not have any cases in 2020. We could further observe that there were 7 LGAs located in inland areas despite the presence of Kyogle LGA near the coastal regions of these 8 LGAs. Compared with cases appearing frequently in other developed LGAs, this consequence was most likely related to the sparse population of these LGAs.

4.6.2 Event-driven Cases

We observed that the first wave of COVID-19 spread across NSW from approximately January to October 2020. Based on the events on the timeline, we divided this period into three phases: uncontrolled, eased, and restricted control. Therefore, we further filtered each LGA portrait into weekly time spans from January 1st to October 20th, 2020. This enabled us to inspect the COVID-19 spreading situations among LGAs during this period, as tracked with the details of these events.

Uncontrolled Phase. After interacting on the Control Panel, we hovered the mouse over each E Protein to spot LGAs, which detected the first case occurring in week 4 in Randwick, Paramatta, Kur-ring-gai, and Burwood.

Concurrently, we noticed the first intervention event started in week 13. Thus, we began by selecting the phase from January 1st, 2020 to March 24th, 2020, to identify the uncontrolled phase. We further utilized the playable timeline window in GIS View to explore the spreading pattern in the Greater Sydney Region and surrounding places. As shown in Figure 4.7(a), besides the case that occurred in week 4, there were no other cases during the period from January 1st, 2020 to February 25th, 2020. However, starting from week 9 to the end of week 12, cases began spreading in the surrounding places of the Greater Sydney Region, and in a short period, the cases spread severely and finally caused outbreaks in almost all postal areas, especially in the Waverley LGA in the darker polygon color.

Drawing an overview of this uncontrolled phase, the visualization results in Figure 4.7(b) show that all LGAs were affected by the first wave of weekly increasing COVID-19 cases, with no restricted events yet. We focused on LGAs with the highest number of COVID-19 infections and interacted with them in GIS View. We located them around the main cities, including Sydney. During this period, Bondi Beach, in the Waverley LGA, had more cases than other LGAs, peaking in only two weeks. We noted that Bondi Beach was crowded with massive gatherings of people, as reported in the news, and health officials announced a crowd ban on March 21st, 2020. NSW residents were facing physical and psychological pressures due to the spread of COVID-19 and bushfires across the state. These factors indirectly contributed to overcrowding in these tourist hotspots. This situation was also reflected in the Northern Beaches LGA, which saw a rapid increase in COVID-19 cases in the first wave of the pandemic. We further examined the other two LGA portraits, Sydney and Woollahra, and concluded that the consequences were likely due to the connection to Waverley in GIS View, resulting from the movement of people in the adjacent LGA. The fifth infection case in the LGA occurred in Sutherland Shire, located in the southern region of Sydney. We could assume that the COVID-19 pandemic had spread to the outer LGAs in this phase.

Eased Phase. We adjusted the timeline for the first NSW lockdown events from March 25th to June 30th, 2020. The visual results of the top 5 LGAs in Figure 4.7(c) show that Northern Beaches had the highest number of cases, followed by Penrith and Blacktown, with Sydney LGA dropping to fourth place. Waverley finished fifth. With the special ban in place at Bondi Beach, the number of infection cases in Waverley and its adjacent LGA, Sydney, significantly decreased. We also noted that the NSW lockdown event had a positive impact on LGA cases in the Greater Sydney Region, as reflected by 0 infection cases in Northern Beaches, Sydney, and Waverley for several weeks in the second half of the selected timeline. Penrith, Blacktown, and Blackburn all had a larger number of low-income residents and are located in the Greater Western Sydney Region, which had more than half the cosine arc of the lower-income RNA and an intense frequency. As a result, we anticipated that the pandemic would spread quickly to other LGAs in the Greater Western Sydney Region.

Restrict Controlled Phase. To test our hypothesis, we further filtered the timeline from July 15th to October 20th, 2020, which was when a new lockdown and tighter restrictions were implemented by the NSW government. The visual results, as shown in Figure 4.7(d), reveal that all of the LGAs are connected in the Greater Western Sydney Region. We observed that the most strictly controlled events showed stable increases in cases, even in Cumberland and Liverpool, which had continuous growth in cases over the next few weeks. This confirms our conjecture about the pandemic situation, with a large number of infection cases in the Greater Western Sydney Region.

We also noticed potential patterns among LGA portraits, particularly in the lower-income RNA, demonstrating that all five LGAs have significant lower-income population groups and high population density. In other words, during the implementation of the tightened lockdown, most people worked from home or were self-isolated, and the number of infection cases in the CBD and tourist hotspots markedly decreased. Thus, the factors in the LGA portraits will dominate and reflect the number of infection cases. Lower-income groups or other population factors in the census may be the most crucial factors affecting the spread of the pandemic.

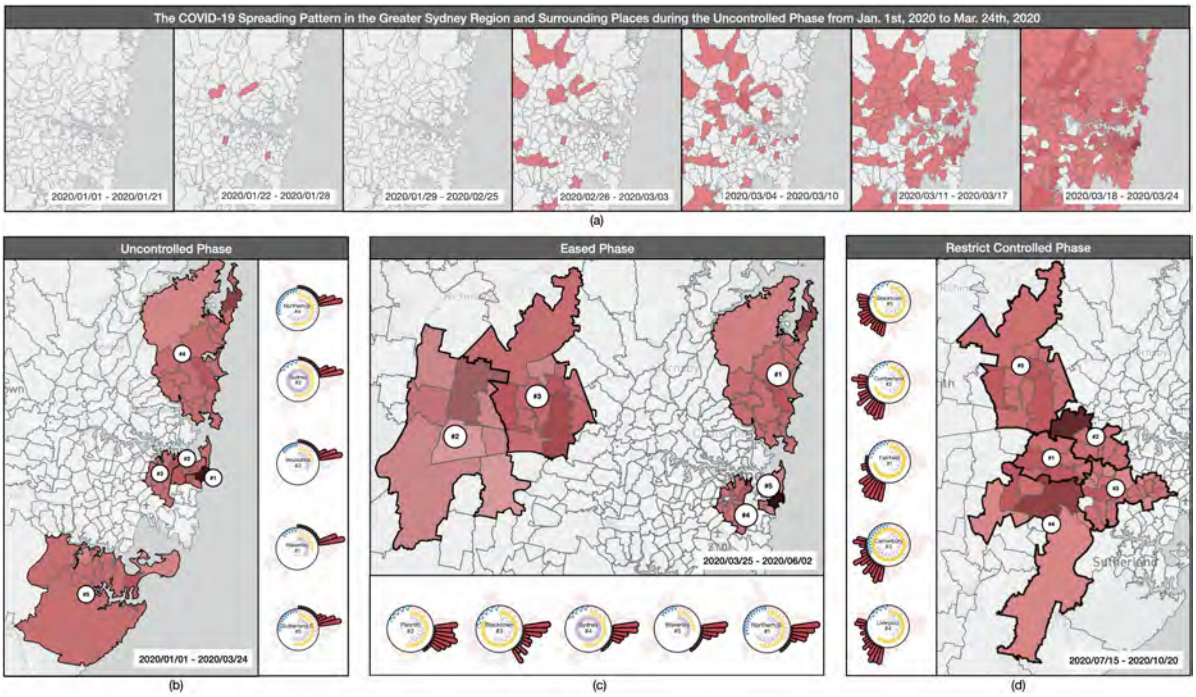


Figure 4.7: The event-driven cases encompass explorations of (a) the COVID-19 spreading pattern in the uncontrolled phase and (b)-(d) three phases of overviews. (a) provides a detailed breakdown by time spans of the COVID-19 spreading pattern in the Greater Sydney Region and surrounding places during the uncontrolled phase from 2020/01/01 to 2020/03/24. In (b)-(d), each marked ranking number is arranged according to geo-location, where (b) contains the top 5 LGA portraits in total infection cases with non-controlled events applied to the same phase as (a); (c) displays the top 5 LGA portraits in total in the eased phase from 2020/03/25 to 2020/06/02; and (d) includes the top 5 LGA portraits in total case numbers during the restricted controlled phase from 2020/07/15 to 2020/10/20.

4.6.3 Portrait-driven Cases

The highlighted interaction in RNAs facilitates the comparison of four key factors within each portrait and multiple attached attributes in MDC View. We selected 12 alternative factors in the census indicators to enable a more comprehensive comparison of the COVID-19 situations based on their portrait factors. To conduct these comparisons, we analyzed portrait-driven cases with different factor influences in collaboration among views.

Prevalence Rate Influence. We began by exploring LGA portraits for the entire time period from January 1st, 2020 to January 11th, 2022, aimed at comparing LGA portraits of coastal areas with inland areas across different case modes. We randomly selected two coastal LGA polygons (Sydney and Mid-coast) and two inland LGA polygons (Cobar and Hay) as representatives and compared them using actual case mode and prevalence rate mode, as depicted in Figure 4.8.

The consequences conveyed that all S Proteins in Sydney and Mid-Coast LGAs decreased in prevalence rate mode, while all S Proteins in inland LGAs increased in height. The opposite occurred in actual case mode. We were able to make comparisons among all LGA portraits in both modes because they had been standardized using the same calculation methods as for the S Protein heights.

These findings suggest that almost all LGAs in inland areas had insufficient 10k population, resulting in the height of S Protein in their portraits growing. We also concluded that the cases per 10k population with standardization were similar in both coastal and inland areas, which validated the strong infectiousness of COVID-19.

Case Trends and Lower-income Influences. Based on our previous findings, there may be potential relationships between COVID-19 cases and lower-income groups in each LGA. Therefore, we kept the COVID-19 outbreak timeline from the first year (January 1st, 2020 to January 5th, 2021) in the Control Panel and selected the top 10 LGAs with the worst COVID-19 pandemic situations (i.e., highest total case numbers).

From the top 10 rankings displayed on each portrait in the first year, as shown in Figure 4.9(a), we identified that areas with RNAs longer than half the range of lengths,

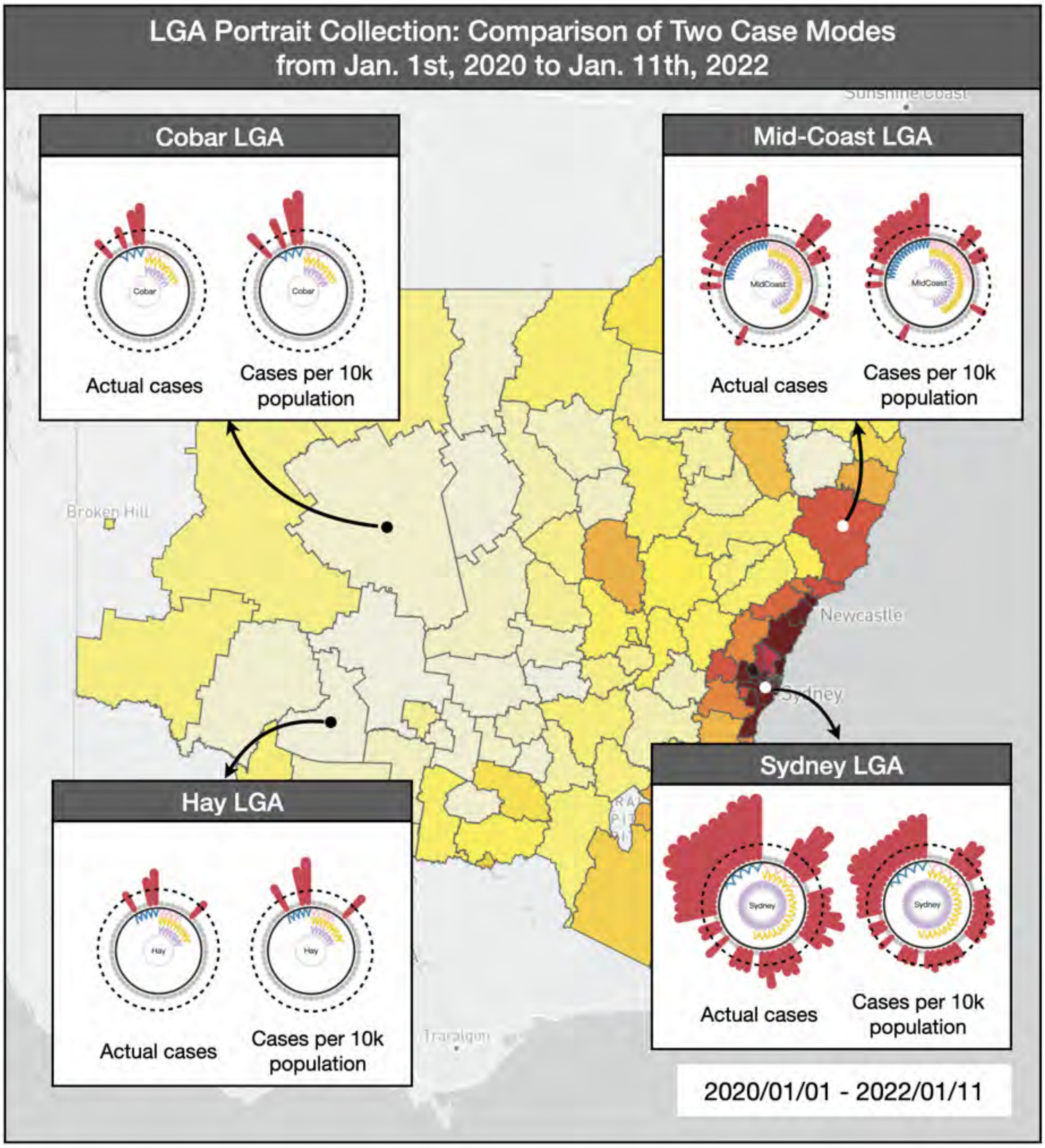


Figure 4.8: Four cohorts of LGA portraits in different modes (actual cases and cases per 10k population) were selected using LGA polygons in GIS View. Two coastal LGAs (Sydney and Mid-Coast) are marked with white pins, while two inland LGAs (Cobar and Hay) are marked with black pins.

or LGAs with the busiest frequency of RNA in one factor, have comparatively worse COVID-19 situations. For example, Fairfield ranked 8th highest in the number of cases with 143 total cases. This area contains the 4th longest lower-income RNA length but the highest tightness frequency. Additionally, the RNA lengths of aged and lone persons in this area do not approach half the length of the full path.

Given that all lower-income RNAs are highlighted in the visual overview, we assume that the lower-income RNA in the LGA portrait is the most influential factor in COVID-19. We selected 9 LGA portraits only based on the appearances of low-income RNA and categorized them into three layers whose lengths are longer than one-half, longer than one-quarter but shorter than one-half, and shorter than one-quarter. As shown in Figure 4.9(b), significant patterns are revealed, showing that the COVID-19 situation improves by layers, where the length of lower-income RNAs achieve shorter from left to right and top to bottom.

Combined Influences on LGA portraits, geo-locations, and events. We reset the time period to encompass both the eased and restricted controlled phases. Upon analyzing the Greater Sydney Region using GIS View, we observed that adjacent LGAs exhibited similar heatmap polygons. Using an iterative approach, we selected the top 5 LGA portraits with the highest case sums during this period and evaluated their case sums, as well as the total case sums throughout the year. These five LGAs are highlighted in Figure 4.9(c), whose pin-points correspond to the colors used in Figure 4.9(d). With the exception of the Waverley LGA, which had been analyzed in previous cases, the other four LGAs were adjacent. While all five LGAs had case ratios exceeding 50% during the selected period, given the duration of our chosen timeframe, which spanned more than half a year, these eased and restricted events effectively suppressed the spread of

COVID-19 in this densely populated region to a certain extent.

Household and Dwelling Influence. The LGA portraits were reverse-selected in the MDC View help to discover the consequences of COVID-19 spread in conjunction with other highlighted census attributes from residents' perspectives. In each column, the median age of each LGA is roughly inversely proportional to financial factors, including the median mortgage, personal income, family income, household income, and rent cost. In most living environments with more than two bedrooms in one household, the majority of people do not meet the standard of one bedroom per person. This indicates that most people still live with others or families, which must also be considered in the analysis of the consequences of the pandemic's spread.

Population Density Influence. The three LGAs with the highest population densities in the Greater Sydney Region are Sydney, Waverley, and Northern Sydney. These areas have all been severely affected by COVID-19. Upon analyzing their trends on coordinates, we discovered that they share similarities such as a young median age, similar financial profiles, high living costs, and a propensity for using public transit.

4.7 Discussion and Interview with Domain Experts

The *+msRNAers* prototype applied with our aggregated COVID-19 dataset was deployed and used in the workplaces of three domain experts from the Australian government. Through interaction with domain experts, we observed the following:

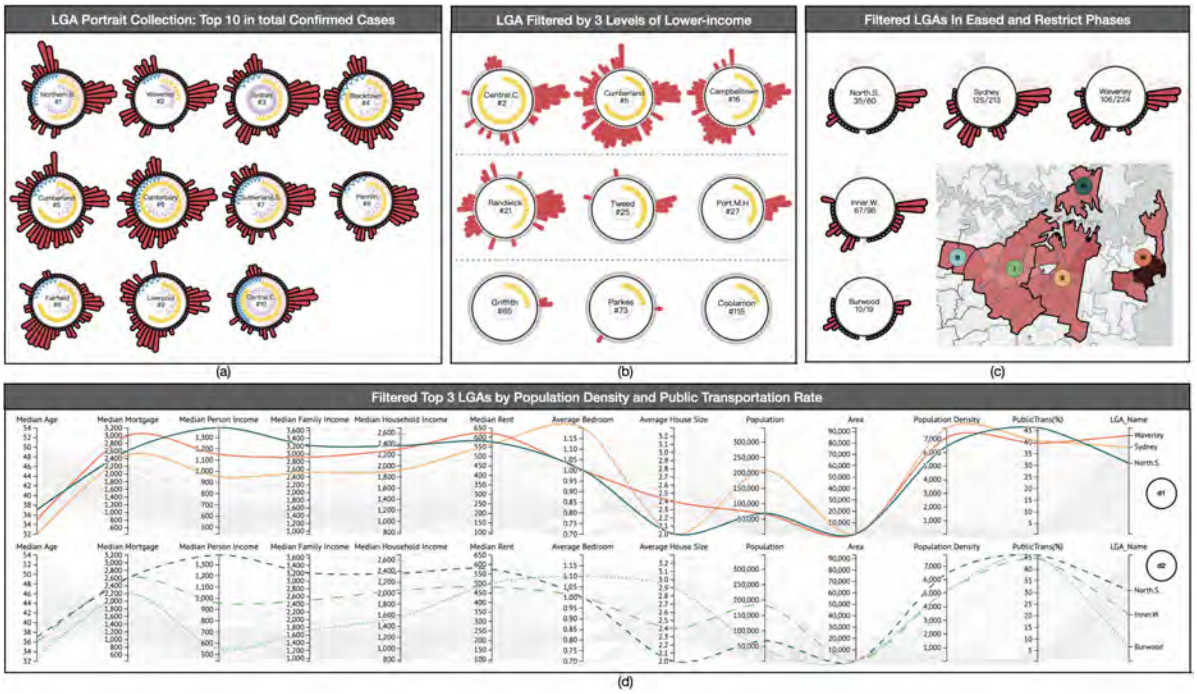


Figure 4.9: The compiled cohorts of cases are organized according to LGA portrait factors. Specifically, (a) an LGA portrait collection showcases the top ten areas with the highest infection case counts across the entire timeline; (b) a visualization shelf of LGAs is presented in three levels, filtered by lower-income factors; (c) detailed portraits of filtered LGAs are provided, and the colors of map pin-points correspond to the colors used in (d); and (d) two portrait cohorts are presented after being filtered by population density and public transportation rate, displayed in multidimensional coordinates and distinguished by colored lines to differentiate LGAs.

4.7.1 Influence of Key Risk Factors

The combined influence of key risk factors has a notable negative relationship with the community's resilience profile against the virus spread. According to infection cases, LGAs with a higher representation in four risk groups are all at the top of the list. For example, among the entire list of factors in NSW, the Central Coast has the highest-ranked key factors, ranking #1 in aged group, #1 in lone person group, and #2 in lower-income group; Canterbury also has the highest-ranked key factors, ranking #2 in aged group, #1 in lower-income group, and #3 in lone person group.

The larger the population in the high-risk group, the higher the COVID-19 cases.

Key factor(aged group): The areas with smaller aged male or female groups among LGAs in NSW have significantly reduced infection cases. In the ranking according to infection cases, the top LGAs generally have older age groups in community representation. There is not much difference in the effect caused by aged male or female groups.

Key factor(lower-income): Among four key risk factors, the lower-income group has the highest impact on COVID-19 risk. The visualization results show that infection cases significantly increase for LGAs in the lower-income group, roughly occupying one-quarter of the maximum of the lower-income group population in NSW. For example, Cumberland, Canterbury, Sutherland, Sydney, Fairfield, Penrith, Blacktown, Liverpool, and Central Coast are the LGAs with the worst COVID-19 situation, which have larger population sizes in the lower-income group from small to large.

Key factor(lone-person): The lone-person group has a certain impact as well. For example, Sydney CBD ranked 3rd according to infection cases and has key risk factor rankings: lone person (#1), low-income (#12), and age group (Male #25, Female #29). Compared with other factors across all LGAs, its influence seems the weakest. However, it would also affect the LGA's resilience if taking into account other risk factors.

Lower risk factors, higher resilience: According to the domain expert, if LGAs have a lower presentation of most risk factors, their resilience appears to be higher. This also showed in the situation where several higher-resilience LGAs would respond quickly and positively to flatten the curve and reduce the risk, even when they were the areas where these early and severe infections happened during outbreaks.

Domain experts pointed out that these implications largely support the guidelines developed by the Australian government during the pandemic. However, they would also include representation of other under-resourced people in future work, who may also be more at risk of exposure, including Aboriginal and Torres Strait Islander people, people living in aged care facilities, and people with disabilities. They plan to share with us more datasets to increase the diversity of higher-risk groups. For the lone-person risk group, domain experts explained that an individual or group's social relationships should be significantly affected and explained that a lone person would have much fewer resources and social support compared to people living together. Australian and most Western societies encourage young adults to move out and live alone. However, this situation may not be beneficial during a pandemic. Apart from living conditions, they would also like to explore diverse relationship statuses as they believe that residents' relationships might have a greater impact on social support and mental well-being.

4.7.2 Influence of Other Factors

The higher population does attract more risk for infection. All LGAs with larger resident populations should be given more attention during all phases of the pandemic. For example, Canterbury, Blacktown, and Central Coast ranked top 3 in population and have higher infection cases than other outbreak LGAs.

Travel restrictions could effectively control the risk caused by activities related to public transportation. For instance, North Sydney, Burwood, and Inner

West are the top 3 in terms of the number of people who frequently use public transit. These LGAs have the earliest cases during outbreaks before the controlled phase due to the high volume of public transit. However, they respond positively and quickly when the government increases travel restrictions.

Geographic factor: From the overview, the total number of infection cases in coastal areas is generally higher than in inland areas for the first 12 months. In particular, those areas with famous beaches had severe infections, for example, the Northern Beaches and Waverley (with Bondi Beach). Analyzing the spreading pattern in the Greater Sydney Region during the uncontrolled phase, domain experts observed that the outbreak initially appeared in scattered locations throughout the LGA and subsequently spread to areas adjacent to existing cases. During the early stage of the outbreak without control, spreading patterns were more apparent in residential areas. However, in the latter half of the phase, the pattern became more prevalent in tourist attractions, ultimately resulting in the peak of infection cases in Waverley.

Domain experts have also discovered that the combination of key risk factors and other factors can have a more significant impact. Taking Northern Beaches as an example, they had the most infection cases in the first year of the COVID-19 outbreak, with a greater effect on geographic factors and the highest presentation in lone-person groups, despite a reasonable economic indicator and low presentation in the aged group. Although most of these implications seem obvious, they would consider strengthening governance based on these factors, as these implications also prove the effectiveness of government intervention. The experts also emphasized the crucial need for clearer regulations about beaches and recreational usage.

After experimenting with experts, we also conducted a follow-up interview. Overall, experts were impressed by the "actionable insights" that *+msRNAers* obtained. They felt excited about the interface design inspired by the viral 2D structure. They expressed that it supported their memory of each data object and their representation because of the familiarity of the metaphor adopted. They commented, "This would help us reduce the effort and time involved in training other staff." They found that "interaction with humans and interface-in-the-loop is intuitive and assists them in manipulating exploration along a timeline and geographic map." They also used the prototype system to make several valid assumptions and discover unexpected implications. For assumption validation, they found that the prototype's capability to help them target higher-risk groups is "specifically useful," and the interactive portrait was "the most useful treasure" considering the magnitude of the data they have.

Experts consistently agreed that the visual modeling of *+msRNAers* can be easily applied in most epidemiological analyses due to the similarity of virus transmission, which is often caused by spatiotemporal and objective factors. The prototype will be reusable because the integrated portrait designs helped them augment awareness by porting most data features. They were swiftly able to identify higher-risk regions with "constrained factors and higher presentation in risk groups using the realistic." The experts mentioned that they expect the prototype to enable them to understand how to build profiles and predictions for communities in the future based on "reasonable and meaningful information provided."

4.8 Limitations of +msRNAers

Our work was constrained by the quality and availability of the data provided, including the related datasets and redefined risk factors with domain experts. These data issues could have led to inaccuracies in our final results. For example, COVID-19 case data collected by the NSW government is based on usual residence, but some records did not track location accurately, which could have compromised the precision of our analysis. Concurrently, the case data we had access to were recorded by LGAs or postal areas, which limited our ability to examine more granular infection trends in communities. Additionally, our use of a weekly time frame, as dictated by the aggregated datasets, may have limited our ability to uncover more nuanced patterns in the pandemic's progression. The Australian Bureau of Statistics conducted the latest census surveys on communities in 2020, but the complete summary will not be released until mid-2023. Thus, we utilized the 2016 census in aggregated datasets, which may pose bias in accurately assessing the impact of the pandemic on population demographics and social indicators during this ongoing period. Moreover, the LGA profiles we used contained a vast amount of data, with over 15,000 variables each. We could not concentrate on all variables, necessitating a reduction to only thirteen based on expert supervision. We considered social and economic indicators and indicators of vulnerability to COVID-19, such as rental and mortgage affordability. However, variables in other categories may have directly or indirectly influenced our results. Furthermore, the intervention events in the COVID-19 dataset were constrained by the Australian government system, with different levels of government responsible for administering public policies and programs, which may have resulted in varying intervention measures implemented across LGAs in NSW.

4.9 Conclusion of +msRNAers

With epidemiological analysis as a backdrop, this paper proposes a new visual modeling method called *+msRNAers* for exploring and comparing spatiotemporal and multidimensional features based on requirements in epidemiology. The method employs a metaphor to assemble portraits that can be used for visualizing each community by combining the visual encoding of time-varying case numbers with objective risk factors that may affect transmissions. The method integrates multiple views, including a Control Panel, GIS, and MDC Views, to provide wide-scope observations on filtering events at different severity levels, geo-based spreading distributions, and multidimensional risk factors for each community.

To evaluate the feasibility and effectiveness of *+msRNAers*, we deployed and applied a two-year-scale aggregated dataset by integrating COVID-19 cases with geo-information, NLP-extracted event division on timelines, and risk factors from NSW census data based on expert supervision. After applying *+msRNAers* to this COVID-19 aggregated dataset, we progressively validated the feasibility and effectiveness of *+msRNAers* by conducting one user study to iteratively improve the applied prototype and comparing visual portraits from profuse perspectives in three subject-driven case studies. We summarized how geography, phases of intervention events, and objective risk factors affected COVID-19 spreading situations during the pandemic.

In further interviews with domain experts, we identified additional objective factors that may be influential in the Australian context. We examined pre-existing community factors and discovered practical implications for potential patterns of established community characteristics against the vulnerability facing this pandemic. Despite some limitations and future work, feedback from domain experts suggests that the *+msRNAers* can be considered a common visual modeling method for exploring community-based spatiotemporal and multidimensional features. This method can be applied to abundant epidemiological analyses, such as investigating case trends and comparisons, geo-distribution and transmission, risk factors explorations and rankings, and other related tasks.

CHAPTER 5

UcVE: User-centered Visual Explorer of In-Process Comparison in Spatiotemporal Space

5.1 Research Scope of UcVE

We cooperated with domain experts and further developed an original User-centered Visual Explorer (*UcVE*) based on unit visualization with scalable aggregated views. The spatiotemporal features are encoded in each visualization unit for their variation. *UcVE* allows users to view, save, and track the outcomes of in-progress exploration to lessen the cognitive load on the user. *UcVE* can facilitate concurrent comparison with multiple visualization units, which are selected from historical and current exploration results in coordination with storage sequence and block tracking views. To maximize the user's exploration ability, *UcVE* provides a flexible geo-based layout, aggregation functions, and temporal views of the timeline with categorized events. Using COVID-19 datasets, we present case studies in various user contexts. We also worked with domain experts to discuss our case findings and provide expert feedback for evaluations.

UcVE makes the following main contributions in three orientations:

- **Design-oriented: A visual metaphor of unit visualization with the customizable aggregated view** to expand the visual representation scalability (C1, C5). The *UcVE* maps unit visualization with the encoded abstraction of spatiotemporal information in three statuses: single, auto-clustered, and custom-merged (C2), which allows for easy tracking and details on dynamic demand for each individual variation (C4), making exploration easier.

- **Human-oriented: User-centered Progressive exploration approach with saving and tracking in-process visualization results** to ease user cognition workload. Users can record each interacted result on a map as a storage sequence for further iterative callback and exploration to deal with spatiotemporal region discretizations (C3, C4); ranking and target tracking can also be done interactively between different storage sequences (C5).

- **Comparison-oriented: Comparative Visualization with user-selected multiple visualization units concurrently,** breaking to geo-barriers to explore implicit relationships among units (C1, C3). Combined with other visualization views and interactions, a comparison matrix view enables a detailed comparison of spatiotemporal attribute values among multiple visualization units in a scalable manner (C5).

5.2 Requirements Analysis

Our goal is to create a visual approach that allows users to progressively explore and compare multiple visualization units with different spatiotemporal properties.

We know from previous research that analyzing spatiotemporal data is one of the most difficult aspects of visual analysis, so how to clearly and completely abstract spatiotemporal features is our main design requirement. Other related properties, for example, multivariable and textual data, are usually present with spatiotemporal data. In addition to visualizing all spatiotemporal features, we must consider designing and implementing user-based interaction methods that are applied to every individual user scenario to assist users in in-depth further complex spatiotemporal dataset analysis.

This work aims to focus on the majority of analytical tasks in spatiotemporal data which can be abstracted from their spatial location accompanied by numerical attribute values in their temporal features. In Section 5.5, we introduced the COVID-19 crisis context as an example of typical spatiotemporal datasets and applied them to *UcVE*. We also discussed the design requirements in the collaboration with domain experts and analysts in the NSW government and distilled the requirements as below.

R1: Draw encoded visualization units in a reasonable manner. The fundamental method of mapping the geographic coordinates of units to a map makes a geo-based layout easy to implement but encoding the adjustable temporal features of each unit while displaying the geographic coordinates remains a design challenge that needs to be addressed. As a current instance, the analysts in NSW Government released a visual interface [201] for COVID-19 basic statistics for each geo-spot which used colored heat-polygons on a playable timeline rather than abstracted as visualization units for each spot. This consequence makes inspection worse when exploring polygons with large size differences on different scales, and the detailed values are not evident to map colors on polygons.

A flexible overview of encoded units within a limited view of map space is required to clearly interact with and explore multiple visualization units with spatiotemporal features in a reasonable manner. The NSW COVID-19 statistics interface only supports inspecting the confirmed cases counted in each spot in the past month, also, it only shows the trend of data changes rather than allowing for numerical comparisons of various geo-spots within a constrained area or attribute comparisons of map overviews across different timelines.

R2: Record encoded visualization units in various analysis results. To support the user-centered exploration of historical and current results, implementing the storage function is needed to allow users to record the results of each interactive exploration and the relevant map parameters for a later callback. Simultaneously, multiple storage steps can break the geographical barrier by allowing two units that are far apart to be checked and to track and compare temporal features underlying different timelines.

R3: Facilitate detailed comparisons of each encoded visualization unit in user-centered scenarios. One of the difficult parts of the comparison task is comparing detailed data that varies in both time and location. It should be capable of facilitating detailed comparisons of the same geo-location underlying different time phases, as well as data attribute value comparisons of multiple geo-locations underlying the same time phase. The timeline should include multiple categorized options.

5.3 Visual Design

To allocate the design requirements outlined in Section 3, we proposed a visual design, which employs visualization units for user in-process exploring and comparing spatiotemporal datasets. Any spatiotemporal datasets with spots in geographic coordinates and numerical values in time-vary attributes for each spot can be applied to this design easily. The visual design of each units is inspired by the Sunburst[195] and Off the Radar[13] visualization of flower-like metaphors[60].

In this section, we demonstrated the visual design of units, as well as user-centered progressive exploration strategies with *UcVE* Storage and *UcVE* Matrix interactions. For each distinct unit, we abstracted the spatiotemporal features into petals, and distinguished three statuses of visualization units by different colors in centers, as shown in Figure 5.1(a). The visualization units with a single unit (navy color), visualization unit with auto-clustered units which are automatically clustered according to different zooming scales (green color), and visualization unit with manually merged units (orange color) are defined based on current exploration scenarios, as shown in Figure 5.1(b).

5.3.1 Visualization Units Design

We investigated redesigning each spot as a visualization unit with timeline attributes and mapping them to each geo-location to fully describe spatiotemporal features. The visualization unit should enable converting each unit into a single unit on the map, allowing each node to depict spatiotemporal features with a compressed adjustable time span.

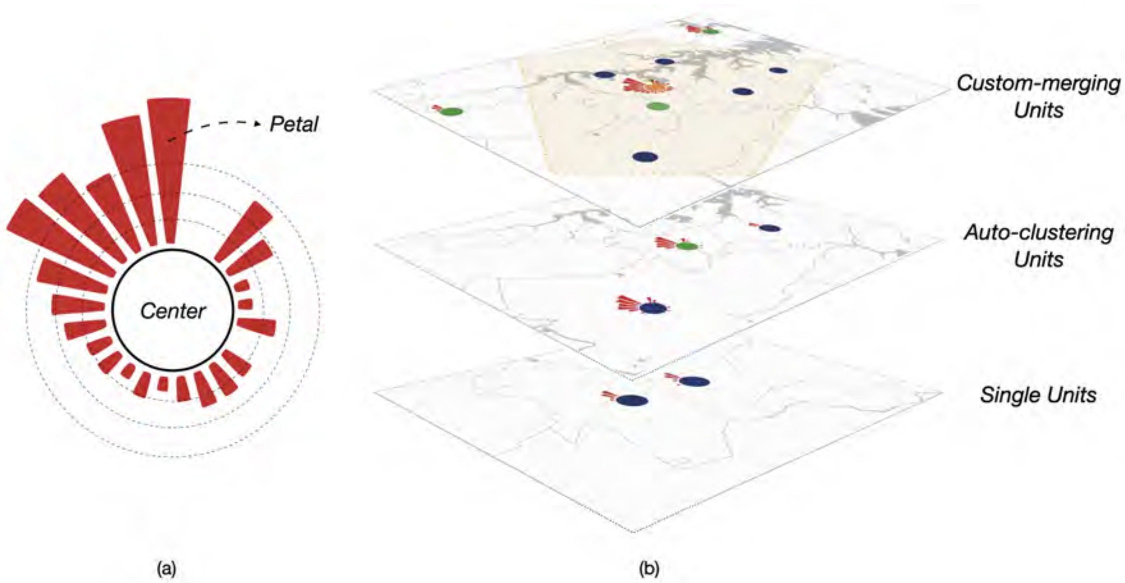


Figure 5.1: The design of visualization units. (a) A basic visualization unit with encoded petals and a colored center; (b) Three exploration scenarios of visualization units based on varying scales: single units, auto-clustering units, and custom-merging units.

Defining single visualization unit. The circular node is commonly used to pinpoint location information on a map. We locate the centroid of the polygon area of each target unit as the geo-locations of a single spot. To symmetrically attach temporal attributes, we redesigned the radial directions of the nodes for visualization. The x-axis of the polar coordinates is defined as the perimeter of a circular node, whereas the y-axis is defined as the vertical direction. Simultaneously, the perimeter of each node is sliced into a predetermined number of identical-sized spans to meet the visual discrimination channels of the user. Simultaneously, the attribute values of the corresponding time spans are allocated to each slice of the perimeter in a radially outward direction with different histogram heights (named petals). The formula for calculating petal *i*'s height in a single unit *s* is defined as:

$$\begin{gathered} P(s_i, w) = a \cdot \ln(F(s_i, w) + b), \\ F(s_i, w), a, \text{and } b \in N^* \end{gathered} \tag{5.1}$$

A growth petal height $P(s_i,w)$ of this visualization unit which refers to a single unit is proportional to the attribute value, in which $F(s_i,w)$ is the function to count the numerical value corresponding to this week w. Both a and b are positive integer parameters that can be adjusted to ensure that the petal height is *0* when the infection number is *0*.

Defining auto-clustered visualization unit. When there are too many rendered nodes, the query results on the map frequently cause overlapping issues, resulting not only in difficulty in recognition, but also vastly increases the client's rendering cost. We employed a progressive map node clustering method to solve this problem, which can automatically cluster nodes based on the provided clustering radius and regenerate a new cluster center to depict multiple nodes within the acceptable visible area, depending on the map's current zooming scale, the cluster unit consists of all clustered single units which defined as $F(c_j,w) = \sum_{i=1}^{n} F(s_i,w)$. The function in *Equation (2)* is applied to calculate the petal *i* in week *w* of clustered visualization unit *c* which is clustered by other singles.

$$P(c_i,w) = a \cdot \ln(\sum_{i=1}^{n} F(s_i,w) + b) \tag{5.2}$$

The recalculated cluster centers should be biased towards the denser areas of the children units, and the clusters under different zooming scales based on map tiles should be weighted separately, so a step-weighted coordinate calculation should be applied.

Defining custom-merged visualization unit. Similarly, the heights of the petals in each clustered node should be reunited, but because simple superposition would result in excessive heights, we reframe the merged petal heights while merging the different statuses of nodes' location sets. At this point, the merged petal height represents the sum of the infection cases in each time span of all the clustered nodes. After solving the node overlap problem by revising the clustering method, we realized that the automatic clustering method can sometimes result in geographical classification errors, which will hinder the user's ability to explore target nodes, so we developed a visual analysis interaction based on drawing a polygon selection to merge customized aggregation.

As a result of the merge, the merged visualization unit m calculates the center and each petal's height using a similar algorithm which sums each time span of infection cases in each merged visualization unit, to the units' set and marks the centroid of the newly merged polygons as its location on the map. The petal i's height of the merged unit is defined as:

$$P(m_i,w) = a \cdot \ln(\sum_{i=1}^{n} F(s_i,w) + \sum_{j=1}^{m} F(c_j,w) + b) \tag{5.3}$$

Where $P(m_i,w)$ represents the petal height in week w of the visualization unit merged by other clustered and single visualization units.

Any polygon G can be divided into n finite simple triangles $T_1,T_2,\ldots,T_n$ where the centroid of these simple triangles is (T_{i_x},T_{i_y}) and the area is A_i. So the coordinates of the centroid of the newly merged polygon are *(G_x,G_y)*.

$$G_x = \frac{\sum C_{i_x} \cdot A_i}{\sum A_i}, G_y = \frac{\sum C_{i_y} \cdot A_i}{\sum A_i}, \tag{5.4}$$

A brushable timeline linked with visualization units should be attached to help filter time phases more efficiently. Each visualization unit utilizes the adjustable segmented circular arc as the time span unit, with radially arranged petal heights displaying the attribute values in this time span.

However, because of the geographic distance between nodes and the radially grown petal directions, comparing the petal height of different visualization units will be more difficult. To demonstrate the differences among visualization units, we introduce comparison strategies in other visualization designs and interactions.

5.3.2 Visual Exploration Strategies

The comparison strategies guide users on supporting exploring and comparing analysis tasks. Most geographic visualizations aim to output analyzed results while users interact with a map in diverse manners. However, the interaction results displayed by the map usually vary in different states. To effectively solve this problem, we implemented a user-centered visual explorer (*UcVE*) based on geo-map with viewing, saving, and tracking in-process visualization results.

UcVE Storage: Ranking. Considering that the user may vary the zooming scale or time phase of the view when exploring the map, we allow the visualization unit states of the current perspective to change.

The visual storage should support the saving of all visualization units and parameters set on the map during each step of exploration for easy reloading later.

Each visualization unit is compressed into a colored block and queued into a sequence with all units on the current map view. Considering different units occurs based on different map scale, we encode each block with a nested block to indicate belonging relationships. For example, block A is made up of blocks B, C, and D, and we have recorded A in sequence 1, and B, C, and D in sequence 2. If we want to track B, the nested block in A will be highlighted with the same color as B. Both Figures 4 and 7 show this situation.

Each storage step records each of the user interaction results in an ordered sequence, which is then ranked from left to right in descending order of each visualization unit's total attribute value sum, and compresses each visualization unit into a nested-able block to depict the containing relations of units among different storage sequences.

We highlighted the associated visualization blocks that have the same spot names or contain spot names in different storage sequences. The time span indicator, which utilized the number of each storage sequence, is marked in front; the outer arc length of each number maps the time span of the storage and can be dragged to swap orders among other storage sequences, as shown in B_6 in Figure 1.

UcVE Matrix: Comparison. The *UcVE* storage function greatly increases users' analytical scalability. To further compare the detailed value difference among visualization units, we introduce an $n \times m$ matrix space. Each grid in the matrix can visualize the detailed attributes of each visualization unit. The parameters n and m can be altered to accommodate users' appropriate resolution. Each draggable grid is listed with comparable visualization units that aim to break through the limitations of objective factors (visualization units with different zooming scales or views, different time spans selected, etc.). However, in juxtaposition, it is difficult to precisely compare visual differences among items that are far apart, even if they can be adjusted in order. Therefore, we employed an explicit encoding on the visualization unit that can extract and represent visual critical differences among juxtaposed visualization units in grids, rather than blinding users to minor changes in their comparison. The comparison strategy of each visualization unit in the grid is a benchmark peer-to-peer visual analysis method. Specifically, this comparison strategy involves selecting one visualization unit as the compared benchmark interactively, while other visualization units automatically vary based on the compared benchmark, and the encoded visual results of the changes show the difference in comparison.

This comparison strategy allows for highlighting the associated petals of the compared visualization units that correspond to the petals of the selected benchmark, as well as encoding the petal shape (raised or depressed) in other compared visualization units to show the compared differences. We listed two pre-options to encode the petal shape: Bezier curve and Polyline.

A Bezier curve, which is usually intended to approximate a real-world shape, is defined by a formula that uses a set of discrete control points to define a smooth, continuous curve. The definition of a Bezier curve indicates that the first and last control points are always the endpoints of the curve. However, the intermediate control points generally do not lie on the curve. This means that if the Bezier curve is used to describe the petal's outer shape, the expected accuracy may not be depicted, as shown in Figure 5.2. Therefore, we rule this option out.

The Polyline, in contrast to the Bezier curve, uses an accurate value to better describe raised or depressed differences. We defined that the radius of this petal at its highest or lowest point equals the radius of the compared benchmark petal P_{bench}, the current petal height $P_{current}$ describes the height of this petal edge, as shown in Figure 5.2.

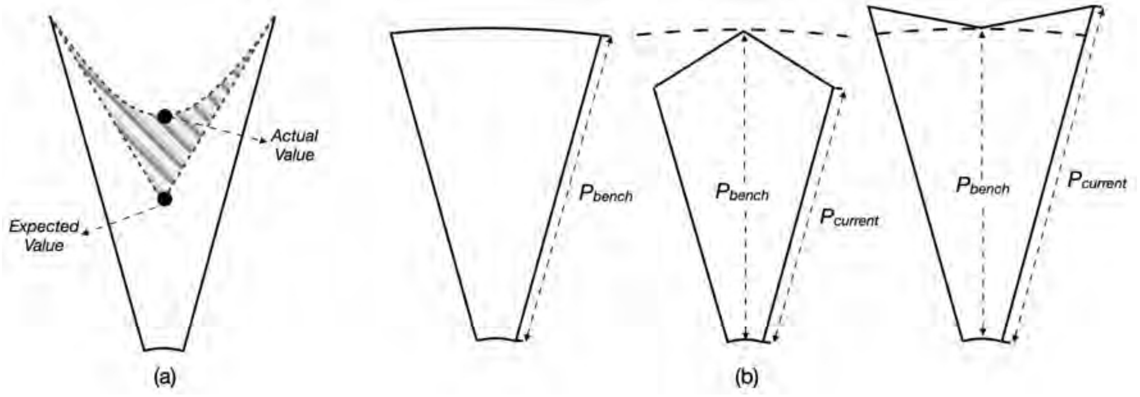


Figure 5.2: The comparison strategies in *UcVE* Matrix design. (a) It will cause value bias when using a Bezier curve to depict the petal's outer shape; (b) It shows what is possible when using Polyline to plot visualization unit petals to compare the value differences.

Furthermore, we added numbers and time span units to the colored center of the visualization unit to indicate the exact value of the entire visualization unit or one specific petal when hovering over it.

5.4 COVID-19 as Spatiotemporal Dataset Example

Spatiotemporal data relates to both space and time [81]. There is typically various information at various times on every spatial spot, and this information is typically accompanied by multi-dimensional attributes. When users encountered the tasks of comparisons among them, variation can bring analysis difficulties. As a result, it is crucial to separately represent the features of various dimensions, including time and space.

In the past two years, we have engaged in research on COVID-19 data in New South Wales (NSW), Australia. Supported by domain experts, we designed different visual tools to reveal useful patterns behind the basic statistics. During these collaborations, we noticed that the uncertainty of the COVID-19 epidemic means that traditional data analysis performs more poorly than effective visual models. In particular, the spatiotemporal features in hot spots, as inseparable parts of the epidemic analysis, played a significant role in analyzing the COVID-19 spread. The following subsections present COVID-19 as our data source, followed by variables of consideration corresponding to the design requirements.

5.4.1 COVID-19 Data Sources

The NSW Government has made available open-source datasets for COVID-19 cases and tests within NSW[1]. Our example dataset provides information on the number of daily infections based on COVID-19 case locations since the first infection was detected in Jan. 2020 to 2022. It excludes 189 cases of crew members who tested positive while onboard a ship docked in NSW at the time of diagnosis. The case aspects include confirmed, tested, recovered, and death by their notification date, location, age range, and likely source of infection, but the analyzing task becomes complicated because the available data variables are continuously adjusted and reduced due to privacy and other reasons. In addition, relevant news articles, alerts, and ministerial media releases issued by the NSW government on COVID were retrieved as events that can be combined with the notification date of the case dataset. As common division methods of usual residence, 128 Local Government Areas (LGAs) and 964 postal zones within NSW boundaries were included in the geographical datasets provided by the NSW Data Analytics Center (DAC), which were introduced with geo-polygon and geo-coordinates data.

5.4.2 Variables Consideration

As domain experts guided us to further meet the requirements, we discussed relevant key variables selected from multiple datasets.

Our COVID-19 case dataset contains information collected from January 1st, 2020 to January 31st, 2022 in NSW on cases by notification date and postcode, local health district, LGA, and likely source of infection. Due to the risk of personal information being leaked, only the notification date of confirmation and the location of each case's usual residence can be accessed in this dataset.

During the data cleaning process, we discovered that issues arose when the same case corresponded to multiple postal zones; the 648 postal zones resulted in overcounts, making identification more difficult. The LGA, which is defined as an official spatial unit that may contain multiple postal zones, represents a reunited geographical neighborhood. Therefore, using LGAs can help to solve the situation where the same case is counted multiple times. Simultaneously, it can improve the map's construction and readability. As a consequence, we chose 128 LGAs as the geo units for the aggregated area.

Considering that different government-released events caused different degrees of impact on the pandemic spread, we organized the retrieved media information of government-conducted events into two categories: intervention and restriction. Each time span unit was considered to cover all signature events in the past 2 years, which consisted of monthly cases summed up daily and listing 25 months of case data from each LGA.

In this study, we finalized 98298 rows of 25-month infection case data from Jan. 1st, 2020 to Jan. 31st, 2022, which we bridged with 2 categories of 12 significant events from 128 LGAs.

5.5 Visual Exploration Workflow

In this section, we propose a user-centered progressive workflow for the interactive visual exploration based on D3.js[28] and Mapbox [133], and further implement a visual analytics system applied to the COVID-19 datasets in NSW, Australia to demonstrate *UcVE* from a workflow perspective, as shown in Figure 5.3.

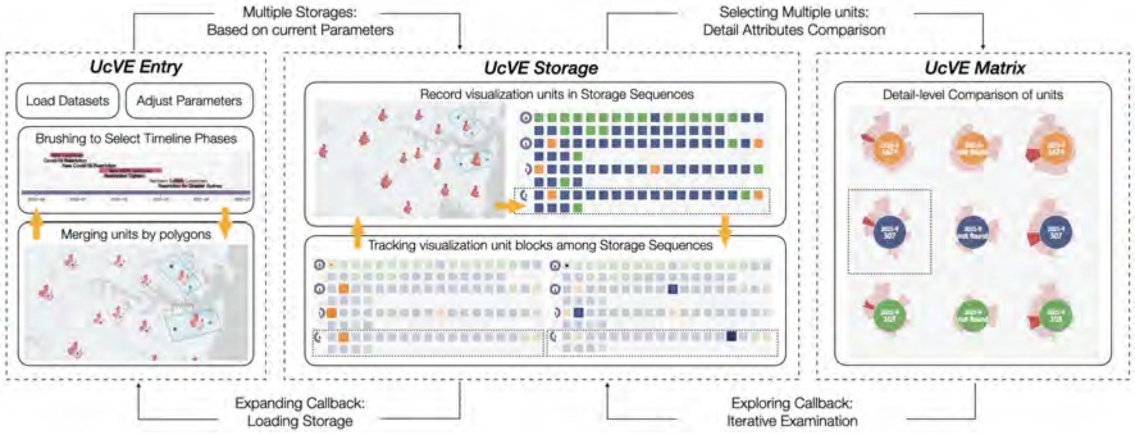

Multiple Storages:
Based on current Parameters
Selecting Multiple units:
Detail Attributes Comparison
UcVE Entry
Load Datasets
Adjust Parameters
Brushing to Select Timeline Phases
Merging units by polygons
UcVE Storage
Record visualization units in Storage Sequences
Tracking visualization unit blocks among Storage Sequences
UcVE Matrix
Detail-level Comparison of units
Expanding Callback:
Loading Storage
Exploring Callback:
Iterative Examination

Figure 5.3: The entire workflow of *UcVE* embodies a user-centered progressive and comparative exploration, which consists of *UcVE* Entry, *UcVE* Storage, and *UcVE* Matrix.

5.5.1 UcVE Entry: Parameter Setting

After the visual analytics system loads the COVID-19 data, the user interacts by setting the parameters of the clustering radius and basic time span unit. The value of the clustering radius equals the width of a tile. The tile on the map is defined as a segmentable basic unit on the current zooming scale; the basic time span unit specifies the time span of the minimal unit represented by each petal in each visualization unit. Users can transfer multiple visualization units and all parameters for further exploration and comparison as entries to *UcVE* storage in sequence, including cluster radius, time span unit, current map view and zooming scale, and all unit data in the selected time span.

5.5.2 UcVE Storage: Iterative Exploration

The scalabilities are expanded by the diverse permutations and combinations of the transferred parameters to represent the various visual exploration results of multiple visualization units in which the user performed multiple storage steps. Users can iteratively zoom in or out to different scales to expand or recluster target units, as well as produce storage of multiples by adjusting the selected time spans. Simultaneously, the ranking strategy in *UcVE* storage supports tracking and highlighting each visualization unit into associated block that refers to the same or contains spots in different storage sequences. Moreover, each historical or current storage sequence can be reloaded on the map as a callback to reflect all the visualization units' situations in this storage sequence.

5.5.3 UcVE Matrix: Detail-level Comparison

Users can drag and drop interesting visualization unit blocks into the *UcVE* matrix view for detailed comparison. Users can compare the similarity-difference patterns showcased by the encoded petals' outer shape in this matrix. (1) visualizing one unit within a selected time phase; (2) tracking the varied trends of one specific unit in different time phases; (3) listing the detailed attribute distribution of different units in different time phases; and (4) comparing the detailed attributes of the same time span among multiple units are some of the possible patterns that can be explored or compared. Users can also progressively explore until they reach the conclusion by iteratively removing any visualization unit and adding new ones from multiple storage sequences.

5.6 Case Studies

In contrast to more sparsely populated areas, the coastal areas of NSW are more densely populated, making them more susceptible to cluster infections. For instance, over the last two years, the government has conducted periodic intervention and control events in all or parts of New South Wales, which may have suppressed the epidemic infection at various times.

In this section, we present three categorized cases: map-driven, storage-driven, and matrix-driven, which we progressively employ to explore and compare important connections and differences between spatiotemporal properties in the NSW COVID-19 dataset, as well as attempt to find the covid behind the influence of LGAs on potential pandemic patterns. Relevant parameters must be pre-entered into the visual analysis system. In these case studies, users can customize the clustering radius, which corresponds to the tile width in the current map perspective, and the basic time span is set to monthly. The number of blocks which are used in each storage sequence depicted in each row is 18, and the size of the *UcVE* comparison matrix is 3 rows by 5 columns.

5.6.1 Map-driven Cases

In the initial map view, users can observe both automatically clustered visualization units and individual units, which are distributed on the map based on their geo-location. Users can adjust the zooming scales by zooming in and out and custom-merge their target visualization units by interactions to gather new visualization units of different statuses for map-driven case exploration. With the help of the brushable categorized timeline, users can adjust different time phases to filter visualization units on the map. In these map-driven cases, the clustering radius is set to 70 to better distinguish auto-clustered visualization units.

Visualization unit comparison of coastal LGAs with inland LGAs. We roughly divide NSW into two sections, coastal and inland. To better display visualization units on the map, we use the merged interactive tool to select four areas, two coastal and two inland as follows: the northeast coastal area near the border of Queensland; the southeast coastal areas near the border of Victoria, which has representative cities like Sydney, Newcastle, Wollongong, etc.; the north-central areas; and the remote western areas, as shown in Figure 5.4.

When users look at the entire timeline over 25 months, they can observe that both the two parts of coastal parts are ranked top 2 with the total number of COVID-19 infection cases in the whole period of 25 months, which is specified as the southeast coastal areas near Victoria win the first with merging 3 auto-clustered visualization units, then the northeast coastal areas near Queensland reached the second with merging 2 auto-clustered visualization units, and followed by the remote western areas containing 6 auto-clustered visualization units and 4 single visualization units, and accompanied by the north-central areas, which is the last one with 2 auto-clustered visualization units and 1 single visualization unit.

The reason for such widespread of COVID-19 cases may be due to the geographical features of New South Wales. New South Wales is Australia's most populous state, with a large number of densely populated cities along its coastal sides; rich forest resources

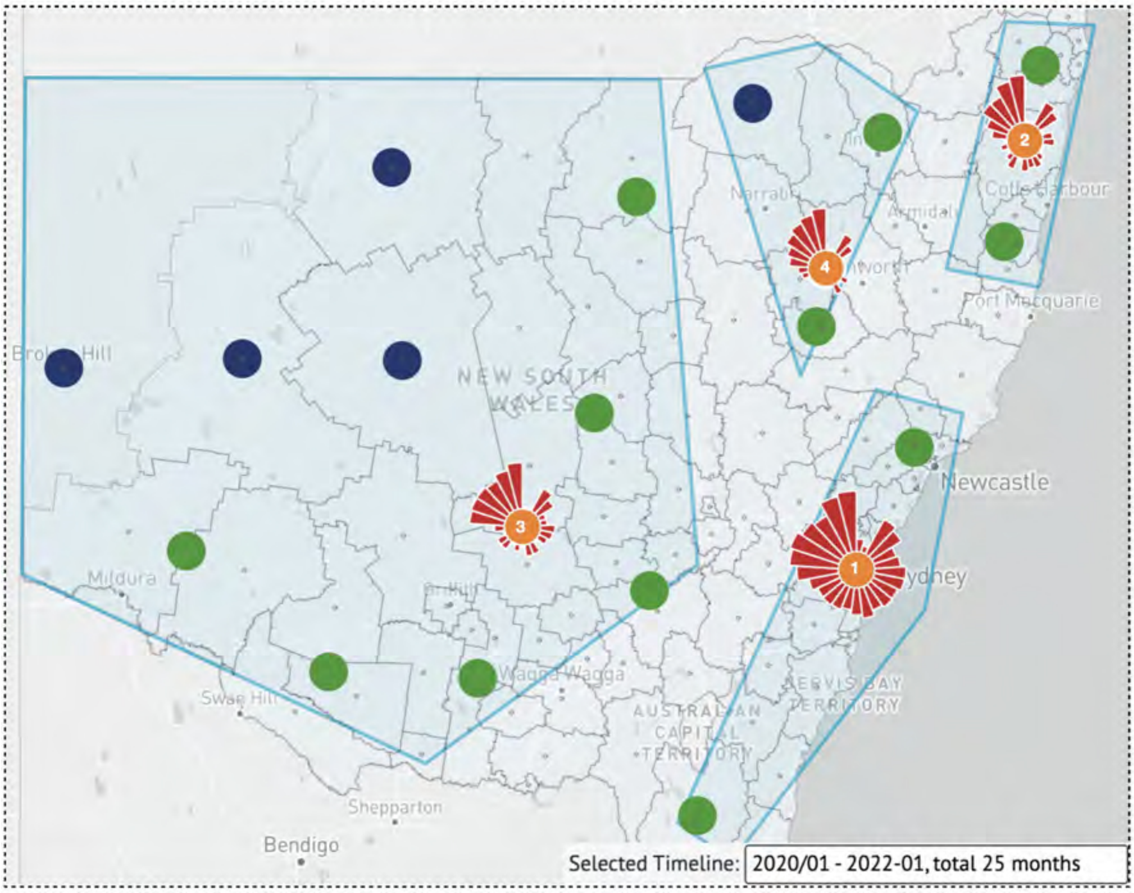


Figure 5.4: A map-based case with rankings in each visualization unit center. Using the interactive merge function, the 128 LGAs were classified into 4 visualization units representing the southeast coastal areas near Victoria (Rank 1); the northeast coastal areas near Queensland (Rank 2); the north-central areas (Rank 3); and the remote western areas (Rank 4) within NSW for exploratory analysis.

in the north-central areas; and less populated remote western areas. We know that the epidemic's spread is directly proportional to population density. This explains why COVID-19 outbreaks are more severe in densely populated coastal areas, despite the fact that the LGA area size is much smaller than in remote western areas.

Visualization units comparison of Northern and Southern Sydney LGAs. Users can brush to select time phases by swiping through the entire timeline. We chose the first 4 months of the COVID-19 outbreak in this case to investigate the epidemic situation in Sydney and surrounding cities. We zoomed into the bottom-most visualization unit displays around Sydney, then merged North Sydney with its adjacent neighbor, Mosman for further analysis on the map, as shown in Figure 5.5. According to the petals of single visualization units representing Ku-ring-gai, Burwood, and Parramatta LGAs, we found that they had the first local cases in January; also, the ranking of the total number of confirmed is Australia's most populous state, with a large number of densely populated cities along its coastal sides; rich forest resources in the north-central areas; and less populated remote western areas. We know that the epidemic's spread is directly proportional to population density. This explains why COVID-19 outbreaks are more severe in densely populated coastal areas, despite the fact that the LGA area size is much smaller than in remote western areas.

Visualization units comparison of Northern and Southern Sydney LGAs. Users can brush to select time phases by swiping through the entire timeline. We chose the first 4 months of the COVID-19 outbreak in this case to investigate the epidemic situation in Sydney and surrounding cities. We zoomed into the bottom-most visualization units displays around Sydney, then merged North Sydney with its adjacent neighbor, Mosman for further analysis on the map, as shown in Figure 6. According to the petals of single visualization units representing Ku-ring-gai, Burwood, and Parramatta LGAs, we found that they had the first local cases in January; also, the ranking of the total number of confirmed cases found that visualization units representing Woollahra and Waverley had the highest total number of infection cases among the 16 visualization units around Sydney over the selected 4 months.

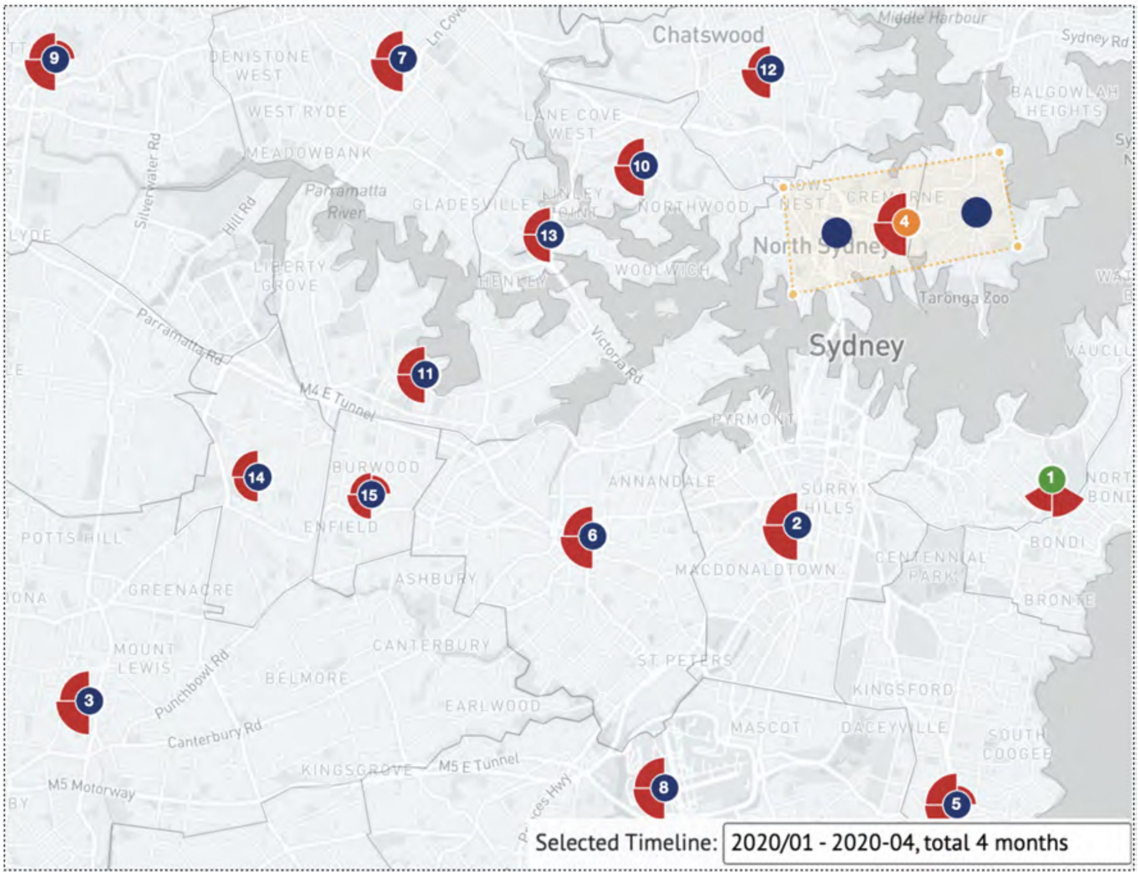


Figure 5.5: A map-based case with rankings in each visualization unit center. The first 4 months of the full timeline month (January 2020 to April 2020); merged Northern Sydney and Mosman from the perspective of the Sydney city map to explore and analyze the epidemic situation in northern and southern Sydney.

We note that the LGAs where the first local covid cases occurred in January 2020 were Ku-ring-gai, Burwood, and Parramatta. These three LGAs are all major residential areas in the Greater Sydney District, which also proves the hypothesis that COVID-19 is prone to outbreaks in these places; and North Sydney and Mosman are the main business districts in Sydney. Compared with its southern neighboring LGA, Sydney (different from the Greater Sydney District), the north of Sydney is the main working area, while the south of Sydney is mainly for entertainment, with a large number of shopping malls and famous beaches in Bondi. As reported on various news outlets, after the Australian wildfires had been contained at the end of 2019, large crowds poured into beach destinations for a vacation, resulting in an extremely high population density, which became the source of the spread of the COVID-19 outbreak in early 2020, indirectly leading to the first lockdown in NSW.

5.6.2 Storage-driven Case

Multi-segment time phases can be divided by brushing the timeline in the 25 months and recorded into the storage view for tracking and the exploration of the number of COVID-19 cases represented in visualization units' petals. In these storage-driven cases, the clustering radius keeps the set at 70.

Tracking the ranking of multiple visualization units. We drew a polygon to select the entire Greater Sydney Districts from the overview and classified the whole timeline into two options. One option is to divide the entire timeline into two years, 2020 and 2021; the other is to divide the entire timeline on a 4-monthly scale, shown in the time span indicator. The ranking track of each using the storage function to record their visualization units in batches is shown in Figure 5.6.

We selected 4 representative visualization blocks, namely Northern Sydney, Southern Sydney, Burwood, and Chatswood, from those recorded for comparative analysis. Our observations reveal that the four blocks are all from the first visualization block in the 0th storage, which is the Sydney LGA's visualization unit block with the largest number of COVID-19 cases at the time of the initialized overview. Sydney's epidemic performance is worse than Burwood and Chatswood from an overall perspective, whether in years or in a 4-month phase. Analyzing the growth trend, we found that Northern Sydney, Southern Sydney, and Burwood all showed signs of a rebound in the number of covid cases from August 2020 to December 2020, while Chatswood experienced a rebound in cases from December 2020 to April 2021. Based on the combined NSW government interventions and restrictions, it was found that during the two lockdown periods (March 2020 to June 2020 and July 2020 to November 2020), there was a short period of unblocking, which led to the second wave of the covid epidemic, and after the second lockdown in November, the third spread of Omicron.

Among their own visualization blocks, the total case rankings in the Northern Sydney region fluctuated greatly, while the Southern Sydney region was always in the top 8; on the contrary, Burwood and Chatswood never entered the top 8. We also combined official information from NSW census data [154] to find that COVID-19 spread quickly in crowded residential areas, such as South Sydney, Burwood, and Chatswood, which are popular tourist destinations, and led to a secondary transmission of the virus when the lockdown period ended whereas, in North Sydney, which acts as a business district, the number of infections responded quickly to government interventions because people worked from home during the lockdown period.

5.6.3 Matrix-driven Cases

The visualization units in the matrix allow for comparison at a fine granular level. The user is not only able to select any visualization unit from the storage view, but they also enable examining the differences in each petal's attribute values by observing the

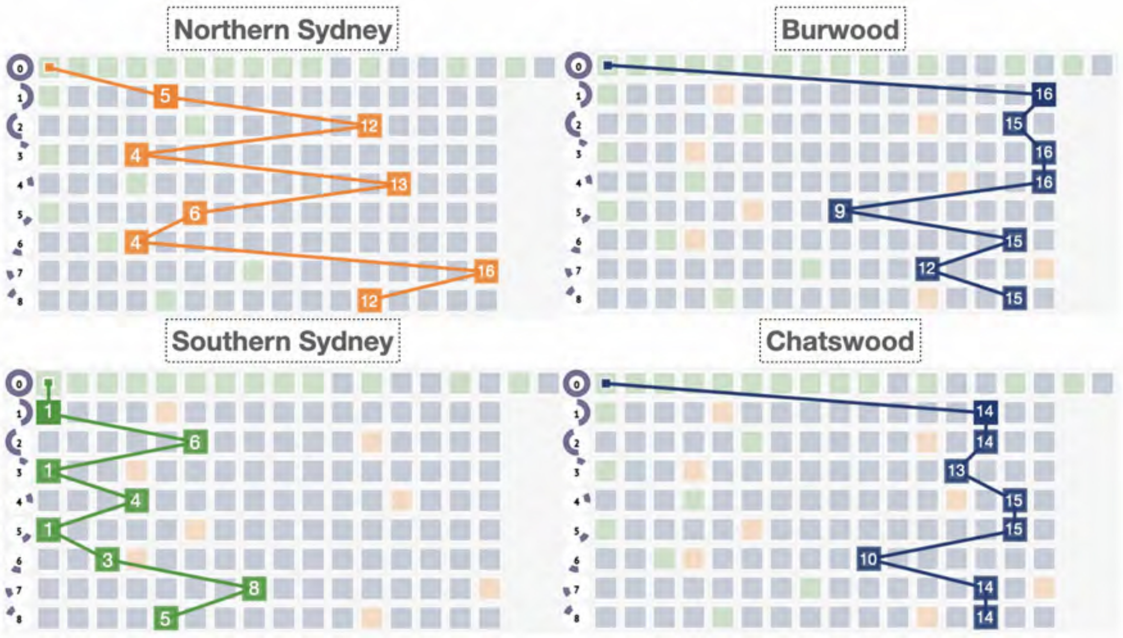


Figure 5.6: The storage-driven case. For visual analysis of their COVID-19 case rankings through two timeline classifications, four representative visualization units, Northern Sydney, Southern Sydney, Burwood, and Chatswood, were selected. Of these, the 0th storage represents all visualization units stored in the overview state; the 1st and 2nd storage steps respectively represent the epidemic data for the years 2020 and 2021 under the zooming scale of Sydney City; the 3rd to 8th storage steps respectively represent the storage of visualization units in four-monthly time span units from January 2020 to December 2021.

various petal's outer encoded shapes.

A comparison of multiple visualization units before and after government interventions. In this case, the clustering radius is set to 70. The intended time period we brushed for this case starts from January 2020 to July 2020, then we recorded the interaction results and picked the top 5 visualization units for a detailed comparison in this matrix after custom-merged the visualization units of Northern Sydney and Mosman. In terms of overall COVID-19 cases, the auto-clustered visualization unit (Woollahra and Waverley) had the highest number of infections, followed by three single visualization units (LGA Sydney, Canterbury-Bankstown, and Cumberland). The combined visualization unit of Northern Sydney and Mosman is fifth, as shown in Figure 5.7.

By comparing the petal shapes in the coming months, we discovered that the auto-clustered visualization unit of Woollahra and Waverley has much lower cases than the other three single visualization units (in decreasing order, these being Cumberland, Canterbury-Bankstown, and Sydney) from May to July 2020. Moreover, Woollahra and Waverley had 0 diagnosed cases in May 2020, and only in June and July did these areas have more than the combined nodes of Northern Sydney and Mosman (Northern Sydney and Mosman had 0 infection cases in May and June 2020).

We note that the time phase we selected from January 2020 to July 2020 includes the vacuum period from January to March 2020, the first lockdown from the end of March to June 2020, the related restriction in June 2020, and the starting point of the second lockdown beginning in July. This visual finding verifies the earlier case analysis that a huge number of individuals flocking to coastal areas acted as a catalyst for the epidemic's early development. COVID-19 spread when the government began intervening with lockdown events in March, causing new cases in the LGAs in residential communities. Because Northern Sydney and Mosman are both in the CBD, there were no new cases in the lockdown from May to July.

Exploration of multiple visualization units during government interventions. In this case, the clustering radius is set to 50. Due to the vast majority of government intervention events targeting the Greater Sydney Districts, we zoomed into the view of the Greater Sydney Districts on the map. We found that the Waverley and

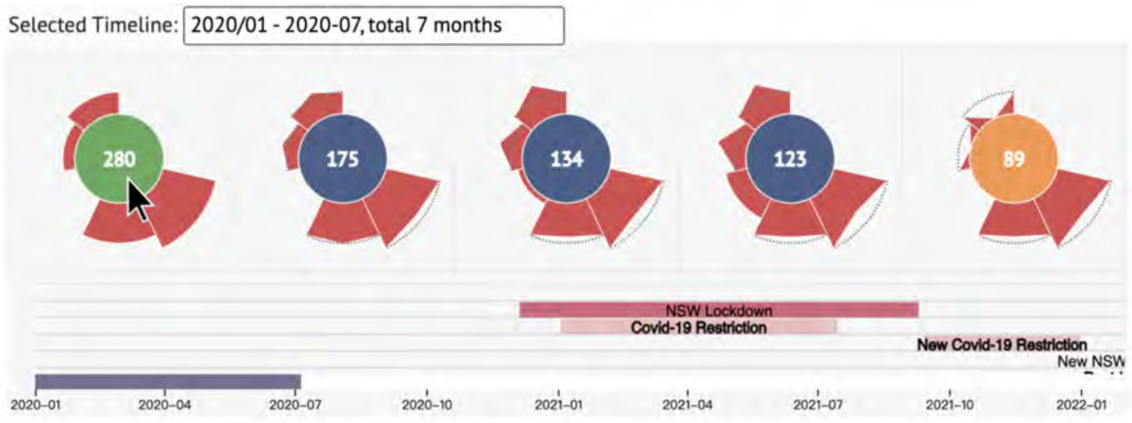

Selected Timeline: 2020/01 - 2020-07, total 7 months

Figure 5.7: The matrix-driven case of exploring multiple visualization units before and after government interventions. Picked in order from the top 1 to 5 of the total infection cases in the zooming scale of the Greater Sydney Districts with the time phase from January to July 2020: auto-clustered visualization unit of Woollahra and Waverley, followed by 3 single visualization units, LGA Syndey, Canterbury-Bankstown, and Cumberland, and a custom-merged visualization unit of Northern Sydney and Mosman.

Woollahra areas that were severely affected by the COVID-19 outbreaks in the early phases were not automatically clustered together, so at that time, we used the merge function to merge them as a new visualization unit. In the current storage sequence, we found that the custom-merged visualization unit (Waverley and Woollahra) did not appear in the top five, and we clicked to select the top five visualization units for a detailed comparison. We hovered the mouse over each petal of the first Canterbury-Bankstown, and the petals of the other visualization units (Cumberland, Sydney, Bayside, Randwick, and Parramatta LGAs) were highlighted correspondingly. From June 2021 to January 2022, the comparison results represent the difference in the number of cases each month for the five selected visualization units.

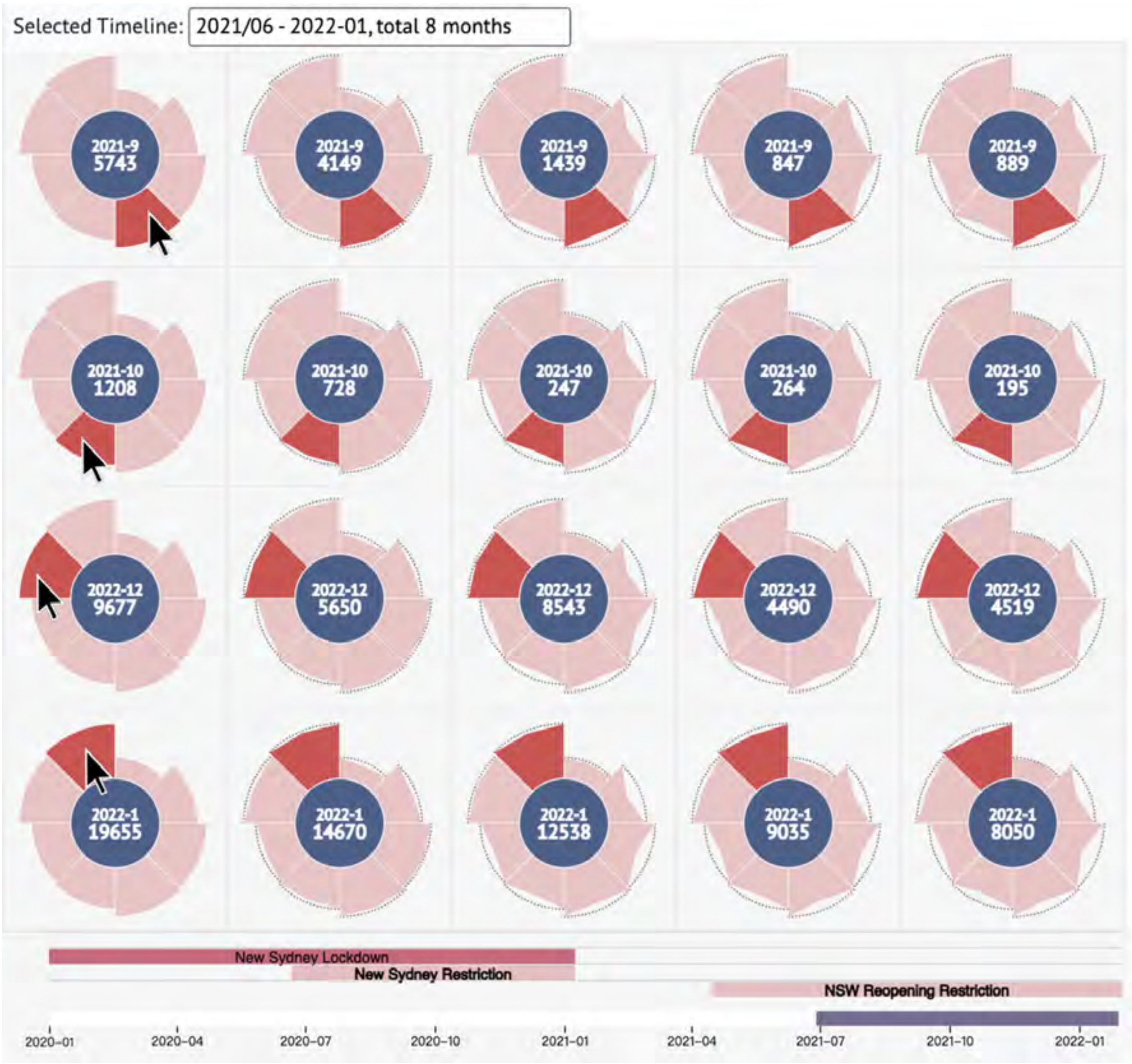
The results in Figure 5.8 show that Canterbury-Bankstown had the highest number of covid cases for each of these 8 months. Although in some months, the ranking of other visualization unit cases is not followed by the total ranking order from the total number of infections, for example, Sydney (with a total ranking of 3rd) has 8543 confirmed cases in December, which is much larger than Cumberland's 5650 (total ranking 2nd), but overall, the ranking order of this selected phase is to a certain extent the same.

In line with its intervention measures, the NSW government announced the third closure of the Greater Sydney Districts in June 2021 to deal with the mutated COVID-19 virus, Omicron. During the lockdown phase, it can be seen that in the first half of the lockdown, that is, from June to September, the number of confirmed cases of the epidemic was still increasing gradually, and there was no downward trend until October, so the government ended the lockdown at the end of October. The epidemic was relatively stable in November after the lockdown ended, but during the Christmas holiday season in December, it increased and reached a peak in January 2022, with three single units in a single month, which is more than 12,000 cases.

5.7 Feedback from Domain Experts

We deployed the visual analytic system based on *UcVE* online and discussed our case study findings with two anonymous domain experts from the Australian government in different departments.

After learning about our methodology and exploration strategies, both experts expressed their enthusiasm for using *UcVE* to analyze the spread of COVID-19 in NSW. We asked the experts to explore on their own using our system, answering their questions and recording their preferences and feedback. By redoing our case studies, they first



Selected Timeline: 2021/06 - 2022-01, total 8 months

Figure 5.8: The matrix-driven case of multiple visualization units during government interventions. Picked in order from the top 1 to 5 of the total infection cases in the zooming scale of the Greater Sydney Districts with the time phase from June 2021 to January 2022. The selected visualization units represented LGAs from left to right are Canterbury-Bankstown, Cumberland, Sydney, Bayside, Randwick, and Parramatta.

affirmed our findings in the different cases before delving into various explorations based on their individual interests in the COVID-19 dataset. They mentioned that the *UcVE*-based visual analysis system was responsive and that the storage function was flexible enough to give them enough analytical confidence. They all appreciated our visual outputs in the universal state that make it possible to connect spatiotemporal to general analysis with rich adjustable parameters and grant possibilities on some very aggregate predictor questions. The expert from the NSW Health Department commented that the system not only displays the full range of COVID-19 with its geographic and timeline features, but it also presents the data in a clear and interactive visualization output. The other expert from the DAC expressed the view that the comparison function of *UcVE* was initially difficult to understand. However, after learning, it was possible to actively compare the distribution of any one or more LGAs within NSW during the past two years. Three experts appreciated *UcVE*'s visual design and anticipated the future benefits of this technology, which allows for precise time comparisons over a single day or week.

5.8 Limitation of UcVE

Our research is generally limited by two aspects. One being is that the spatiotemporal features may come from various data sources among different analytical tasks. For example, the COVID-19 datasets in NSW were applied to *UcVE*, which includes diverse data sources from official releases, social media, and unstructured textual data sets obtained by the researchers of this chapter, which constrained our work. Manually integrating these datasets may produce inaccurate results. Relevant COVID-19 case data was collected by the NSW government based on case locations, yet some records failed to specify the locations where the infections were acquired. Moreover, although the COVID-19 case data released by the NSW government contains open-sourced postal areas and LGAs in the geographic information, we found that there are a lot of errors and null values in the postal area data, which prevents us from using a more detailed geo-scale to explore their locations.

Secondly, our work is also constrained by parameter settings. Our *UcVE*-based visual analytics system delivers strong scalability for user-centered analysis. However, it requires prior learning of how to use the system and how to set relevant parameters, which may incur additional learning costs. For instance, setting a proper clustering radius may make exploration easier; the size of the parameters of the visual cues in the storage and comparison matrix must be set by users based on their screen resolution; the parameters for calculating the petal height in *UcVE* need to be set based on data magnitude for a better display.

5.9 Conclusion of UcVE

In this chapter, we present *UcVE*, a user-centered visual explorer for in-process exploring and comparing spatiotemporal features by visualization units. The visualization unit uses a geo-based layout to display several visualization units by abstracting the attribute values of time-varying properties into encoded petals that compress at each unit's geo-locations. We adjusted the aggregation purposes to define three distinct visualization unit statuses: single, auto-clustered, and custom-merged, keeping map space efficient and enhancing user exploration capabilities while avoiding overlapping issues. The visualization unit allows users to visualize, save, and track in-process exploration results to reduce user cognitive load. We explain it as two exploration strategies: one preserves users' visualization unit interaction output as steps in the current map perspective for historical and current exploration results callback, and the other allows users to compare the attribute values of temporal properties by the encoded petal outer shape of multiple visualization units at a detailed level.

We further implement a visual analytic system that offers *UcVE* for visual exploration and comparison of COVID-19-related datasets. We abstracted each LGA with a coordinated location and its monthly COVID-19 infection cases to encode on the map and coupled the map with a brushable timeline chart to filter different time phases. Simultaneously, two other charts in the system responded to two exploration strategies for tracking visualization unit blocks and comparing encoded visualization units in the matrix. We used three separate case studies in the COVID-19 dataset to demonstrate the efficiency of visualization units, each driven by a different function and explored the causes of the visual results. We also undertook a review in cooperation with domain experts and prioritized our future work based on their feedback.

CHAPTER 6

ClinicLens: Visual Analytics for Exploring and Optimizing the Testing Capacity of Clinics given Uncertainty

6.1 Research Scope of ClinicLens

Recapping the last three years, a valuable lesson learned from COVID-19 is the vital role of the clinic in responding to the pandemic [155]. During this period, numerous clinics were involved in sampling, testing, and diagnosing the disease in addition to giving vaccinations and triaging hospitalizations [5]. However, too many of the current visual approaches emphasize case trends, such as location-based transmission analysis [19] or studies on human mobility [24]. Few tools concentrate on the clinics themselves, which ignores the impact that the clinics have had in preventing the spread of COVID-19.

These clinic features (which are distinguished from clinical features) include information like the clinic's location, business hours, the medical services provided, and their testing capacities. Clinical features are primarily based on the patient and include information such as sign levels and mortality [226].

"Testing is our window onto the pandemic and how it is spreading. Without testing, we have no way of understanding the pandemic." Quoted statement from Ritchie et al.[177]. As we dig deeper into the testing data, it is clear to see that high-quality and high-quantity testing can have significant economic and social ramifications during a pandemic. For example, during COVID-19, large-scale rapid testing was an essential part of the pandemic management strategy for many countries. Through these programs, cases were diagnosed and treated quickly, and the close contacts of positive cases were isolated to slow down the spread. Mass testing, isolation, and population screening in high-risk areas were all crucial public health measures that both helped to control the outbreak and provided information to guide social restrictions [132]. We also know that accurate and prompt testing can reduce the duration of isolation and quarantine, making it faster to get back to work and resume other activities [140]. However, clinic testing capacities are characterized by underlying uncertainty. For example, the demographics of the area in which the clinic is located may not be fully known, or social restrictions might alter the test numbers. However difficult, it is crucial to investigate and maximize clinic testing capacities in the face of these uncertainties if we are to contain the COVID-19 outbreak.

Accordingly, we developed *ClinicLens*, a visual analytics system that allows domain experts to investigate, forecast, and optimize the testing capacities of clinics. At present, *ClinicLens* has been driven by an aggregated dataset of COVID-19 data pertaining to New South Wales (NSW), Australia. However, we have plans to expand the system to support other states and territories in Australia. We also note that the framework *ClinicLens* is built upon could be used to provide visual analytics for other infectious diseases with a few modifications. The architecture comprises Back-end Engine and Front-end Visualization that not only provides dynamic overviews of COVID-19 trends in NSW but also includes regression models that can forecast testing capacities. Notably, the framework is robust to uncertainty. Overall, the system helps domain experts to explore the features that may affect clinic's testing capacities through a visual interface that provides analytics from multiple perspectives: spatiotemporal, location-based demographics, interventions, service factors, etc. Particularly, with the AI-empowered-oriented interactions, domain experts can investigate feature details across collaborative visual views in steps, analyze the cascading relationships between features as they develop, and iteratively update features as a way to optimize the testing capacity of clinics. Three case studies and expert interviews validate the usefulness of *ClinicLens* as contributing new angles through which to explore both the trends associated with COVID-19 and optimize the testing capacities of clinics. In sum, *ClinicLens* offers the following detailed contributions.

- *ClinicLens* assesses and forecasts clinic testing capacities despite uncertainty in the data through feature modeling methods inside the Back-end Engine.
- This visual analytics system amalgamates the Back-end Engine with Front-end Visualization, enabling users to interactively explore COVID-19 trends across NSW

and optimize the testing capacities of clinics accordingly.

- Three real-life case studies with expert interviews cross-evaluate the usefulness and effectiveness of *ClinicLens*.

6.2 Prior Study

In this section, we present our prior study conducted in consultation with a panel of three domain experts. This past work includes conventional studies on the bottlenecks and gaps in COVID-19 research, a background on the available datasets, the concept of uncertainty as it applies to clinics and testing, and the needs and expectations of a useful and valuable analytics system, as specified by our experts.

6.2.1 Conventional Research Bottlenecks and Gaps

Over the past three years, we have been working closely with three COVID-19 experts. Each has a Ph.D. degree and more than 20 years of experience in their field. Two works for the NSW Government: expert E1 focuses on data science and visualization, while E2 is a policy consultant. Expert E3 works at a clinic.

When talking about the pandemic generally, all our experts stated that it has been a difficult period for people in Australia and for humanity as a whole. They also commented that it is unfortunate that people are still getting COVID-19 and passing away today. Although the overall momentum of the epidemic has been contained, the ever-evolving virus variants still trouble people and professionals. "*We all respect and value the work of* *front-line medical professionals, and I want to do whatever I can to help the community improve things*," said E1. "*We already have made lots of research on CT diagnosis with algorithms that aid in pre-therapy, but the conventional research on COVID-19 data stands still on numbers and simple visualizations*," he added. He also noted that he anticipates gaps and bottlenecks in the existing COVID-19 research simply because the data is multivariate and is rife with uncertainty. Existing methodologies might be helpful for analysis, but they struggle in certain difficult situations.

E1 and E2 shared with us the current visualization interface released on the NSW Government's website. At a base level, the interface provides basic statistics for cases, tests, and deaths over the previous seven days plus overall. It also provides line charts of cases, tests, and vaccine rates as trend visualizations for the whole period. A map application keeps track of cases and test data in geo-distributions over a 30-day timeline. Nevertheless, the systems do not support any interactions of selection, filtering, or storage. Hence, the system cannot really be used to support deep decision-making. "*We need intelligent decision-making aids*," E2 and E3 stated from different perspectives. "*both policymakers and clinic managers need to stand by objective bases before making a decision, such as data simulation and what-if analysis, rather than acting impulsively*." "*That's exactly what visual analytics can provide*." According to E1-E3's thinking, visualization and visual analytics can greatly aid in the overall analysis of COVID-19 data. Hence, good visual analytics should be a positive step forward that should be easily combined with interactivity and machine learning algorithms to improve the overall analysis experience with specific COVID-19 tasks. It is worth noting that clinical data research makes up a sizable portion of E3's work, and he also has experience in hospital settings. This offered us a great transformational path for connecting epidemiological data analysis to facts. He raised concerns about the entire health system being overwhelmed over the past three years ‚Äì a sentiment the other experts shared. "*The current healthcare system does not appear to be sufficiently intelligent or optimized. Clinics operate independently and have limited connections with one another, indicating room for improvement.*," E3 emphasized.

6.2.2 Data Background and Uncertainties

Much information has been provided by the COVID-19 Data Program [3] to improve the NSW Government's coordinated COVID-19 response. Headed by the NSW Data Analytics Center (DAC), this program assembles open-source datasets of COVID-19-related information, such as tests, cases, and clinic information. The datasets are updated weekly and made available to the public on the NSW Government website. However, in the post-COVID era, analysis tasks have changed from straightforward numerical analysis to more intricate aftermath work. Therefore, some of the statistical items in the datasets have been modified given that the epidemic has largely been suppressed. Likely infection sources, for instance, are no longer identified, but geo-based clinic information has been added to help people locate where to obtain a test more quickly.

Our first focus as researchers was on the datasets that are still being continuously updated. These included the tests, cases, and clinic datasets from the NSW Government website. E2 pointed out that government interventions affecting COVID-19 restrictions should also be taken into account: "*Overlapped interventions should have significant impacts on test and case amounts*". Hence, we also crawled news on government interventions, which we combined with the case information on confirmed infections, the test results, their locations, and the notification dates. Aside, the clinics' dataset contained details of the COVID-19 testing and assessment centers, such as their geographic location, services list, and business hours. Although these three datasets appear straightforward and manageable, all domain experts still noted that they are actually complex and filled with uncertainty.

In clarifying these uncertainties, we discovered three main areas of issue: 1) the method used to count the tests; 2) daily test attributions; and 3) each clinic's daily testing numbers.

In terms of the counting methods, our domain experts noted that *"The NSW Government has a specific counting method on tests."* The NSW Government determines counting every negative COVID-19 test on any day separately (i.e. 🧍 ⇒ *n*·🔬 = n). That means an individual's first positive test is counted, along with each negative test they've had on previous days. But, after the first positive test, no additional tests are included (i.e. 🧍 ⇒ *n*· 🔬 = 0). All three experts acknowledged that this form of counting guarantees the authenticity of the tests to at least some extent. Plus, it reduces errors and makes statistical analysis easier.

Second, *"We must realize that not all daily tests summed each day are from the previous day."* The truth is that not all clinics have the ability to test COVID-19 nucleic acid results, and shipping every test from clinics to a specific laboratory for testing delays the release of data. We were reminded that, usually, all the test results would be made public within 3 days, although some may still take up to seven days. Thus, the second uncertainty is that the actual daily tests released are dispersed but should be reported in the data releases of the previous days. Hence, as researchers, we needed to consider how to handle this uncertain consequence when modeling clinic features.

Third, our experts reported that *"we cannot monitor the number of tests or testing capacities provided by each clinic on a daily basis."* because the basic unit under which the testing numbers are released is either a local government area (LGA) or postcode, not each individual clinic, which also leads to uncertainty.

On the whole, the first uncertainty, counting methods, can be managed in the data processing. To resolve the last two uncertainties, dispersed daily test counts and the lack of attribution to particular clinics, the domain experts concurred that the testing capacities for each clinic should be abstracted by dividing them into daily test numbers for each clinic. Then machine learning algorithms could be used to establish models for predicting the daily test count amounts allocating this uncertainty.

6.2.3 Requirement Analysis

We finalized the brainstorming with E1-E3 by gathering their concerns and expectations and summarizing their requirements. We had all settled on visual analytics as a significant theme, and all experts agreed that they expected to receive a visual analytics system that used the continuously updated NSW COVID-19 data. After several rounds of discussion, we concentrated on the following four requirements.

R1. Establish feature modeling for clinic testing capacities given uncertainty. Our approach should first collect and sort the existing data. This would involve untangling the selected features to make them reasonable given the uncertainty in the data. Next, each data feature should be ported into an aggregated dataset. Any data scaling that may affect a clinic's testing capacities would be formulated, and feature models would be established through feature engineering. Last, we would undertake an evaluation to confirm the model's reliability. To accomplish all this, it would be necessary to implement Back-end Engine that included all the above processes as components.

R2. Visualize COVID-19 information and clinic-based test capacities based on multiple features. On the basis of our prior study, the domain experts expected that we would implement Front-end Visualization that could display the time series and location-based trends found in the data, as this is what was required to support detailed and dynamic visualizations of daily test and positive case numbers. Additionally, machine learning algorithms and visual analytics techniques should be used to provide multiple different user-friendly views as the system's outputs. As the foundation for visualizing clinic-based information in the COVID-19 background with numerous features, these views need to cooperatively support exploration navigation and intuitive interactions.

R3. Provide the ability to investigate the impact of possible features on the clinic's testing capacity. The experts also asked for the flexibility to interact with any of the features that could affect testing capacity. "Advanced interacting" might include inspecting each feature's importance, filtering the time periods, selecting the locations to be included, adjusting the clinic features included in the analysis, and saving the results of exploration. They also wanted a system that allowed iterative investigations of these impacts and one that, in addition, offered a well-designed visual view of the ground truths, initial prediction results, and updated prediction results in any given analysis.

R4. Enable strategies for optimizing a clinic's testing capacities. Above all, the experts expected a comprehensive system where the knowledge gained from their iterative investigations could be used to optimize clinics' testing capacity and inform upcoming decisions. Adopting reasonable strategies requires the application of case studies in real-world scenarios to enhance the resilience of clinic testing capacity for sustainable balances and to demonstrate the effectiveness of our system.

6.3 ClinicLens Overview

ClinicLens was developed to meet the above four requirements, which is an interactive visual analytics system based on regression analysis that was designed to explore and optimize clinic testing capacities using data that contains uncertainty. As outlined in Fig. 6.1, the *ClinicLens* pipeline is based on the experts' requirements. It consists of Back-end Engine and Front-end Visualization. The Back-end Engine is composed of *Data Processing Component*, *Feature Modeling Component*, and *Regression Model Component*. The regression model quantifies the data columns to predict the daily testing capacity of each clinic as a fundamental implementation of R1. As the assembled dataset driving *ClinicLens*, we combined the raw COVID-19 test and case information published by the NSW Government with the clinic data, data on the government interventions, and the

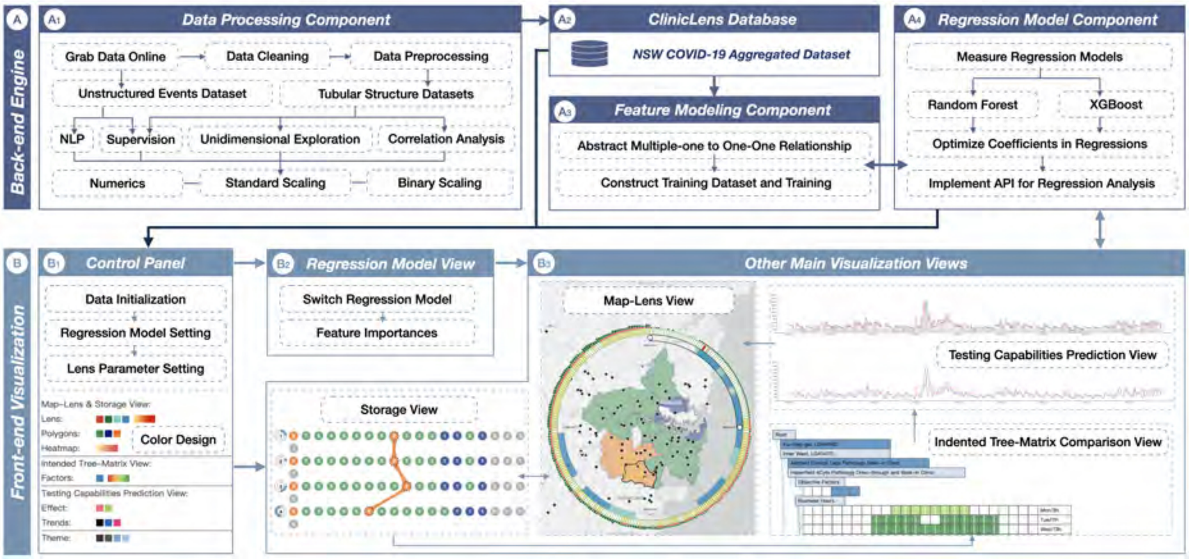

A
Back-end Engine
A1
Data Processing Component
Grab Data Online
Data Cleaning
Data Preprocessing
Unstructured Events Dataset
Tubular Structure Datasets
NLP
Supervision
Unidimensional Exploration
Correlation Analysis
Numerics
Standard Scaling
Binary Scaling
A2
ClinicLens Database
NSW COVID-19 Aggregated Dataset
A3
Feature Modeling Component
Abstract Multiple-one to One-One Relationship
Construct Training Dataset and Training
A4
Regression Model Component
Measure Regression Models
Random Forest
XGBoost
Optimize Coefficients in Regressions
Implement API for Regression Analysis
B
Front-end Visualization
B1
Control Panel
Data Initialization
Regression Model Setting
Lens Parameter Setting
Map–Lens & Storage View:
Lens:
Polygons:
Heatmap:
Color Design
Intended Tree–Matrix View:
Factors:
Testing Capabilities Prediction View:
Effect:
Trends:
Theme:
B2
Regression Model View
Switch Regression Model
Feature Importances
Storage View
B3
Other Main Visualization Views
Map-Lens View
Testing Capabilities Prediction View
Indented Tree-Matrix Comparison View

Figure 6.1: The *ClinicLens* framework. *ClinicLens* consists of (A) Back-end Engine and (B) Front-end Visualization. The entire pipeline of the framework is depicted here, which begins with importing the data in A1 via the *Data Processing Component*. The data processing procedure involves multiple steps but results in an aggregated NSW COVID-19 dataset that contains three different feature types. These data are then stored in a database (A2). In A3, the *Feature Modeling Component* loads the aggregated dataset, abstracts one-to-one relationships, and constructs the training set. The *Regression Model Component* then establishes two appropriate models (RF and XGBoost) for data prediction in A4. The entire Back-end Engine is always ready to be called upon by the Front-end Visualization system (B). A suite of parameters can be adjusted in the Control Panel (B1). Then users are transported to the regression model view (B2). Additionally, four other main visualizations can be rendered by the front-end system for further interpretation, exploration, and interaction (B3).

demographic census data. These datasets were then cleaned, extracted, aligned, and rescaled into usable features that could be used to train the regression models. The Front-end Visualization conjuncts multiple views guided by elaborate colors [190]. Users can interact with these views to explore trends in COVID-19 tests and cases based on location (to meet R2) or clinic and clinic features (to meet R3). Additionally, the system can generate iterative forecasts of testing capacity based on any of the features available in the system (to meet R4).

6.4 Back-end Engine

The Back-end Engine in *ClinicLens* involves an exhaustive data pathway, from accepting the raw data to processing and storing that data in database to finalizing the output of the two trained models. Notably, the Back-end Engine is robust to data uncertainty. The models, based on random forest (RF) and XGboost, are also bridged to the Front-end Visualization to both display the data in a way that can be iteratively analyzed and to offer predictions of clinic testing capacities. The overarching goal of the framework is to predict each clinic's testing capacity using the daily LGA testing numbers as a ground truth. Additionally, each component of the Back-end Engine connects to the front-end system via an API port. Each of the components is described in more detail next.

6.4.1 Data Processing Component

The purpose of the *Data Processing Component* is to load the raw datasets retrieved from Data.NSW, NSW Health, and the Australian Bureau of Statistics (ABS) into the framework. These datasets comprise both structured and unstructured data. In terms of structured data, three datasets contain tabular data. 1) The NSW government COVID-19 tests and cases data, which include the number of daily COVID-19 tests conducted [1] and the number of confirmed cases [2], are released on a weekly basis. These data are classified by LGA and postcode. 2) COVID-19 clinic data. This dataset contains the service details of the authorized clinics [3], hereafter termed the "clinic features". These details describe

[1] NSW COVID-19 tests: https://data.nsw.gov.au/nsw-covid-19-data/tests

[2] NSW COVID-19 cases: https://data.nsw.gov.au/nsw-covid-19-data/cases

[3] COVID-19 clinics: https://data.nsw.gov.au/nsw-covid-19-data/covid-19-clinics

such things as each clinic's business hours, services offered, and testing requirements. Again, these data are classified by LGA and postcode. 3) The 2021 Census dataset, which contains location-based demographic information gathered from the most recent census [4]; the unstructured data pertain to government interventions. These data were crawled and then processed using a natural language processing (NLP) model as part of prior research [233].

To prepare *ClinicLens*, each dataset was next pre-processed to remove all self-testing counts and self-reported cases from the test and case numbers. In tandem with confirming that there were no empty values in any of the datasets, this process aligned the data with the case counts released by the clinics. The COVID-19 tests and case datasets were then combined with the clinic features and the government interventions, yielding an aggregated dataset of *252,350* rows. These data were stored in a database spanning *128* LGAs (*612* postcodes) and *248* clinics and *10* government interventions over a *1030-day* period from 1 January 2020 to 28 October 2022.

Before feature modeling, the aggregated dataset was aligned and rescaled by feature. We were guided by the domain experts, optimizing and selecting the features by assembling and recalculating them in a series of iterative experiments. Through unidimensional exploration and correlation analysis, we ensured that every feature was both reasonable and distinguishable. Thus, our assumptions about which features could affect a clinic's testing capacity were built up over time and guided by expert input. Overall, these features fall into three main categories. 1) Numeric features (X_N), such as LGA

[4] The 2021 Census: https://www.abs.gov.au/census

population densities, daily test, and case numbers, the clinics' business and break hours, location-based clinic counts, and day counts over the whole period. 2) Standard scaling features (X_S), which consist of the current day of the week (1-7), the current season (1-4 represent Spring to Winter), and three levels of government interventions (0-3, where 0 means no intervention and 3 means the strictest interventions). 3) Objective factors of clinic features (X_{Bm}), mostly binary, which comprise factors such as "Referral Required" (0/1 indicating whether a GP referral was required to conduct the test), "Age Limit" (whether the clinic tested infants), "Booking Required" (whether an online booking was required before testing), "Walk-in Allowed" and "Drive-through Allowed", and "Wheelchair Accessible", which are all self-explanatory.

6.4.2 Feature Modeling Component

The *Feature Modeling Component* begins with abstracting the aggregated dataset and preparing the training dataset, all for the purpose of predicting the testing capacities of each clinic. To illustrate the process as simple as possible, we used LGAs as our basic unit; however, postcodes could also be used as an alternative unit.

Compared to the one-clinic-one-LGA relationship (one-to-one), a multiple-clinics-to-one-LGA relationship (multiple-to-one) will introduce uncertainty into the clinics' daily test counts. As previously discussed, we did not have information on the daily test counts of each clinic to directly support multiple-to-one relations, so we redefined a new one-to-one relationship for each clinic in each multiple-to-one relationship for the purposes of training. Specifically, we reconstructed a new one-one relationship for each multiple-one relationship to contain all clinics in a new LGA-based entity by rescaling their clinic features to LGA-based wholes so that they could correspond to the daily released test amounts on each LGA as ground truth. Hence, each clinic binary scaling feature and clinic count in multiple-one are rescaled, to sum up as a whole feature: $X_{B(m,n)} = \{\sum_0^n x_1, \sum_0^n x_2, \ldots, \sum_0^n x_m, n\}$, where m represents the number of binary factors and n stands for the number of clinics in this multiple-one relation. Other features are automatically transcribed to the previous value based on LGAs, which aligns the ground truth of daily test amounts with the LGAs of a single clinic. Thus, the assumption of aiming regression training in each LGA is defined as $Y_{LGA} = \{X_N, X_S, X_{B(m,n)}\}$.

Table 6.1: Regression Model Performance

Models	$RMSE$	$MAPE$	R^2
Linear	770.65	141.86	0.27
GBDT	135.20	81.50	0.97
XGBoost	**69.90**	**66.61**	**0.99**
CatBoost	197.16	85.95	0.95
ExtraTree	379.57	73.16	0.82
LightGBM	184.45	79.88	0.96
DecsionTree	337.89	73.98	0.86
RandomForest	**59.62**	**70.23**	**0.99**

6.4.3 Regression Model Component

As COVID data involve time-series information, regression models serve as the common choices for forecasting the testing capacities [36]. We thus tested various models, including linear regressor, GBDT, XGBoost, CatBoost, ExtraTree, LightGBM, Decision Tree, and RF on our constructed training dataset. The results are shown in Table 6.1, which are evaluated by $RMSE$, $MAPE$, and R^2. Of all models tested, the comparison result suggests that both RF and XGBoost are the most appropriate regression models for this analysis task because they have representative performance on predicted accuracy and feature importance. We determined to employ these two regression models and applied them to the aggregated dataset to predict the possible testing amount for each clinic per day, defined as y_{clinic}. The regression estimates the daily clinic test by a regressor f like: $y_{clinic} = f(\{x_N, x_S, x_{B(m,1)}\})$. y_{clinic} is evaluated per the ground truth of Y_{LGA}, as $Y_{LGA} \approx \sum_1^n y_{clinic}$. We adjusted and optimized the coefficients to ensure the outputs are reasonable, such as ensuring the test amounts not to be negative and keeping two decimal digits.

6.5 Front-end Visualization

We introduce the Front-end Visualization of *ClinicLens*, which enables iterative closed-loop explorations of clinic testing capacity across multiple collaborative visual views. The in-process visual exploration is followed by requirements and started from inspiration by the visual metaphor "Lens on Map [128]", concurrently implementing other views and interactions that dynamically bridge the connection to the Back-end Engine.

In addition to the four main visualization views, there are two supplemental views: the Control Panel and the Model Features View. The Control Panel allows users to select particular datasets, filter timelines, nominate which unit to use, select a preset set of features or all features to include in the regression, adjust the lens size, and change the color representations. The Model Features View shows the RF and XGBoost switches, along with the importance attached to each feature to help determine which model is most appropriate for the user.

Turning to the four main views, the *Map-Lens View* offers options for navigating, inspecting trends in the pandemic, selecting LGAs, and a heatmap of the clinics' testing capacities given the current settings. The Storage View is where users can save their exploration results. The results of explorations can be saved at each iteration in the form of ranking or tracking LGAs by their testing capacities. The *Indented Tree-Matrix Comparison View* is designed to help users interactively inspect, compare, and configure the clinic features to generate testing capacity predictions, while the *Testing Capacities Prediction View* depicts the ground truth trends before and after any predictions are made. More detail on each of these views is offered next.

6.5.1 Map-Lens View

The *Map-Lens View* is designed to serve as a trigger that instructs users to begin their exploration. As such, this view offers a variety of buttons that allow interactivity, as illustrated in Fig. 6.2. A well-designed lens, consisting of three nested layers, is attached to the map's scope once a size parameter has been selected in the Control Panel. We specify these three nested layers to proceed clockwise, starting from the vertical and stopping at the circular end.

Inner: The intervention timeline layer. The inner layer assembles the features of the time series and the interventions over the selected period, including the level of intervention from 0-3. A two-sided draggable slider is available, which can be used to filter the timeline by day, and two colors (turquoise blue 🟦 for eased events and saxe blue 🟦 for restrictions) to distinguish 0-3 level-divided intervention events, where null means no event and the superimposed two colors represent the strictest intervention.

Outer: The test and case count layer. The outer layer expands the space to evenly distribute and radially express the daily test and case counts based on the chosen time frame. To balance space utilization and data range, the radial height V of both test and case amounts H is considered to combine piece functions that are equal to the same calculation: $H_{(day,tests,cases)} \propto a \cdot \ln V(tests,cases) + \frac{V(tests,cases)}{b}$. $V(tests,cases)$ means the amount number of either cases or tests, a and b are adjustable parameters used to achieve a better screen fit. Different representations were painted, where true green 🟩 stands for tests and true red 🟥 represents the cases.

Middle: The positive case rate layer. The daily rate of positive cases helps users to better understand the relationship between testing capacities and cases. It is essentially a highlighted color scale (from canary yellow to scarlet red 🟧) to map the daily COVID-19 severity of LGAs in the current Lens scope. The positive case rate for any given day is calculated by $R_{day} = \frac{V_{cases}}{V_{tests}}$.

In addition to basic navigation tools, the *Map-Lens View* likewise offers some buttons for users to interact with the map. Users can draw polygons to select LGAs or use nearby buttons to toggle each map layer on or off. The map's scope is made up of three superimposable layers: a base layer with pins representing clinic locations, a status polygons layer, and a heatmap layer of testing capacities. The status polygons and the heatmap of average testing capacities interact simultaneously with lens pans and zooms of the map.

The map base layer is used for locating each clinic in NSW. Each clinic is included in an LGA. The basic setting of the status polygons layer is based on the number of clinics in an LGA. To make it simpler for users to visually recognize multiple-one or one-one relationships, we encoded the sea green 🟩 for LGAs with multiple clinics and the wisteria purple 🟦 for an LGA with a single clinic. In addition, we commit to utilizing tangerine yellow 🟧 for user-selected LGAs through interactions. The testing capacities layer shows the average testing capacities of each clinic on a heatmap scale from moon yellow to strong red ▬. The average testing capacity of a specific clinic T_{clinic} is calculated by: $T_{clinic} = \frac{\sum_{d_1}^{d_2} y_{clinic}}{d_2 - d_1}$, where d_2 and d_1 defined as time period from d_1 to d_2.

6.5.2 Storage View

User-centered design is being incorporated into an increasing number of visual analytics systems. In *ClinicLens*, we employ a user-centered design that is similar to our previous design in *UcVE* [233] to support the ability to save the results of users' explorations for later recall. The *Storage View* receives the save command and further compresses and lists the LGAs that *Map-Lens View* had selected for sequential allocation, according to the exploration results on *Map-Lens View*. With the exception of the first circle in each sequence, which is called the location-unit circle and shows the sequence number and the selected period in *Map-Lens View*, each LGA queried in the sequence is represented by an LGA circle with the same colors as the LGA on the map. The precise clinic numbers in each LGA, represented by the number-in-center LGA circles, are queued in descending

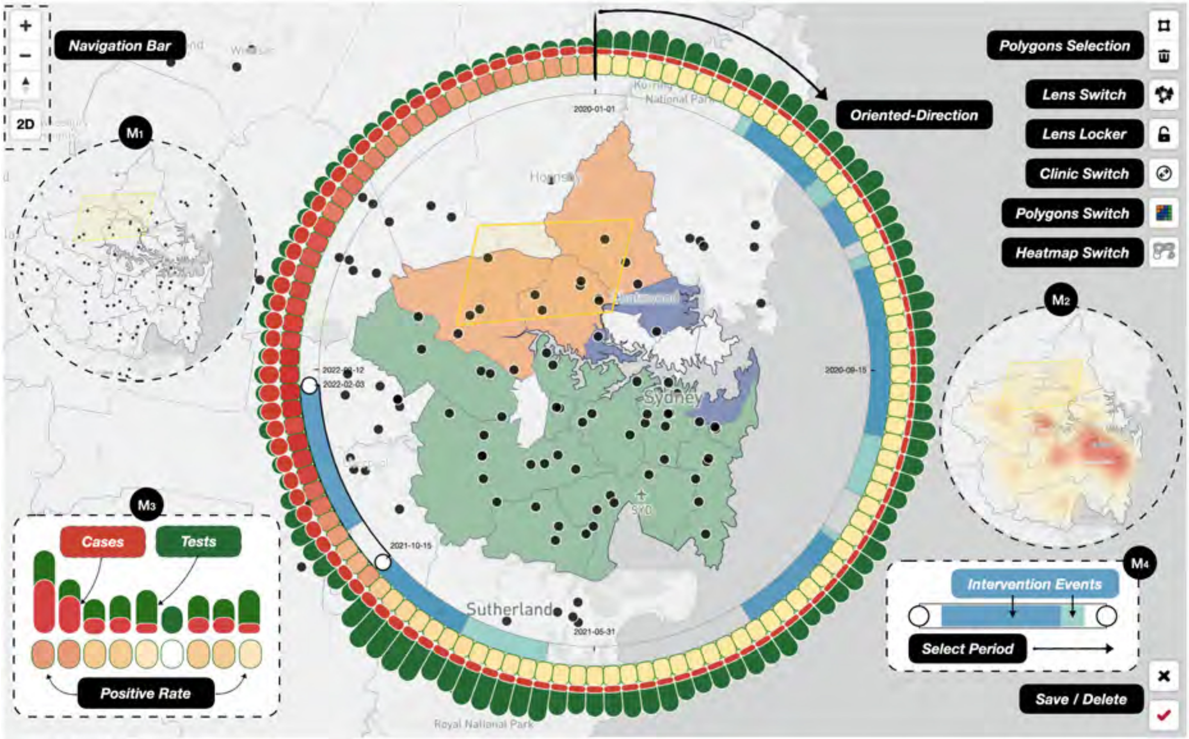


Figure 6.2: The Map-Lens comprises three nested layers: (1) The intervention timeline layer (M4), which encodes interventions and is adjustable through draggable sliders; (2) The test and case count layer (M3), which are displayed through color bars; and (3) the positive case rate layer (M3), rendered as a heatmap. The four buttons (Lens Switch, Clinic Switch, Polygons Switch, and Heatmap Switch) control the lens and the three superimposable map layers. These three layers are the base layer (M1), the status polygons layer, and the heatmap layer (M2). The lens can also be locked in place on the active scope using the Lens Locker button. When finished exploring the *Map-Lens View*, users can preserve the results with the Save button or Delete the callback from the Storage View.

order based on the total number of tests in the sequence.

As depicted in Fig. 6.1, the *Storage View* additionally allows for the reordering, tracking, and highlighting of every sequence with associated links if they point to the same LGA across different sequences (e.g., LGA Canterbury-Bankstown is tracked rankings by link connections among sequences). The situations of all LGA circles in each historical or current storage sequence can also be updated on the *Map-Lens View* as callbacks.

6.5.3 Indented Tree-Matrix Comparison View

Inspired by the hierarchical confusion matrix in Neo [74], we created and implemented an indented tree-matrix structure of the clinics' features by LGA. Abstracting a tree structure, as shown in Fig.6.3, the hierarchy progressively represents LGAs, clinics, and features by level.

The matrix on each leaf node shows the specific features of the clinics, while the hierarchical structure in Fig.6.3(a) illustrates the progressive relationship between the LGAs, clinics, and features. Subtrees can be expanded or collapsed, and features can be revised by clicking. For a more comprehensive analysis, users can add LGAs to the tree-matrix by clicking the LGA circles in the Storage View.

Additionally, each clinic's leaf-fused features can be split into two blocks, as shown in Fig.6.3(b). One block is a 1 × 6 vector of blocks representing the clinics' objective factors. (From left to right, these are Referral Required, Age Limit, Booking Required, Walk-in Allowed, Drive-through Allowed, and Wheelchair Accessible). Salvia blue 🟦 is used to distinguish whether the (binary) value is "Yes" or "No"). The other block is a 7 × 48 matrix of blocks detailing the business hours across a week from 0:00 to 24:00 for each day of the week from Monday to Sunday. Here, a scale of red to green indicates from few to many hours, with each block in the matrix representing half an hour as the basic unit to allow for precise adjustments. Users can compare and edit the various features of the adjacent vectors or matrices among the clinics using the *Indented Tree-Matrix Comparison View*.

The prediction button queries the Back-end Engine for updated features and performs regression analyses of the clinics' testing capacities.

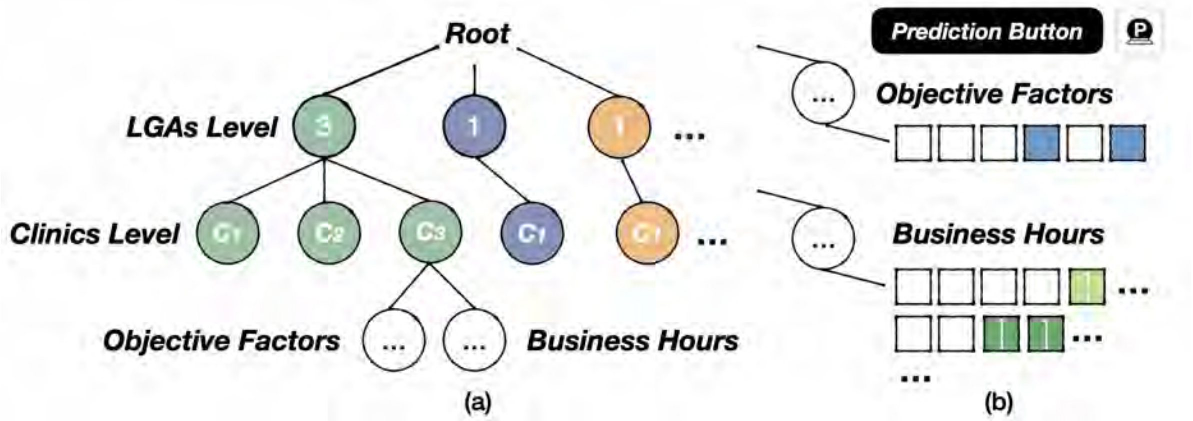


Figure 6.3: The *Indented Tree-Matrix Comparison View* combines the benefits of a tree structure and matrices to interactively expand the hierarchy and balance any adjustments to the objective factors and the business hours. Here, (a) shows the entire structure of the Tree-Matrix View, while (b) illustrates how a unified vector and matrix represent both the features of the objective factors and the business hours.

6.5.4 Testing Capacities Prediction View

The *Testing Capacities Prediction View* relies on two regression models. These models provide users with insights to help interpret specific trends and correlation variations in the testing capacities of each clinic in a given LGA over the chosen time period. This view offers two modes of inspection across four views, delineated by five colors, as shown in Fig.6.4. The prediction results can be saved as a graph.

The two modes of inspection: 1) a step-line chart that conveys the ground truth against the predicted daily testing capacity for each clinic (along with its possible variations); and 2) a curve-line chart that shows smooth trends and discrepancies between the ground truths and the predictions.

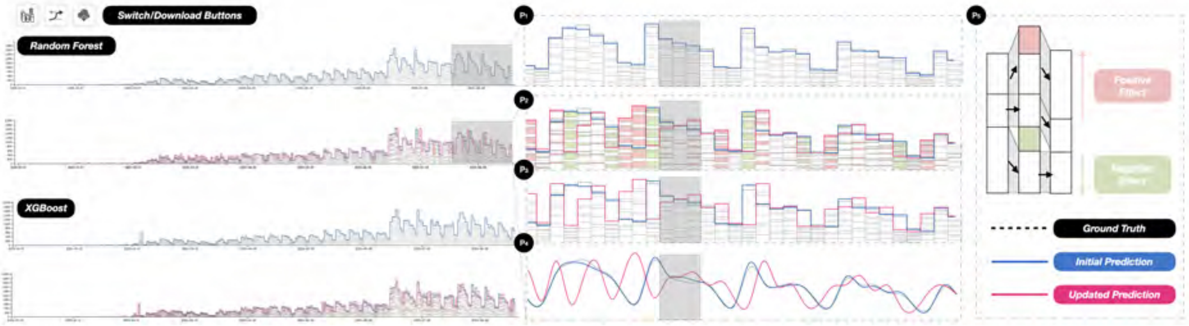


Figure 6.4: The *Testing Capacities Prediction View* provides two alternative regression models, RF and XGBoost, which are each based on different feature importances. The four statuses (P1-P4) are based on different data inspection angles and are controlled by corresponding buttons. These can be summarized into two modes: (P1-P3) Step-line mode and (P4) Curve-line mode. (P5) shows the five colors, which designate: positive and negative effects, the ground truths, the initial predictions, and the updated predictions for clinic testing capacities by LGA.

The four switchable views: Users can switch between the four views via buttons with each offering a different data inspection angle. The first view, the *Testing Capacities Prediction View*, shows the predicted test counts for each clinic as bars under a daily total cap versus actual testing capacities (check P1 of Fig.6.4). The Second is a view of the updated predictions. This view displays the forecast test capacities for each clinic after revisions to their features have been made. Total capacities are shown as step-line charts for negative effects, and these are connected to step-line charts of the original predictions and the ground truth (see P2). The third view shows the updated predictions of the test counts for each clinic based on any changes made to the features. This view also includes step-line charts of the original predictions and the ground truths without positive and negative effect bars, as shown in P3. Last are the three curve-line charts representing the ground truth, the initial predictions, and the updated predictions, as depicted in P4.

The five colors: These colors need to be reasonable and distinguished by other colors used, we borrow true black ■ the dashed lines representing the ground truth trends; ultramarine blue 🟦 for the solid lines representing the initial predictions; fuchsia red 🟥 for the solid lines representing the updated predictions; fuchsia red 🟥 for the positive effects of any changes made to the clinic features; and lime green 🟩 for the negative effects of any changes made to the features.

6.6 Evaluation

To evaluate ClincLens, we conducted three case studies, after which we gathered and synthesized feedback through interviews with domain experts.

6.6.1 Case Study I: Overview of COVID-19 Broad Trends and Regional Clinic Average Testing Capacities

We first conducted a general exploration of broad COVID-19 trends across all of NSW using *ClinicLens*. After initializing the parameters to include the entire period from 1 January 2020 to 28 October 2022, we drove the lens to include all LGAs. The broad trends returned are shown in Fig. 6.5(A).

Overall, we observed the following:

1) Generally, the daily test counts had a periodic 7-day trend, except for two irregular patterns: a) a pulse at the beginning of the third wave of the Alpha and Beta variants; and b) a peak and a sustained high level of testing beginning in the second half of

2021. Then, from the middle of the NSW Government's final intervention, testing counts decreased. These were the reopening restrictions put in place between 8 November 2021 and 31 January 2022.

2) In the early days of the outbreak, daily case numbers were generally low. There were even scattered days when there were no new cases. However, the situation started deteriorating in 2022, peaking at *20.9k+* cases on 6 January. The spreading of the Omicron and other multiple subvariants led to a sharp increase in case numbers, which maintained high until the end of the study period.

3) Because logarithmic mappings of the test and case numbers are not intuitive at small values, we instead examined the daily positive case rates. First, we divided the entire time period into five segments, each with its own specific COVID-19 background. Phase 1 covered the period of the outbreak before any restrictions were put in place, this being from 1 Jan 2020 to 24 Mar 2020. Positive cases were reported on January 25 and 27, 2020, followed by an approximately one-month absence of cases before COVID-19 finally broke out in force at the beginning of March. After reaching a peak of nearly *1,000* cases per day in late March, the government decided to implement a limited intervention. Phase 2 covers a period of three main restriction programs enacted between 25 March 2020 and 26 February 2021. These interventions successfully suppressed the spread, each within a month, giving rise to a 10-month period of diagnosis rates under *0.2%* against three waves of the virus. Phase 3 spans a flat recovery period with no interventions from 27 February 2021 to 24 June 2021. During this period, NSW's clinics conducted over *10,000* tests per day. Generally, positive case rates stayed under *0.1%*, except for one day

on 4 March 2021, when it reached *0.15%*. In Phase 4, from 25 June 2021 to 31 January 2022, the Delta and Omicron variants successively overtook all measures to stop them. In this period, the testing capacities of all clinics were overwhelmed to full loads. After two temporary bursts of spread around the Christmas and New Year holidays, positive case rates crossed the *40%* mark and, unsurprisingly, continued to rise to a top rate of *58.16%*. By 8 January 2022, case numbers had reached *149,033* from *256,229* tests. The final phase covers the period from 1 February 2022 to 28 October 2022, where multiple subvariants coexisted in the population. During this period, the clinics' test capacities gradually restored to levels allowing over *10k* tests per day. However, given the voracity of the subvariants, positive case rates stayed high. Fortunately, none exceeded *40%*, and the majority were below *25%*.

We then explored the clinic distribution across NSW, observing that clinics had been established based on populated areas. For example, most of the LGAs near the coast had more than one clinic with fairly concentrated clinic densities. By contrast, LGAs in the more inland areas had few authorized clinics. For example, Broken Hill, a regional service town, had only two clinics to serve the entire Far West Region ‚Äì an area of around *95,000* km^2. This observation gave us some insight into the huge differences in test counts between the coastal and inland regions. Hence, we used the heatmap to decompose the major trends around the Greater Sydney Area by the five identified phases of the pandemic, as shown in Fig.6.5(B). These heatmap results again confirmed our findings of the broad changes in periodic testing capacities across the clinics in the Greater Sydney Area.

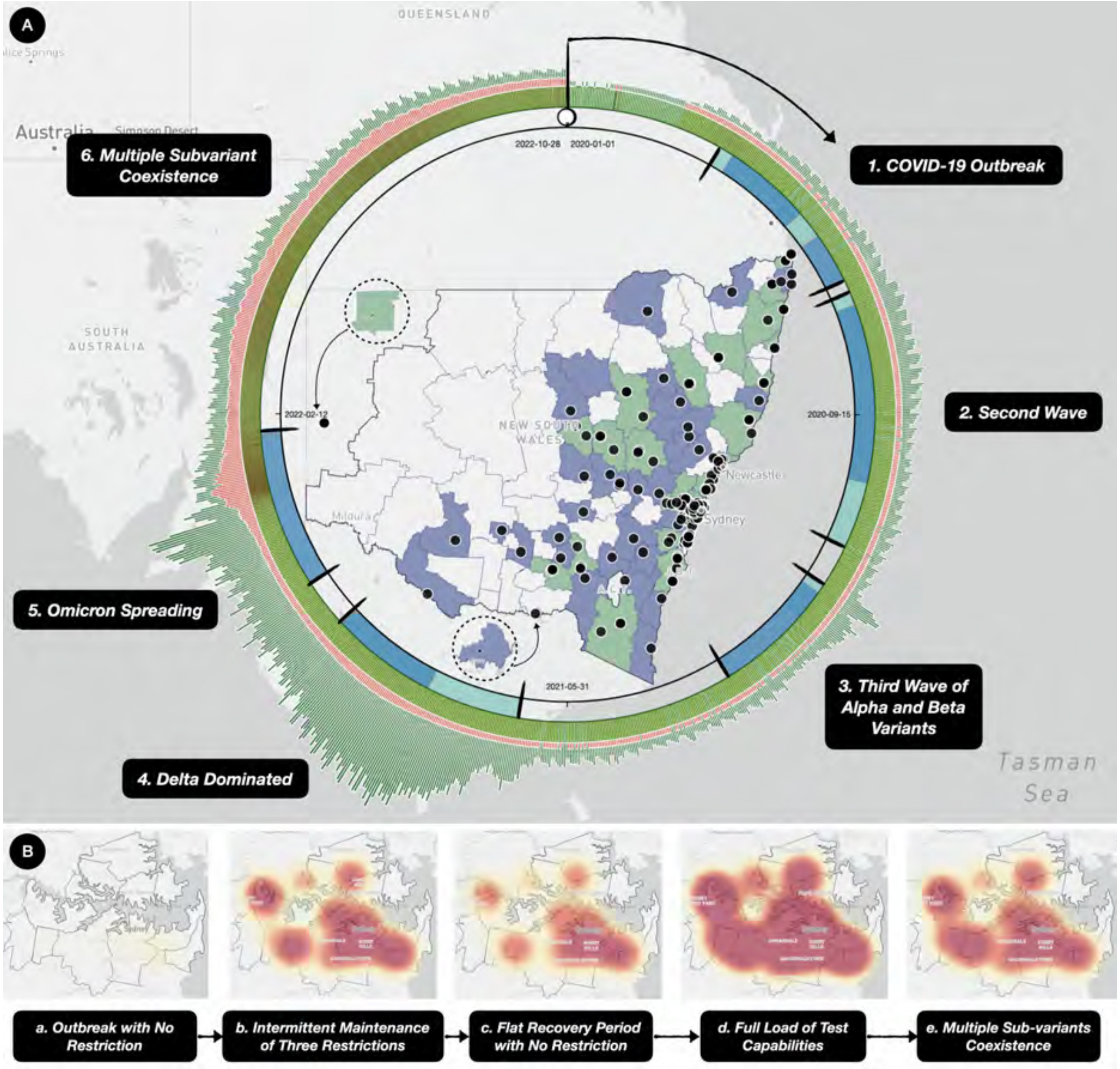

A
QUEENSLAND
Australia
Simpson Desert
SOUTH AUSTRALIA
NEW SOUTH WALES
Newcastle
Sydney
Mildura
Tasman Sea
2022-10-28
2020-01-01
2020-09-15
2021-05-31
2022-02-12
1. COVID-19 Outbreak
2. Second Wave
3. Third Wave of Alpha and Beta Variants
4. Delta Dominated
5. Omicron Spreading
6. Multiple Subvariant Coexistence
B
a. Outbreak with No Restriction
b. Intermittent Maintenance of Three Restrictions
c. Flat Recovery Period with No Restriction
d. Full Load of Test Capabilities
e. Multiple Sub-variants Coexistence

Figure 6.5: Case Study I: (A) The COVID-19 broad trends in NSW from 1 January 2020 to 28 October 2022, where intervention events in the timeline represented government intervention events against the virus variants in varied periods and the scatter points represented the varied distribution of clinics. (B) A five-piecewise periodic heatmap chain of the average testing capacities of clinics in the Greater Sydney Area.

6.6.2 Case Study II: Investigating the Impact of Clinic Features on Testing Capacities

In this case study, we explored how the clinic features influence testing capacities. As discovered in Case Study I, most clinics were operating at full capacity from 25 June 2021 to 31 January 2022. Hence, we began our investigation by adjusting the map lens to this period.

The LGA of Sydney, which includes the central business district and close surrounds, is a significant place in the Greater Sydney Area. It is an artery of communication to northern Sydney, the southern airport, the eastern coastal beaches, and the western residential areas. Zooming into the Greater Sydney Area, as shown in Fig. 6.6(A), we drew a polygon to highlight the Sydney LGA and saved the current navigation results to the Storage View. Seven clinics are located in this LGA (see Fig. 6.6(B)), ranking fourth in the total number of tests performed during this time period. Clicking on the LGA Sydney circle, we then browsed all the clinics' features in the *Indented Tree-Matrix Comparison View*. Here, we discovered that all clinics allowed walk-ins and all were wheelchair accessible. Only one clinic required a booking before tests. Five out of the seven clinics were open seven days a week and for more than eight business hours per day. The other two only operated on weekdays and only for 4.5 hours per day and 8 hours per day, respectively, as indicated in Fig. 6.6(C). Looking to check for the impact of these clinics on Sydney's testing capacities, we delved further into these two subcases.

Exploring the impact of a clinic's business hours. Initially, we attempted to slightly adjust the business hours of each clinic, helped by the RF model. The predictions that followed showed that this feature can indeed affect a region's test capacities. Test numbers are positively correlated with business hours and inversely correlated with break hours. Additionally, over several iterative adjustments to the business hours, we observed that test capacities were not sensitive to sudden increases or decreases in the operating hours on a certain day. Rather, because there is uncertainty in the released figures, the regression model adds together the testing numbers of the previous few days. In addition, we also found that simply changing the business hours of a certain clinic affected the predicted test numbers of other clinics on the same day. In fact, any changes to the business hours on a certain day would cause variations in test counts over a period of 7 days. As an example, to represent this phenomenon, we changed our two weekday-only clinics, Central and 4Cyte, to open on the weekend with the same hours (see Fig. 6.6(D)), while keeping the clinics' other objective factors unchanged. Fig. 6.6(F) shows the resulting predictions.

Within Fig. 6.6(F), Panel F1 shows the daily test numbers for both the ground truth and the predictions for 3 December 2021 to 31 December 2021. Panels F2 and F3 show alternate views of these data. Overall, these results confirm the fact that extending these clinics' business hours to Saturday and Sunday changes the predicted test numbers for each clinic not only on those days but also for the previous few days. Additionally, from the detailed tooltips, we discovered that both these clinics were significantly more affected (positively) in terms of test number predictions than any of the other clinics in the marked zone on the day of 11 December 2021. Generally speaking, opening on the weekend seemed to have a positive impact on the clinics from Thursday to Saturday, most evidently on Saturday, and a negative impact from Monday to Wednesday. However, from a periodic perspective, extending the business hours to the weekend had a positive impact on overall test amounts for the week.

Panels F2 and F3 also reveal some other interesting patterns. For instance, test numbers in the week of 11 December 2021 are different from the following week in that extending the business hours had a greater impact on test numbers this week than the next. The predicted peak was postponed by one day and, even though the clinic's test capacities would have already reached the regional limit, the highest peak of the predicted values versus the actual values are essentially the same. This indicates that the RF model learned from the time series features in the training dataset that the test numbers during the Christmas and New Year holidays were not sensitive to a clinic's business hours. As such, the prediction results for the following week remained fairly stable compared to the ground truth.

How different objective factors affect test capacities. Through multiple progressive explorations, we discovered that only a few clinics supported both walk-ins and drive-through testing. The "Walk-in Allowed" clinics were mostly located in densely populated areas, while the "Drive-through Allowed" clinics were mostly located in remote areas, some of which required booking before tests. Additionally, we learned that test numbers were correlated to both of these factors. In other words, because most clinics were either walk-in or drive-through, not both, different combinations of factors would cause different variations in the testing capacities predicted. Thus, we found that every clinic needed to set a "Yes" to at least one of "Walk-in Allowed" or "Drive-through Allowed", plus a "Yes" to "Referral Required", "Age Limit", "Booking Required", or "Wheelchair Accessible". Not doing this had a negative effect on test counts, and vice versa. Further, we investigated the effects of adding break hours to the clinics' business hours. The results show that adding a few break hours (say, 1-2 hours) has no discernible impact on testing capacities.

However, what did intricately impact the predictions was changing confluences of the clinic features. As an example, Fig.6.6 shows a case where we set "Age limit" and "Requires Booking" for both the Central Clinic and 4Cyte to "Yes" (see Fig.6.6(E)). Additionally, we adjusted the business hours to include an hour break each weekday and extended the Central clinic's weekend hours by 2.5 hours and the 4Cyte clinic's hours by 4 hours each day. The predictions, shown in Fig.6.6(G), are located in the same time period as the previous subcases presented in curve-line mode. Although some trends in both the initial and updated predictions are staggered, there were not too many differences in the overall testing counts for the week because most trends matched properly. However, the pattern shown in Fig.6.6(G1) illustrates the differences in trend over a five-week period from 3 December 2021 to 7 January 2022. Inspecting the values more closely using the tooltips, we discovered the predicted test numbers each day for the first three weeks were generally lower than the initial predictions. This situation can perhaps more clearly be seen from a side-by-side comparison of Fig.6.6(F3) and (G1), where the updated clinic features cause a negative influence that outweighs the benefits of extending the weekend business hours in terms of testing capacity. Moreover, Fig.6.6(G2) shows us that updating

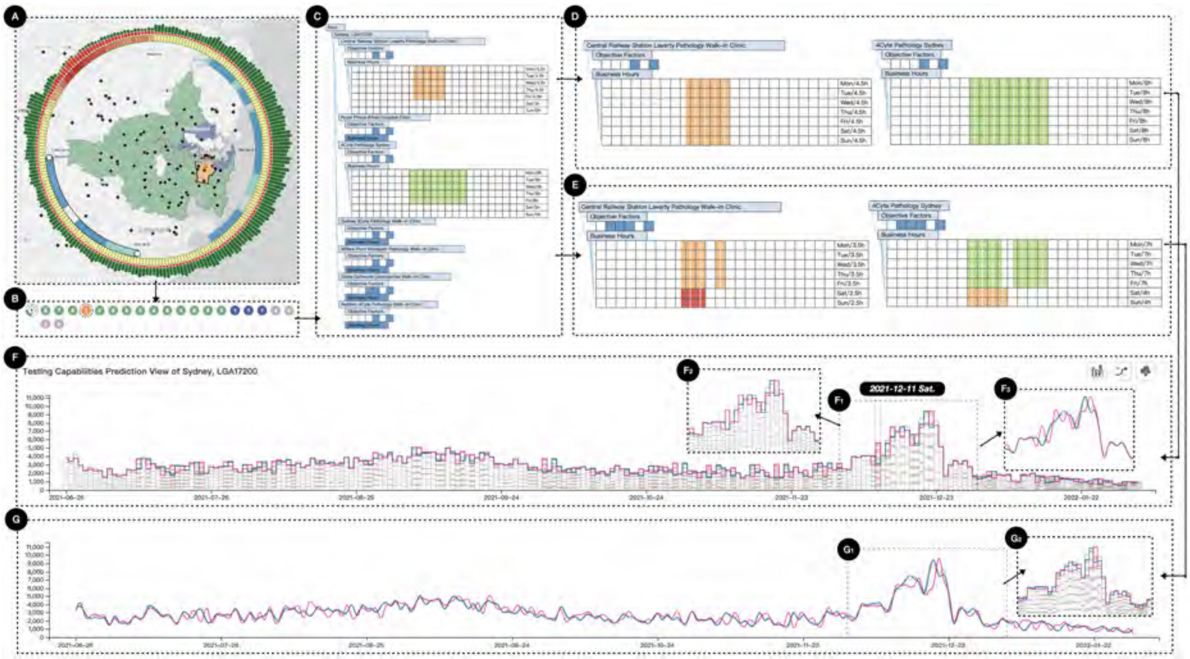


Figure 6.6: Case Study II: (A) Driving Map-Lens to the Greater Sydney Area within a time constraint of 26 June 2021 to 1 February 2022. (B) Saving the exploration result to the Storage View and selecting the LGA Sydney circle to (C) the *Indented Tree-Matrix Comparison View* for further observation. (D)-(F) and (E)-(G) provide two different subcases of predicted testing capacities with updated clinic features.

the features for our two focus clinics, Central and 4Cyte, affected the forecast testing numbers for all clinics in the Sydney LGA on certain days. However, the most major effects were felt in these two clinics. This result is consistent with our findings from the previous case study. We also found it convincing that the predicted test numbers changed in the first three weeks before the holidays and then remained relatively constant for the next two weeks. This indicates that the models are accurately reflecting people's testing behaviors.

6.6.3 Case Study III: Exploring the Optimization Strategy with LGA-based Testing Capacities under Uncertainties

While exploring the aggregated COVID-19 dataset, we realized that making reasonable updates to the clinics' features could reduce the pressure on certain periods of daily test capacities. Additionally, maintaining a similar volume of tests seemed to achieve the most sustainable balance between clinics' testing volume tolerance and efficient testing services. Thus, we sought to explore patterns and optimize the clinics' features to prevent the spread of cases and inform policies for coexisting with the virus in the post-COVID era. We discussed the previous two case studies with our domain experts before undertaking the third case study, and, with their feedback in mind, we set out to find strategies for optimizing the testing capacities of the clinics across one LGA by adjusting their clinic factors. This included considering the geographical location of each clinic, their business hours, and their other service details. The complete process is shown in Fig. 6.7. Our focus points for this case study were three main city scopes in NSW: the LGAs' of Greater Sydney; the Central Coast, a regional area just north of Greater Sydney, and the third most populous region of NSW; and the city of Newcastle, north of the Central Coast, which is NSW's second most populous region (see Fig. 6.7(1) and (2)). Fig. 6.7(3) and (4) show the polygon drawn to highlight the Central Coast LGA. We explored and stored four time periods, as shown in Fig. 6.7(5) before finally opting for the period from 1 February 2022 to 28 October 2022. In this period, there were no government interventions, while positive cases remained consistently high. In fact, the positive rates did not start to decline until the end of August which accordingly reflected on the daily test that counts reflect a similar trend with the average weekly total of tests performed reaching *40k* prior to September and eventually dropping to approximately *20k* after.

As Fig. 6.7(6) shows, the Central Coast LGA and its eight clinics reported the third highest test counts in the state during this period. We first examined the average testing capacities of each clinic in the Central Coast LGA through the heatmap. Next, we selected this LGA and gathered statistics on the features of the clinics from the *Indented Tree-Matrix Comparison View*. Fig. 6.7(7) shows the clinics' features.

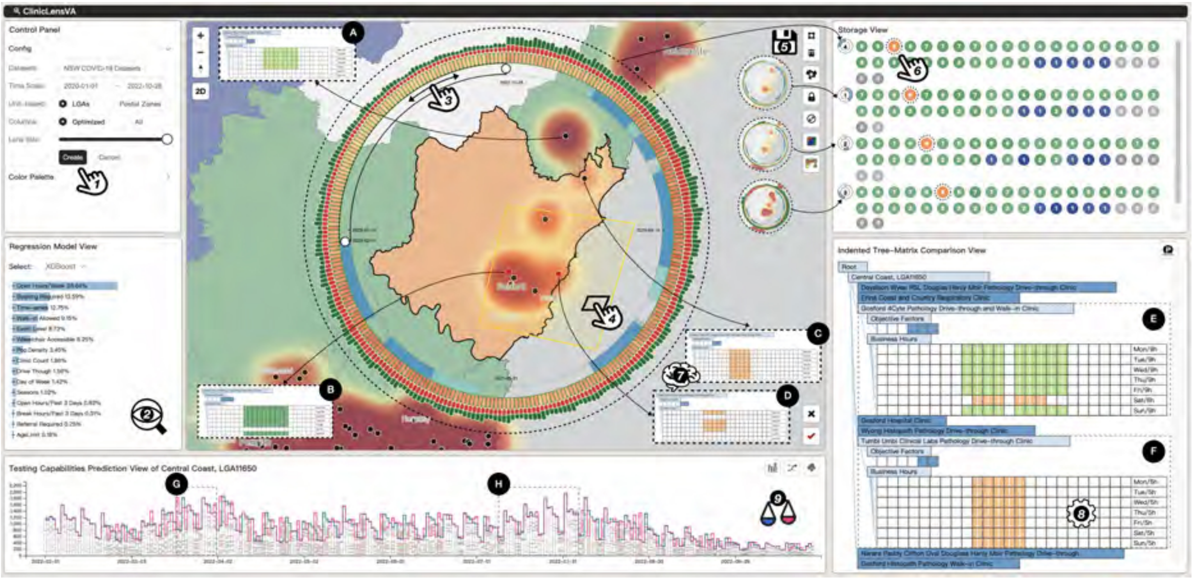
From this Fig.6.7(7), we can see that the Central Coast boasts one "almighty" clinic, namely Gosford 4Cyte Pathology Clinic (B), which is open 10 hours per day and supports both walk-in and drive-through clients. It is wheelchair accessible and does not require referrals or bookings. Further, there is no age limit for the patients tested. The remaining seven clinics in this LGA either support walk-ins or drive-throughs, not both, and most operate for only 8 hours per day. From their locations on the *Map-Lens View*, we observed that the Gosford clinic was grouped with three other adjacent clinics that could conduct tests for each other. The other four were located in relatively independent regions.

Our formulated strategy for optimizing the testing capacities in this LGA is shown in Fig.6.7(8). We made the decision to keep the entire business hours the same by appearing the Saturday hours to six hours and appending one hour of break time on the other days, as found in (E). From the *Map-Lens View*, we observed that the Doyalson clinic (C) was geographically close to the Morisset clinic (A) in another LGA (Lake Macquarie) and that Morriset had stronger average test capacities but did not operate on weekends. Hence, the Doyalson Clinic, to some extent, appeared to be an alternative to Morriset for weekend testing. Both clinics were checked as "Drive-through Allowed" and "Wheelchair Accessible" but, we ruled out setting Doyalson Clinic as Walk-in Allowed by checking its location and surroundings on Google Maps with an abundance of caution. Thus, we did not change any of Doyalson's features. We further discovered that the Trubi Umbi Clinical Lab (D), which was located in a populated area, was open for 5 hours for all weeks but only provided 5 hours on the COVID-19 test on Monday, Wednesday, Thursday, and Friday in our database. After confirming on HealthDirect [5] that the hours for Trubi Lab were not an error in our database, we included updating (D) to (F) as suggestions in the optimization strategy.

Fig. 6.7(9) shows the updated predictions. However, although there appeared to be differences between the initial and the updated predictions, it was clear that roughly half of the predicted daily test numbers were nearly identical for the time period, as shown more clearly in (G). At the same time, we discovered that the initial and updated predictions may be staggered during a period of trend. The total forecast test counts in the updated predictions were slightly higher than the counts for the initial predictions in certain weeks, as shown in (H). In other words, the updated predictions were higher than the initial predictions for most days from 19 March 2022 to 1 August 2022.

Exploring deeper, we hovered over each clinic bar to read the tooltips and identify which clinics contributed the most significant test amounts to the updated predictions

[5] A government-funded service updated clinic information in real-time: https://www.healthdirect.gov.au/



Control Panel

Config

Dataset: NSW COVID-19 Datasets

Time Scale: 2020-01-01 ~ 2022-10-28

Unit-based: ● LGAs ○ Postal Zones

Columns: ● Optimized ○ All

Regression Model View

Select: XGBoost

Open Hours/Week 25.64%
Booking Required 12.59%
Time-series 12.75%
Walk-in Allowed 9.15%
Event Level 8.72%
Wheelchair Accessible 6.25%
Pop Density 3.45%
Clinic Count 1.86%
Drive Though 1.58%
Day of Week 1.42%
Seasons 1.02%
Open Hours/Past 3 Days 0.82%
Break Hours/Past 3 Days 0.31%
Referral Required 0.25%
AgeLimit 0.18%

Testing Capabilities Prediction View of Central Coast, LGA11650

Storage View

Indented Tree-Matrix Comparison View

Root
Central Coast, LGA11650
Objective Factors
Business Hours
Gosford Hospital Clinic
Tumbi Umbi Clinical Labs Pathology Drive-through Clinic

Figure 6.7: Case Study III demonstrates various steps and visualizations conducted using the system. (1) The *Control Panel* shows the time scale, unit, feature columns, and lens size settings. (2) The *Regression Model View* examines the importance of features after focusing on specific city scopes. (3) The *Map-Lens View* allows filtering of time periods. (4) Drawing a polygon highlights a specific area. (5) The *Storage View* saves time period sequences. (6) The *Indented Tree-Matrix Comparison View* focuses on a specific region to compare clinic features. (7) Initial clinic features are compared with (8) updated features. (9) The *Testing Capacities Prediction View* visualizes predictions and compares daily test capacities.

on specific days. The statistics reveal that the Gosford and Tumbi clinics were primarily responsible for the increases. This positive result accords with our strategy of extending Tumbi's business hours to include Tuesdays, Saturdays, and Sundays and increasing Gosford's business hours from no hours to 6 hours on Saturdays. Notably, the one-hour break set at the Gosford clinic each day did not cause any unexpected fluctuations. Overall, this proves that *ClinicLens* can be used to develop a reliable strategy for optimizing testing capacities across an LGA.

6.6.4 Expert Interview

After conducting the case studies, we arranged one-on-one structured interviews with our three domain experts (E1-E3). From a discussion of our findings from the three case studies, we received affirmation from the domain experts that *ClinicLens* archived their initial expectations. We then provided them with tutorials and encouraged them to use *ClinicLens* independently to explore issues of their own interest. Their feedback and qualitative preferences are summarized below.

Reliability of the Back-end Engine. According to all the experts, *ClinicLens* provides productive results when exploring and optimizing clinic testing capacities given uncertainty. All the information included in the Back-end Engine was reasonable and, after selecting the appropriate regression model based on performance, the experts felt confident in performing their analysis tasks. This was, in part, thanks to *ClinicLens*'s assistance in providing the RF and XGBoost models with automatic feature extraction as steady streams to the Front-end Visualization. Additionally, *ClinicLens* also offers all-feature regression, which means that new feature inputs can easily be added in the future.

Effectiveness of the Front-end Visualization. The consensus among all domain experts was that each view in the Front-end Visualization was considered in terms of its visual form. They were all impressed with the *ClinicLens* design. E1 commented on *ClinicLens* overall as being a valuable visual analytics system. He was especially impressed with the visual expressions in the *Testing Capacities Prediction View*. E2 highlighted that *"including a mobile lens on the map is a smart move. In addition to* *representing rich information, such as combining interventions, the Map-Lens View can be persuasively explored using several different interaction techniques and serves as a trigger for other views for collaborative exploration".* E3, on the other hand, particularly admired the *Indented Tree-Matrix Comparison View*. He thought the indentations of the statistics not only effectively conveyed the structure of the data in a constrained area but also allowed the user to adjust the clinic features for prediction in a worry-free manner. Cycling through the user-friendly interactions in each view, the experts agreed that *ClinicLens* had a low learning curve and could offer quick responses to COVID-19-related questions through visual analytics and machine learning techniques that were not just limited to clinic testing capacities. In general, they felt the whole visual pipeline ran smoothly over the COVID-19 aggregated dataset, and the experts thought it could significantly increase the efficiency of their analyses.

6.7 Discussion and Limitation of ClinicLens

The purpose of exploring testing capacities. The first and most important thing to discuss is that we must accept the fact that predictions inevitably contain bias or variance [34]. Any slight change in a clinic's factors might lead to changes in testing capacities. The predictions may vary with different attempts to change the same feature, but few, if any, changes will result in a change of large magnitude. Further, some changes resulted in staggered variations between the initial and updated prediction results. This and the inherent uncertainty in the data are the primary reasons we recapped. We cannot trace the total test numbers announced for a certain day to the actual tests conducted by each clinic. Rather, we can only estimate a clinic's test capacity based on its given features with machine learning models. However, this does offer a good overview from a visual analytics perspective. Further, our aim stretches beyond simply predicting test capacities. We have designed *ClinicLens* to help users optimize the testing capacity of a clinic, a postcode, an LGA, or an entire state by giving them the option to adjust multiple different clinic features.

Optimization Strategies. There is no limit to the number of optimization strategies that can be explored through *ClinicLens*. Many strategies formulated will increase the testing capacity of the selected region. However, considering that there is always bias and uncertainty in predictions, we must experiment iteratively and combine real situations with actual experience to find the most appropriate strategies. Generally speaking, a few solid and preferred strategies should ultimately be reached after observing and closely examining a given situation. Our case studies inspired and motivated the domain experts to devise their own feasible strategies. For example, Case Study II taught them that they could tweak various combinations of clinic factors, including just one clinic's business hours, to create a positive effect on a whole region's testing capacity. Case Study III showed them that ensuring each clinic was open consistently during the same business hours each day also increased testing capacities, whereas adding break hours to days had little impact.

Generalizability and Scalability. The generalizability and scalability of our *ClinicLens* can be considered in terms of three perspectives. First, the aggregated COVID-19 dataset is only an example of the data that could be used to power *ClinicLens*. Our demonstrated visualizations ran perfectly both over subsegments of the data and over the whole time period under study. Second, the architecture of *ClinicLens* is designed to purposefully separate the Back-end Engine from the Front-end Visualization. This means that different models can be added or revised very easily. In addition, the framework provides all-feature regression, which means that *ClinicLens* is flexible. It can easily accommodate future additions of new features or input. Lastly, even though *ClinicLens* in its demonstration form is limited to exploring COVID-19 data, all our domain experts agree that this could be a productive visual analytics system for other epidemiological analyses. With slightly adjusted parameters and by adding a few more visualization views to help analyze spatiotemporal features, *ClinicLens* could be used to analyze clinic testing capacities for other infectious diseases.

Limitations. Our approach remains constrained by certain implicit features of the clinic information. First, some implicit features may cause fluctuations in a clinic's predicted test capacity. For example, uneven population density and the size of each clinic in terms of the number of employees and the quality of the medical services provided may influence a person's preference for where they get tested. Additionally, we did not account for any differences in business hours that may have impacted the prediction results. For example, there would clearly be different forecasts for clinics that were open at 5 am and not 5 pm. We also did not demarcate clinics authorized to conduct rapid testing for international airline departures (where the results are made available within 48 hours) as these features give rise to more complex regression analysis. One feature that the experts requested after using *ClinicLens* was the ability to add or delete clinics from an area and forecast the impact on the other clinics in the region. *ClinicLens* does currently not do this, but it may be a function we add in future work.

6.8 Conclusion of ClinicLens

In this paper, we presented *ClinicLens*, an interactive visual analytics system for use by domain experts to both explore location-based COVID-19 case and test statistics as well as optimize the testing capacities of the clinics in selected locations. Motivated by the current challenges in preparing for infectious disease outbreaks and informed by expert requirements, *ClinicLens* comprises Back-end Engine that is automatically driven by AI to identify and extract features that may affect clinic testing capacities and Front-end Visualization system that offers multiple perspectives on the data to support planning and decision-making. Importantly, the framework is robust to the uncertainty inherent in the available COVID-19 datasets. Three real-world case studies along with expert interviews validate the usefulness and effectiveness of *ClinicLens*. As such, we believe *ClinicLens* offers a fresh perspective on the decision-making surrounding clinic testing capacities for infectious disease management. In the future, we intend to address the current limitations of *ClinicLens* as well as deliver more visual assistance for real-world analysis tasks to domain experts.

CHAPTER 7

CONCLUSION AND FUTURE DIRECTION

This thesis focuses on conducting a series of visual comparative studies on remarkable data features in multivariate data. The study begins with a comprehensive literature review of related works in the areas of methodology-based, multivariate-based, and scenario-based research. Based on the current research problems and gaps, the significance and objectives of the research are concisely outlined. The research pathmap consists of four main approaches in a progressive pipeline, each of which focuses on exploring and comparing different aspects of multivariate data in various domains. The four approaches offer different emphasis on enhancing the understanding and comparison of datasets, including:

7.1 Conclusion and Future Direction of PansyTree

The *PansyTree*, is a novel approach for comparing hierarchical structures and node numerical multi-attributes by merging union trees. In this study, we introduce a unique icon called "Pansy" to represent each merged node in the hierarchical structure. The Pansy icon is characterized by three colors that map data items from three distinct datasets at the same hierarchical position or tree node. The petals and sepals on the Pansy icon are designed to display the values of each attribute and the hierarchical information. Moreover, we redefine the links in force layout encoded by width and animation to improve the communication of hierarchical information. We apply the *PansyTree* to CNCEE datasets and present two use cases to demonstrate its effectiveness.

Although *PansyTree* provides a visualization solution for comparing both hierarchical structures and attribute values by merging three hierarchies, its potential for multiple hierarchy comparisons in various scenarios is worth exploring. In future research and development, we aim to investigate and evaluate the extent to which *PansyTree* can effectively handle merging larger datasets and accommodating a higher number of attribute values. This exploration will involve assessing the scalability and performance of *PansyTree* under different conditions.

Moreover, we will emphasize enhancing the ability to compare attributes and hierarchical structures through visual encoding and interaction techniques. This includes refining the visual encodings used for representing attribute values and finding ways to facilitate intuitive interactions that allow users to explore and analyze the hierarchical data efficiently. By enhancing the visual encoding and interaction capabilities, we aim to provide users with more powerful tools for making insightful comparisons and gaining a deeper understanding of complex datasets.

7.2 Conclusion and Future Direction of +msRNAers

The *+msRNAers*, which takes into account the spatiotemporal features of virus transmission patterns and multidimensional features of objective risk factors in communities, enabling portrait-based exploration and comparison in epidemiological analysis. We applied *+msRNAer* to aggregate COVID-19-related datasets in New South Wales, Australia, combining COVID-19 case numbers, geo-information, intervention events, and expert-supervised risk factors extracted from LGA-based censuses. We demonstrated the *+msRNAer* workflow and evaluated the feasibility, effectiveness, and usefulness by applying aggregated COVID-19-related datasets via one user study and three subject-driven case studies. According to positive feedback from experts, *+msRNAer* provides a general understanding of analyzing comprehension that not only compares relationships between cases in time-varying and risk factors through portraits but also supports navigation in fundamental geographical, timeline, and other factor comparisons. By adopting interactions, experts discovered functional and practical implications for potential patterns of long-standing community factors against the vulnerability faced by the pandemic. Experts confirm that *+msRNAer* is expected to deliver visual modeling benefits with spatiotemporal and multidimensional features in other epidemiological analysis scenarios.

Based on the feedback from participants in the user study and discussions with domain experts, we have targeted four aspects to focus on in our future work of *+msRNAer*.

Firstly, we plan to update our prototype by importing refined COVID-19 case data and new census data later next year, when community datasets are collected during the pandemic.

Secondly, to increase the diversity of community profiles, we will include more variables, such as relationships, ethnic groups mentioned by domain experts, and education level. We will continue to apply machine learning algorithms, such as Principal Components Analysis (PCA) for dimensionality reduction to add more categorical indicators and time series predictions for infectious cases affected by risk factors. We also expect to propose a rating system for all variables with assistance and evaluation from domain experts, which would be able to quantify the characteristics of communities. By designing an appropriate measurement matrix for indicators, we will be able to create index metrics for future use in other epidemiological analysis scenarios. These metrics will not only assist decision-makers in making pandemic prevention measurements but will also educate the public on personal influence, and eventually, how to work together to tackle this great challenge through their own efforts.

Thirdly, with the improved approach, we intend to conduct a systematic review with more government staff. As suggested, we will offer an automated reporting function for storytelling purposes. This function would help disseminate the right messages to the public.

Finally, we will consider applying our prototype to more transmission datasets in epidemiological analyses to validate the scalability of *+msRNAers*.

7.3 Conclusion and Future Direction of UcVE

The *UcVE*, a user-centered visual explorer for progressive comparing multiple visualization units in spatiotemporal space. We create unique unit visualization with the customizable aggregated view based on the visual metaphor of flower bursts. Each visualization unit is encoded with the abstraction of spatiotemporal properties. To reduce user cognition load, *UcVE* allows users to visualize, save, and track in-the-process exploration results. In coordination of storage sequence and block tracking views, *UcVE* can facilitate comparison with multiple visualization units concurrently, selected from historical and current exploration results. *UcVE* offers a flexible geo-based layout, with aggregation functions and temporal views of the timeline with categorized events, to maximize the user's exploration capabilities. Finally, we demonstrate the usefulness by using COVID-19 datasets, case studies with different user scenarios, and expert feedback.

In light of these limitations and the feedback from the domain experts via *UcVE*'s evaluation, we will update our system by expanding the diversity of related COVID-19 datasets and updating the supporting functions. We will take into consideration the NSW government follow-up updates and obtain more precise and thorough location data. For example, we can drill down from LGAs to postal areas and even construct a visual analytic system based on travel paths if permitted. Simultaneously, we will consider incorporating census data for each sampling location, such as local population, economic indexes, etc., to better understand the relationship between these factors and the epidemical spread. We will also improve the system storage function to enable the collaborative exploration of multi-users online as well as an analysis of multiple user interaction behaviors. With these improvements, we also plan to conduct a systematic evaluation with more government staff. As suggested, we will offer an automated analysis function of real-time data, which would help relevant policymakers apply proper decisions.

7.4 Conclusion and Future Direction of ClinicLens

The *ClinicLens*, an interactive visual analytics system for exploring and optimizing the testing capacities of clinics in spatiotemporal and multidimensional features given uncertainties. *ClinicLens* houses a range of features based on an aggregated set of COVID-19 data. It comprises Back-end Engine and Front-end Visualization that take users through an iterative exploration chain of extracting, training, and predicting testing-sensitive features and visual representations. It also combines AI4VIS and visual analytics to demonstrate how a clinic might optimize its testing capacity given the impacts of a range of features. Three qualitative case studies along with feedback from subject-matter experts validate that *ClinicLens* is both a useful and effective tool for exploring the trends in COVID-19 and optimizing clinic testing capacities across regions.

Future work on *ClinicLens* involves several key areas of development. First, we plan to enhance the Back-end Engine by incorporating additional implicit features and integrating state-of-the-art algorithms. For instance, we will explore the utilization of advanced algorithms outlined in [121], which can enhance the system's robustness to uncertainty and improve feature modeling and predictions. These advancements will contribute to more accurate and reliable results, empowering decision-makers with better insights for managing infectious disease outbreaks.

Another aspect of our future work involves the development of a clinic editor as a new system function. This feature will provide users with the ability to edit and update clinic information, facilitating efficient management of clinic data within the system. By incorporating this editor function, we aim to enhance the usability and versatility of our system, enabling users to maintain accurate and up-to-date clinic information.

Furthermore, we have plans to expand the coverage of *ClinicLens* beyond its current implementation in New South Wales. We aim to extend the system's capabilities to cover other states and territories in Australia, creating a comprehensive and real-time system for managing future outbreaks of infectious diseases nationwide. This expansion will enable health authorities and policymakers to have a unified and coordinated approach to effectively respond to potential outbreaks across different regions.

By incorporating implicit features, integrating advanced algorithms, developing a clinic editor function, and expanding *ClinicLens* to cover other states and territories, we are committed to continuous improvement and ensuring that our system remains at the forefront of disease management and public health initiatives.

7.5 Conclusion

The *PansyTree*, *+msRNAers*, and *UcVE* have potential capacities that are designed as general visual analytical approaches that can be easily expanded to other eligible multivariate datasets. In contrast, the *ClinicLens* pays more attention to providing visual solutions for the urgent needs of clinics' testing capacities. All these approaches provide flexible and interactive methods for users to explore, compare, and understand complex data in different application scenarios and have received positive feedback from experts in relevant fields.

In conclusion, the research presented in this thesis focuses on conducting a closed loop of visual comparative studies on multivariate data analysis. The four main approaches proposed offer different emphases on enhancing the understanding and comparison from different angles of multivariate data. Future direction on this topic could include further refinement and expansion of these approaches, improvements on exploring and comparing functions, integration with real-world users, and the development of new visual analytical methods to pursue more accurate, intelligent, and real-time responses, especially on the 3E (effectiveness efficiency, and experience) of visualization performance in multivariate data.

Chapter 8

Appendix

All four related approaches' codes in this thesis have been posted available on Github:

1. *PansyTree*: https://github.com/YuDong5018/PansyTree.

2. +*msRNAers*: https://github.com/YuDong5018/msRNAers.

3. *UcVE*: https://github.com/YuDong5018/UcVE.

4. *ClinicLens*: https://github.com/YuDong5018/clinic-lens.

Bibliography

[1] DATA.NSW , *NSW COVID-19 cases data* .
https://data.nsw.gov.au/nsw-covid-19-data/cases, March 2021.

[2] NSW HEALTH , *Omicron variant in confirmed nsw cases*.
https://www.health.nsw.gov.au/news/Pages/, November 2021.

[3] THE NSW GOVERNMENT, *NSW Government COVID-19 Data Program*.
https://data.nsw.gov.au/nsw-government-covid-19-data-program.

[4] THE NSW GOVERNMENT , *Who are very low to moderate income earners?*
https://www.facs.nsw.gov.au/providers/housing/affordable/about/chapters/who-are-very-low-to-moderate-income-earners, September 2019.

[5] P. ABRAHAM, J.-A. MANSKI-NANKERVIS, R. BIEZEN, C. M. HALLINAN, K. B. GIBNEY, L. SANCI, AND J. RIDE, *Costing of an australian general practice covid-19 drive-through testing and respiratory clinic*, BMC primary care, 23 (2022), pp. 1–7.

[6] S. AFZAL, S. GHANI, H. C. JENKINS-SMITH, D. S. EBERT, M. HADWIGER, AND I. HOTEIT, *A visual analytics based decision making environment for covid-19 modeling and visualization*, in 2020 IEEE Visualization Conference (VIS), IEEE, 2020, pp. 86–90.

[7] S. AFZAL, R. MACIEJEWSKI, AND D. S. EBERT, *Visual analytics decision support environment for epidemic modeling and response evaluation*, in 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), 2011, pp. 191–200.

[8] AGDH, *Queensland covid-19 statistics*.
Australian Government | Department of Health, March 2021.

[9] AIHW, *Health expenditure australia 2018-19*. Australian Government, November 2020.

[10] S. AL-DOHUKI, Y. WU, F. KAMW, J. YANG, X. LI, Y. ZHAO, X. YE, W. CHEN, C. MA, AND F. WANG, *Semantictraj: A new approach to interacting with massive taxi trajectories*, IEEE transactions on visualization and computer graphics, 23 (2016), pp. 11–20.

[11] H. ALBAZZAZ, X. Z. WANG, AND F. MARHOON, *Multidimensional visualisation for process historical data analysis: a comparative study with multivariate statistical process control*, Journal of Process Control, 15 (2005), pp. 285–294.

[12] D. ALBERS, C. DEWEY, AND M. GLEICHER, *Sequence surveyor: Leveraging overview for scalable genomic alignment visualization*, IEEE transactions on visualization and computer graphics, 17 (2011), pp. 2392–2401.

[13] Y. ALBO, J. LANIR, P. BAK, AND S. RAFAELI, *Off the radar: Comparative evaluation of radial visualization solutions for composite indicators*, IEEE transactions on visualization and computer graphics, 22 (2015), pp. 569–578.

[14] T. ALJREES, D. SHI, D. WINDRIDGE, AND W. WONG, *Criminal pattern identification based on modified k-means clustering*, in 2016 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2, IEEE, 2016, pp. 799–806.

[15] 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*, in Computer Graphics Forum, vol. 29, Wiley Online Library, 2010, pp. 913–922.

[16] G. ANDRIENKO, N. ANDRIENKO, G. FUCHS, AND J. WOOD, *Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data*, IEEE transactions on visualization and computer graphics, 23 (2016), pp. 2120–2136.

[17] M. ANGELINI AND G. CAZZETTA, *Progressive visualization of epidemiological models for covid-19 visual analysis*, in Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications, Springer, 2020, pp. 163–173.

[18] D. ANTWEILER, D. SESSLER, S. GINZEL, AND J. KOHLHAMMER, *Towards the detection and visual analysis of covid-19 infection clusters*, 2021.

[19] D. ANTWEILER, D. SESSLER, M. ROSSKNECHT, B. ABB, S. GINZEL, AND J. KOHLHAMMER, *Uncovering chains of infections through spatio-temporal and visual analysis of covid-19 contact traces*, Computers & Graphics, (2022).

[20] ANU, *New data visualisation tool to help track covid-19*.
Australia National University, April 2020.

[21] B. BACH, C. SHI, N. HEULOT, T. MADHYASTHA, T. GRABOWSKI, AND P. DRAGICEVIC, *Time curves: Folding time to visualize patterns of temporal evolution in data*, IEEE transactions on visualization and computer graphics, 22 (2015), pp. 559–568.

[22] P. BACHTIGER, N. S. PETERS, AND S. L. WALSH, *Machine learning for covid-19,Äîasking the right questions*, The Lancet Digital Health, 2 (2020), pp. e391–e392.

[23] H. S. BADR, H. DU, M. MARSHALL, E. DONG, M. M. SQUIRE, AND L. M. GARDNER, *Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study*, The Lancet Infectious Diseases, 20 (2020), pp. 1247–1254.

[24] H. BAO, X. ZHOU, Y. XIE, Y. ZHANG, AND Y. LI, *Covid-gan+: Estimating human mobility responses to covid-19 through spatio-temporal generative adversarial networks with enhanced features*, ACM Transactions on Intelligent Systems and Technology (TIST), 13 (2022), pp. 1–23.

[25] M. BEHRISCH, J. DAVEY, S. SIMON, T. SCHRECK, D. KEIM, AND J. KOHLHAMMER, *Visual comparison of orderings and rankings*, in EuroVis, 2013.

[26] A. BHASKAR, J. CHANDRA, H. HASHEMI, K. BUTLER, L. BENNETT, J. CELLINI, D. BRAUN, AND F. DOMINICI, *A literature review of the effects of air pollution on covid-19 health outcomes worldwide: Statistical challenges and data visualization*, Annual Review of Public Health, 44 (2022).

[27] BING, *Conronavirus australia - live map tracker from microsoft bing*.
bing, March 2021.

[28] M. BOSTOCK, V. OGIEVETSKY, AND J. HEER, *D^3 data-driven documents*, IEEE transactions on visualization and computer graphics, 17 (2011), pp. 2301–2309.

[29] E. Bowe, E. Simmons, and S. Mattern, *Learning from lines: Critical covid data visualizations and the quarantine quotidian*, Big data & society, 7 (2020), p. 2053951720939236.

[30] M. Brehmer, B. Lee, B. Bach, N. H. Riche, and T. Munzner, *Timelines revisited: A design space and considerations for expressive storytelling*, IEEE transactions on visualization and computer graphics, 23 (2016), pp. 2151–2164.

[31] S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, and K. Hamacherk, *Interactive visual comparison of multiple trees*, in 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2011, pp. 31–40.

[32] C. Brunsdon, J. Corcoran, and G. Higgs, *Visualising space and time in crime patterns: A comparison of methods*, Computers, environment and urban systems, 31 (2007), pp. 52–75.

[33] CA, *Covid-9 case tracker australia.*
COVID-19-au.com, March 2021.

[34] L. Cao, *Data Science Thinking: The Next Scientific, Technological and Economic Revolution*, Data Analytics, Springer International Publishing, 2018.

[35] L. Cao, *Ai in combating the covid-19 pandemic*, IEEE Intelligent Systems, 37 (2022), pp. 3–13.

[36] L. Cao and Q. Liu, *Covid-19 modeling: A review*, medRxiv, (2022).

[37] Cao, Longbing, *A new age of ai: Features and futures*, IEEE Intelligent Systems, 37 (2022), pp. 25–37.

[38] L. N. Carroll, A. P. Au, L. T. Detwiler, T.-c. Fu, I. S. Painter, and N. F. Abernethy, *Visualization and analytics tools for infectious disease epidemiology: a systematic review*, Journal of biomedical informatics, 51 (2014), pp. 287–298.

[39] W. W.-Y. Chan, *A survey on multivariate data visualization*, Department of Computer Science and Engineering. Hong Kong University of Science and Technology, 8 (2006), pp. 1–29.

[40] B. Chen, M. Shi, X. Ni, L. Ruan, H. Jiang, H. Yao, M. Wang, Z. Song, Q. Zhou, and T. Ge, *Visual data analysis and simulation prediction for covid-19*, arXiv preprint arXiv:2002.07096, (2020).

[41] L. Chen, Y. Ouyang, H. Zhang, S. Hong, and Q. Li, *Riseer: Inspecting the status and dynamics of regional industrial structure via visual analytics*, IEEE Transactions on Visualization and Computer Graphics, 29 (2022), pp. 1070–1080.

[42] Y. Chen, Y. Dong, Y. Sun, and J. Liang, *A multi-comparable visual analytic approach for complex hierarchical data*, Journal of Visual Languages & Computing, 47 (2018), pp. 19–30.

[43] N. A. Christakis and J. H. Fowler, *Social network visualization in epidemiology*, Norsk epidemiologi= Norwegian journal of epidemiology, 19 (2009), p. 5.

[44] J. Chua, B. Lim, E. K. Fenwick, A. T. L. Gan, A. G. Tan, E. Lamoureux, P. Mitchell, J. J. Wang, T. Y. Wong, and C.-Y. Cheng, *Prevalence, risk factors, and impact of undiagnosed visually significant cataract: The singapore epidemiology of eye diseases study*, PLoS One, 12 (2017), p. e0170804.

[45] K. K. Chui, J. B. Wenger, S. A. Cohen, and E. N. Naumova, *Visual analytics for epidemiologists: understanding the interactions between age, time, and disease with multi-panel graphs*, PloS one, 6 (2011), p. e14683.

[46] T. Dang and A. Forbes, *Cactustree: A tree drawing approach for hierarchical edge bundling*, in 2017 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2017, pp. 210–214.

[47] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna, *Mauve: multiple alignment of conserved genomic sequence with rearrangements*, Genome research, 14 (2004), pp. 1394–1403.

[48] S. Das and A. Endert, *Legion: Visually compare modeling techniques for regression*, in 2020 Visualization in Data Science (VDS), IEEE, 2020, pp. 12–21.

[49] Data.NSW, *NSW Administrative Boundaries - SEED* . https://datasets.seed.nsw.gov.au, Nov 2022.

[50] S. N. de Melo, D. V. Pereira, M. A. Andresen, and L. F. Matias, *Spatial/temporal variations of crime: A routine activity theory perspective*, International journal of offender therapy and comparative criminology, 62 (2018), pp. 1967–1991.

[51] C. DelCourt, G. Moreau, and A. Cougnard-Gregoire, *The potential of cardiovascular risk factors for reducing visual impairment: a pooled analysis of european epidemiological studies*, Investigative Ophthalmology & Visual Science, 58 (2017), pp. 2209–2209.

[52] Z. Deng, D. Weng, Y. Liang, J. Bao, Y. Zheng, T. Schreck, M. Xu, and Y. Wu, *Visual cascade analytics of large-scale spatiotemporal data*, IEEE Transactions on Visualization and Computer Graphics, (2021).

[53] Z. Deng, D. Weng, S. Liu, Y. Tian, M. Xu, and Y. Wu, *A survey of urban visual analytics: Advances and future directions*, Computational Visual Media, 9 (2023), pp. 3–39.

[54] Z. Deng, D. Weng, X. Xie, J. Bao, Y. Zheng, M. Xu, W. Chen, and Y. Wu, *Compass: Towards better causal analysis of urban time series*, IEEE Transactions on Visualization and Computer Graphics, 28 (2021), pp. 1051–1061.

[55] A. Desai, P. Nouvellet, S. Bhatia, A. Cori, and B. Lassmann, *Data journalism and the covid-19 pandemic: opportunities and challenges*, The Lancet Digital Health, 3 (2021), pp. e619–e621.

[56] S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader, *Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis approach*, Journal of medical virology, 92 (2020), pp. 632–638.

[57] E. Dimara, A. Bezerianos, and P. Dragicevic, *Conceptual and methodological issues in evaluating multidimensional visualizations for decision support*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 749–759.

[58] K. Dinkla, M. A. Westenberg, H. Timmerman, S. A. van Hijum, and J. J. van Wijk, *Comparison of multiple weighted hierarchies: visual analytics for microbe community profiling*, in Computer Graphics Forum, vol. 30, Wiley Online Library, 2011, pp. 1141–1150.

[59] E. Dong, H. Du, and L. Gardner, *An interactive web-based dashboard to track covid-19 in real time*, The Lancet infectious diseases, 20 (2020), pp. 533–534.

[60] Y. Dong, A. Fauth, M. Huang, Y. Chen, and J. Liang, *Pansytree: Merging multiple hierarchies*, in 2020 IEEE Pacific Visualization Symposium (PacificVis), 2020, pp. 131–135.

[61] T. Dwyer, S.-H. Hong, D. Koschützki, F. Schreiber, and K. Xu, *Visual analysis of network centralities*, in Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60, Citeseer, 2006, pp. 189–197.

[62] J. Dykes, A. Abdul-Rahman, D. Archambault, B. Bach, R. Borgo, M. Chen, J. Enright, H. Fang, E. E. Firat, E. Freeman, et al., *Visualization for epidemiological modelling: challenges, solutions, reflections and recommendations*, Philosophical Transactions of the Royal Society A, 380 (2022), p. 20210299.

[63] G. Dzemyda, O. Kurasova, and J. Zilinskas, *Multidimensional data visualization*, Methods and applications series: Springer optimization and its applications, 75 (2013), pp. 10–5555.

[64] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi, *The state of the art in integrating machine learning into visual analytics*, in Computer Graphics Forum, vol. 36, Wiley Online Library, 2017, pp. 458–486.

[65] D. Engel, L. Hüttenberger, and B. Hamann, *A survey of dimension reduction methods for high-dimensional data analysis and visualization*, in Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[66] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini, *The influence of contour on similarity perception of star glyphs*, IEEE transactions on visualization and computer graphics, 20 (2014), pp. 2251–2260.

[67] T. Fujiwara, J.-K. Chou, S. Shilpika, P. Xu, L. Ren, and K.-L. Ma, *An incremental dimensionality reduction method for visualizing streaming multidimensional data*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 418–428.

[68] B. GHARIZADEH, J. YUE, M. YU, Y. LIU, M. ZHOU, D. LU, AND J. ZHANG, *Navigating the pandemic response life cycle: molecular diagnostics and immunoassays in the context of covid-19 management*, IEEE reviews in biomedical engineering, 14 (2020), pp. 30–47.

[69] G. GIORDANO, F. BLANCHINI, R. BRUNO, P. COLANERI, A. DI FILIPPO, A. DI MATTEO, AND M. COLANERI, *Modelling the covid-19 epidemic and implementation of population-wide interventions in italy*, Nature medicine, 26 (2020), pp. 855–860.

[70] M. GLEICHER, *Considerations for visualizing comparison*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 413–423.

[71] M. GLEICHER, D. ALBERS, R. WALKER, I. JUSUFI, C. D. HANSEN, AND J. C. ROBERTS, *Visual comparison for information visualization*, Information Visualization, 10 (2011), pp. 289–309.

[72] E. GOETSCHEL, J. SEKARAN, W. REN, M. HE, N. OGBONNAYA, M. NKEREUWEM, I. MAPFUNDE, C. MARTIN, C. COGBURN, AND S. FEINER, *Coviz: Visualization of effects of covid-19 on new york city through socially impactful virtual reality*, in 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), IEEE, 2021, pp. 703–704.

[73] S. GOODWIN, J. DYKES, A. SLINGSBY, AND C. TURKAY, *Visualizing multiple variables across scale and geography*, IEEE Transactions on Visualization and Computer Graphics, 22 (2015), pp. 599–608.

[74] J. GÖRTLER, F. HOHMAN, D. MORITZ, K. WONGSUPHASAWAT, D. REN, R. NAIR, M. KIRCHNER, AND K. PATEL, *Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels*, in CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–13.

[75] M. GRAHAM AND J. KENNEDY, *Exploring multiple trees through dag representations*, IEEE transactions on visualization and computer graphics, 13 (2007), pp. 1294–1301.

[76] G. G. GRINSTEIN AND M. O. WARD, *Introduction to data visualization*, Information visualization in data mining and knowledge discovery, 1 (2002), pp. 21–45.

[77] GSA, *New data visualisation tool to help track covid-19*.
Government of South Australia, March 2021.

[78] J. GUERRA-GOMEZ, M. L. PACK, C. PLAISANT, AND B. SHNEIDERMAN, *Visualizing change over time using dynamic hierarchies: Treeversity2 and the stemview*, IEEE Transactions on Visualization and Computer Graphics, 19 (2013), pp. 2566–2575.

[79] J. HAGENAUER, M. HELBICH, AND M. LEITNER, *Visualization of crime trajectories with self-organizing maps: a case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots*, in Proceedings of the 25th conference of the International Cartographic Association, Paris, 2011.

[80] A. HAMDI, K. SHABAN, A. ERRADI, A. MOHAMED, S. K. RUMI, AND F. D. SALIM, *Spatiotemporal data mining: a survey on challenges and open problems*, Artificial Intelligence Review, 55 (2022), pp. 1441–1488.

[81] J. HAN, M. KAMBER, AND J. PEI, *Data mining trends and research frontiers*, Data Min, (2012), pp. 585–631.

[82] K. A. HASSAN, N. RÖNNBERG, C. FORSELL, M. COOPER, AND J. JOHANSSON, *A study on 2d and 3d parallel coordinates for pattern identification in temporal multivariate data*, in 2019 23rd International Conference Information Visualisation (IV), IEEE, 2019, pp. 145–150.

[83] X. HE, Y. TAO, Q. WANG, AND H. LIN, *Multivariate spatial data visualization: a survey*, Journal of visualization, 22 (2019), pp. 897–912.

[84] C. G. HEALEY, S. J. SIMMONS, C. MANIVANNAN, AND Y. RO, *Visual analytics for the coronavirus covid-19 pandemic*, Big Data, 10 (2022), pp. 95–114.

[85] J. HEER, M. BOSTOCK, AND V. OGIEVETSKY, *A tour through the visualization zoo*, Communications of the ACM, 53 (2010), pp. 59–67.

[86] O. S. HEMIED, M. S. GADELRAB, E. A. SHARARA, T. H. A. SOLIMAN, A. TSUJI, AND K. TERADA, *A covid-19 visual diagnosis model based on deep learning and gradcam*, IEEE Transactions on Electrical and Electronic Engineering, (2022).

[87] D. HOLTEN AND J. J. VAN WIJK, *Visual comparison of hierarchically organized data*, in Computer Graphics Forum, vol. 27, Wiley Online Library, 2008, pp. 759–766.

[88] J. Hua, M. L. Huang, C. Zhao, S. Hua, and C. Shih, *An initial visual analysis of the relationship between covid-19 and local community features*, in 2020 24th International Conference Information Visualisation (IV), IEEE, 2020, pp. 718–722.

[89] D. Huang, M. Tory, S. Staub-French, and R. Pottinger, *Visualization techniques for schedule comparison*, in Computer graphics forum, vol. 28, Wiley Online Library, 2009, pp. 951–958.

[90] M. L. Huang, T.-H. Huang, and J. Zhang, *Treemapbar: Visualizing additional dimensions of data in bar chart*, in 2009 13th International Conference Information Visualisation, IEEE, 2009, pp. 98–103.

[91] M. L. Huang, Z. Yue, Q. V. Nguyen, J. Liang, and Z. Luo, *Stroke data analysis through a hvn visual mining platform*, in 2019 23rd International Conference in Information Visualization–Part II, IEEE, 2019, pp. 1–6.

[92] T.-H. Huang, M. L. Huang, Q. V. Nguyen, and L. Zhao, *A space-filling multidimensional visualization (sfmdvis for exploratory data analysis*, in Proceedings of the 7th International Symposium on Visual Information Communication and Interaction, 2014, pp. 19–28.

[93] C. Hurter, O. Ersoy, and A. Telea, *Graph bundling by kernel density estimation*, in Computer graphics forum, vol. 31, Wiley Online Library, 2012, pp. 865–874.

[94] A. Inselberg and B. Dimsdale, *Parallel coordinates: a tool for visualizing multi-dimensional geometry*, in Proceedings of the First IEEE Conference on Visualization: Visualization90, IEEE, 1990, pp. 361–378.

[95] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, *Visualization as seen through its research paper keywords*, IEEE Transactions on Visualization and Computer Graphics, 23 (2016), pp. 771–780.

[96] S. Jadhav, G. Deng, M. Zawin, and A. E. Kaufman, *Covid-view: Diagnosis of covid-19 using chest ct*, IEEE transactions on visualization and computer graphics, 28 (2021), pp. 227–237.

[97] S. JAMONNAK, Y. ZHAO, X. HUANG, AND M. AMIRUZZAMAN, *Geo-context aware study of vision-based autonomous driving models and spatial video data*, IEEE transactions on visualization and computer graphics, 28 (2021), pp. 1019–1029.

[98] B. JIANG, X. YOU, K. LI, T. LI, X. ZHOU, AND L. TAN, *Interactive analysis of epidemic situations based on a spatiotemporal information knowledge graph of covid-19*, IEEE Access, (2020).

[99] B. JOHNSON AND B. SHNEIDERMAN, *Tree-maps: A space filling approach to the visualization of hierarchical information structures*, tech. rep., 1998.

[100] P. KAHN, *Covic project summary 011621.* https://mprove.de/script/20/covic/_media/COVICProjectSummary011621.pdf, January 2021.

[101] D. A. KEIM, *Designing pixel-oriented visualization techniques: Theory and applications*, IEEE Transactions on visualization and computer graphics, 6 (2000), pp. 59–78.

[102] B. M. KELLER, A. P. REEVES, T. V. APANOSOVICH, J. WANG, D. F. YANKELEVITZ, AND C. I. HENSCHKE, *Quantitative assessment of emphysema from whole lung ct scans: comparison with visual grading*, in Medical Imaging 2009: Computer-Aided Diagnosis, vol. 7260, SPIE, 2009, pp. 74–81.

[103] R. KERRY, P. GOOVAERTS, R. P. HAINING, AND V. CECCATO, *Applying geostatistical analysis to crime data: Car-related thefts in the baltic states.*, Geographical analysis, 42 (2010), pp. 53–77.

[104] H. KIIVERI AND T. SPEED, *Structural analysis of multivariate data: A review*, Sociological methodology, 13 (1982), pp. 209–289.

[105] J. KIM AND K. WOOD, *Visualizing hierarchical structures in multivariate data: A survey*, Journal of Visualization, 25 (2022), pp. 123–139.

[106] E. KLEIBERG, H. VAN DE WETERING, AND J. J. VAN WIJK, *Botanical visualization of huge hierarchies*, in IEEE Symposium on Information Visualization, 2001. INFOVIS 2001., IEEE, 2001, pp. 87–94.

[107] H. KOBAYASHI, T. FURUKAWA, AND K. MISUE, *Parallel box: Visually comparable representation for multivariate data analysis*, in 2014 18th International Conference on Information Visualisation, IEEE, 2014, pp. 183–188.

[108] B. Lee, G. G. Robertson, M. Czerwinski, and C. S. Parr, *Candidtree: visualizing structural uncertainty in similar hierarchies*, Information Visualization, 6 (2007), pp. 233–246.

[109] C. Lee, Y. Kim, S. Jin, D. Kim, R. Maciejewski, D. Ebert, and S. Ko, *A visual analytics system for exploring, monitoring, and forecasting road traffic congestion*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 3133–3146.

[110] R. A. Leite, V. Schetinger, D. Ceneda, B. Henz, and S. Miksch, *Covis: Supporting temporal visual analysis of covid-19 events usable in data-driven journalism*, in 2020 IEEE Visualization Conference (VIS), IEEE, 2020, pp. 56–60.

[111] C. K. Leung, Y. Chen, C. S. Hoi, S. Shang, Y. Wen, and A. Cuzzocrea, *Big data visualization and visual analytics of covid-19 data*, in 2020 24th International Conference Information Visualisation (IV), 2020, pp. 415–420.

[112] N. Levine et al., *Crimestat iv: a spatial statistics program for the analysis of crime incident locations, version 4.0*, Ned Levine & Associates: Houston, TX, USA, (2013).

[113] C. Li, G. Baciu, Y. Wang, J. Chen, and C. Wang, *Ddlvis: Real-time visual query of spatiotemporal data distribution via density dictionary learning*, IEEE Transactions on Visualization and Computer Graphics, 28 (2021), pp. 1062–1072.

[114] D. Li, Y. Wang, S. Wu, J. Qi, and T. Wang, *An visual analytics approach to explore criminal patterns based on multidimensional data*, in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2017, pp. 5563–5566.

[115] G. Li, Y. Zhang, Y. Dong, J. Liang, J. Zhang, J. Wang, M. J. McGuffin, and X. Yuan, *Barcodetree: Scalable comparison of multiple hierarchies*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 1022–1032.

[116] R. Li, *Visualizing covid-19 information for public: Designs, effectiveness, and preference of thematic maps*, Human Behavior and Emerging Technologies, 3 (2021), pp. 97–106.

[117] Y. LI, G. LI, AND X. LI, *Survey on visualization of tree comparison*, Journal of Software, 27 (2016), pp. 1074–1090.

[118] J. LIANG, Q. V. NGUYEN, S. SIMOFF, AND M. L. HUANG, *Angular treemaps-a new technique for visualizing and emphasizing hierarchical structures*, in 2012 16th International Conference on Information Visualisation, IEEE, 2012, pp. 74–80.

[119] G. LIU, Y. LIAO, F. WANG, B. ZHANG, L. ZHANG, X. LIANG, X. WAN, S. LI, Z. LI, S. ZHANG, ET AL., *Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning*, IEEE Transactions on Neural Networks and Learning Systems, 32 (2021), pp. 3786–3797.

[120] J. LIU, T. DWYER, G. TACK, S. GRATZL, AND K. MARRIOTT, *Supporting the problem-solving loop: Designing highly interactive optimisation systems*, IEEE Transactions on Visualization and Computer Graphics, 27 (2020), pp. 1764–1774.

[121] Q. LIU AND L. CAO, *Modeling time evolving covid-19 uncertainties with density dependent asymptomatic infections and social reinforcement*, Scientific Reports, 12 (2022), pp. 1–14.

[122] Q. LIU, Q. LI, C. TANG, H. LIN, X. MA, AND T. CHEN, *A visual analytics approach to scheduling customized shuttle buses via perceiving passengers‚Äô travel demands*, in 2020 IEEE Visualization Conference (VIS), IEEE, 2020, pp. 76–80.

[123] Q. LIU, Z. ZHENG, J. ZHENG, Q. CHEN, G. LIU, S. CHEN, B. CHU, H. ZHU, B. AKINWUNMI, J. HUANG, ET AL., *Health communication through news media during the early stage of the covid-19 outbreak in china: digital topic modeling approach*, Journal of medical Internet research, 22 (2020), p. e19118.

[124] S. R. LORD AND J. DAYHEW, *Visual risk factors for falls in older people*, Journal of the American Geriatrics Society, 49 (2001), pp. 508–515.

[125] J. LUKASCZYK, R. MACIEJEWSKI, C. GARTH, AND H. HAGEN, *Understanding hotspots: A topological visual analytics approach*, in Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, 2015, pp. 1–10.

[126] J. Lv, S. Tu, and L. Xu, *Detection of phenotype-related mutations of covid-19 via the whole genomic data*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18 (2021), pp. 1242–1249.

[127] S. LYi, Q. Wang, F. Lekschas, and N. Gehlenborg, *Gosling: A grammar-based toolkit for scalable and interactive genomics data visualization*, IEEE Transactions on Visualization and Computer Graphics, 28 (2021), pp. 140–150.

[128] C. Ma, Y. Zhao, S. AL-Dohuki, J. Yang, X. Ye, F. Kamw, and M. Amiruzzaman, *Gtmaplens: Interactive lens for geo-text data browsing on map*, in Computer Graphics Forum, vol. 39, Wiley Online Library, 2020, pp. 469–481.

[129] A. Maalej, N. Rodriguez, and O. Strauss, *Survey of multidimensional visualization techniques*, in CGVCVIP'12: Computer Graphics, Visualization, Computer Vision and Image Processing Conference, 2012, pp. N–A.

[130] R. Maciejewski, P. Livengood, S. Rudolph, T. F. Collins, D. S. Ebert, R. T. Brigantic, C. D. Corley, G. A. Muller, and S. W. Sanders, *A pandemic influenza modeling and visualization tool*, Journal of Visual Languages & Computing, 22 (2011), pp. 268–278.

[131] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, *A visual analytics approach to understanding spatiotemporal hotspots*, IEEE Transactions on Visualization and Computer Graphics, 16 (2009), pp. 205–220.

[132] C. R. MacIntyre, *Case isolation, contact tracing, and physical distancing are pillars of covid-19 pandemic control, not optional choices*, The Lancet Infectious Diseases, 20 (2020), pp. 1105–1106.

[133] Mapbox.
https://www.mapbox.com, 2022.

[134] J. Matute and L. Linsen, *Visual stratification for epidemiological analysis.*, in EuroVis (Posters), 2017, pp. 81–83.

[135] M. Meyer, T. Munzner, and H. Pfister, *Mizbee: a multiscale synteny browser*, IEEE transactions on visualization and computer graphics, 15 (2009), pp. 897–904.

[136] T. MIKOLOV, K. CHEN, G. CORRADO, AND J. DEAN, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781, (2013).

[137] F.-B. MOCNIK, P. RAPOSO, W. FERINGA, M.-J. KRAAK, AND B. KÖBBEN, *Epidemics and pandemics in maps–the case of covid-19*, Journal of Maps, 16 (2020), pp. 144–152.

[138] H. MOHAMMADIGOUSHKI AND S. J. MULLER, *A flow visualization and superposition rheology study of shear-banding wormlike micelle solutions*, Soft matter, 12 (2016), pp. 1051–1061.

[139] S. F. MØLLER, J. VON FRESE, AND R. BRO, *Robust methods for multivariate data analysis*, Journal of Chemometrics: A Journal of the Chemometrics Society, 19 (2005), pp. 549–563.

[140] K. MOLONEY AND S. MOLONEY, *Australian quarantine policy: From centralization to coordination with mid-pandemic covid-19 shifts*, Public Administration Review, 80 (2020), pp. 671–682.

[141] T. MORIMAE AND A. SHIMIZU, *Visualization of superposition of macroscopically distinct states*, Physical Review A, 74 (2006), p. 052111.

[142] T. MUNZNER, *A nested model for visualization design and validation*, IEEE transactions on visualization and computer graphics, 15 (2009), pp. 921–928.

[143] K. MUTO, I. YAMAMOTO, M. NAGASU, M. TANAKA, AND K. WADA, *Japanese citizens' behavioral changes and preparedness against covid-19: an online survey during the early phase of the pandemic*, PLoS One, 15 (2020), p. e0234292.

[144] T. NAKAYA AND K. YANO, *Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics*, Transactions in GIS, 14 (2010), pp. 223–239.

[145] U. NASEEM, I. RAZZAK, M. KHUSHI, P. W. EKLUND, AND J. KIM, *Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis*, IEEE Transactions on Computational Social Systems, 8 (2021), pp. 1003–1015.

[146] Q. V. NGUYEN AND M. L. HUANG, *Space-optimized tree: a connection+ enclosure approach for the visualization of large hierarchies*, Information Visualization, 2 (2003), pp. 3–15.

[147] T. Nijssen, O. Kramer, P. de Moel, J. Rahman, J. Kroon, P. Berhanu, E. Boek, K. Buist, J. van der Hoek, J. Padding, et al., *Experimental and numerical insights into heterogeneous liquid-solid behaviour in drinking water softening reactors*, Chemical Engineering Science: X, 11 (2021), p. 100100.

[148] C. Nobre, M. Meyer, M. Streit, and A. Lex, *The state of the art in visualizing multivariate networks*, in Computer Graphics Forum, vol. 38, Wiley Online Library, 2019, pp. 807–832.

[149] L. G. Nonato and M. Aupetit, *Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment*, IEEE Transactions on Visualization and Computer Graphics, 25 (2018), pp. 2650–2673.

[150] NSW Health , *Latest media releases from nsw health*.
https://https://www.health.nsw.gov.au/news/Pages/default.aspx, March 2022.

[151] NSW Health, *Fighting the delta outbreak with new restrictions for local government areas (lgas) of concern*.
https://www.health.nsw.gov.au/news/Pages/, July 2021.

[152] NTG, *Current status coronavirus (covid-19)*.
Northern Territory Government, March 2021.

[153] D. of Health and A. G. Aged Care, *Coronavirus (covid-19) case numbers and statistics*.
https://www.health.gov.au/health-alerts/covid-19/case-numbers-and-statistics.

[154] A. B. of Statistics, *Census data*.
https://www.abs.gov.au/census, 2016.

[155] N. I. on Aging, *Why covid-19 testing is the key to getting back to normal*, National Institutes of Health, (2020).

[156] G. W. H. Organization, *Who covid-19 dashboard*.
https://https://covid19.who.int/, October 2022.

[157] W. H. Organization, *Who coronavirus disease (covid-19) dashboard*.
https://covid19.who.int/.

[158] S. Ou, X. He, W. Ji, W. Chen, L. Sui, Y. Gan, Z. Lu, Z. Lin, S. Deng, S. Przesmitzki, et al., *Machine learning model to project the impact of covid-19 on us motor gasoline demand*, Nature Energy, 5 (2020), pp. 666–673.

[159] L. Padilla, R. Fygenson, S. C. Castro, and E. Bertini, *Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations*, IEEE Transactions on Visualization and Computer Graphics, 29 (2022), pp. 12–22.

[160] H.-G. Pagendarm and F. H. Post, *Comparative visualization: Approaches and examples*, Delft University of Technology Faculty of Technical Mathematics and Informatics, 1995.

[161] J. Panovska-Griffiths, B. Swallow, R. Hinch, J. A. Cohen, K. Rosenfeld, R. M. Stuart, L. Ferretti, F. Di Lauro, C. Wymant, A. Izzo, et al., *Statistical and agent-based modelling of the transmissibility of different sars-cov-2 variants in england and impact of different interventions*, medRxiv, (2022), pp. 2021–12.

[162] M. Park, A. R. Cook, J. T. Lim, Y. Sun, and B. L. Dickens, *A systematic review of covid-19 epidemiology based on current evidence*, Journal of clinical medicine, 9 (2020), p. 967.

[163] N. Perdigao, T. G. Soldatos, K. S. Sabir, and S. I. O'Donoghue, *Visual analytics of gene sets comparison*, in 2015 Big Data Visual Analytics (BDVA), IEEE, 2015, pp. 1–2.

[164] V. V. Pham and T. Dang, *Mtdes: Multi-dimensional temporal data exploration system*, in 2018 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2018, pp. 100–101.

[165] C. M. Pooley, A. B. Doeschl-Wilson, and G. Marion, *Estimation of age-stratified contact rates during the covid-19 pandemic using a novel inference algorithm*, medRxiv, (2022).

[166] A. F. Porter, N. Sherry, P. Andersson, S. A. Johnson, S. Duchene, and B. P. Howden, *New rules for genomics-informed covid-19 responses–lessons learned from the first waves of the omicron variant in australia*, PLoS Genetics, 18 (2022), p. e1010415.

[167] B. Preim and K. Lawonn, *A survey of visual analytics for public health*, in Computer Graphics Forum, vol. 39, Wiley Online Library, 2020, pp. 543–580.

[168] QG, *Queensland covid-19 statistics*.
Queensland Government, March 2021.

[169] J. H. Ratcliffe, *Aoristic analysis: the spatial interpretation of unspecific temporal events*, International journal of geographical information science, 14 (2000), pp. 669–679.

[170] Ratcliffe, Jerry H, *A temporal constraint theory to explain opportunity-based spatial offending patterns*, Journal of Research in Crime and Delinquency, 43 (2006), pp. 261–291.

[171] L. G. S. Real, R. Bueno, and M. X. Ribeiro, *Evaluating boundary conditions and hierarchical visualization in cbir*, in 2019 23rd International Conference Information Visualisation (IV), IEEE, 2019, pp. 68–73.

[172] P. Regulski, P. Wendykier, K. Kantiem, and W. Murdzek, *Advanced methods of visual analysis and visualization of various aspects of the covid-19 outbreak in poland*, Procedia Computer Science, 192 (2021), pp. 4194–4199.

[173] A. Reinert, L. S. Snyder, J. Zhao, A. S. Fox, D. F. Hougen, C. Nicholson, and D. S. Ebert, *Visual analytics for decision-making during pandemics*, Computing in Science & Engineering, 22 (2020), pp. 48–59.

[174] S. Renton, *Covid19 phase3 report 2020*.
https://mccrindle.com.au/wp-content/uploads/reports/COVID19-Phase3-Report-2020.pdf, October 2020.

[175] K. Rezaee, H. G. Zadeh, C. Chakraborty, M. R. Khosravi, and G. Jeon, *Smart visual sensing for overcrowding in covid-19 infected cities using modified deep transfer learning*, IEEE Transactions on Industrial Informatics, (2022).

[176] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, *Coronavirus pandemic (covid-19)*, Our World in Data, (2020).
https://ourworldindata.org/coronavirus.

[177] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, *Coronavirus pandemic (covid-19)*, Our world in data, (2020).

[178] G. G. Robertson, J. D. Mackinlay, and S. K. Card, *Cone trees: animated 3d visualizations of hierarchical information*, in Proceedings of the SIGCHI conference on Human factors in computing systems, 1991, pp. 189–194.

[179] O. Romero and A. Abelló, *A survey of multidimensional modeling methodologies*, International Journal of Data Warehousing and Mining (IJDWM), 5 (2009), pp. 1–23.

[180] A. N. Roy, J. Jose, A. Sunil, N. Gautam, D. Nathalia, and A. Suresh, *Prediction and spread visualization of covid-19 pandemic using machine learning*, (2020).

[181] E. Rydow, R. Borgo, H. Fang, T. Torsney-Weir, B. Swallow, T. Porphyre, C. Turkay, and M. Chen, *Development and evaluation of two approaches of visual sensitivity analysis to support epidemiological modeling*, IEEE Transactions on Visualization and Computer Graphics, (2022).

[182] N. Saeed, H. Nam, M. I. U. Haq, and D. B. Muhammad Saqib, *A survey on multidimensional scaling*, ACM Computing Surveys (CSUR), 51 (2018), pp. 1–25.

[183] H. Samet, Y. Han, J. Kastner, and H. Wei, *Using animation to visualize spatio-temporal varying covid-19 data*, in Proceedings of the 1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19, 2020, pp. 53–62.

[184] P. Sanz-Leon, L. H. Hamilton, S. J. Raison, A. J. Pan, N. J. Stevenson, R. M. Stuart, R. G. Abeysuriya, C. C. Kerr, S. B. Lambert, and J. A. Roberts, *Modelling herd immunity requirements in queensland: impact of vaccination effectiveness, hesitancy and variants of sars-cov-2*, Philosophical Transactions of the Royal Society A, 380 (2022), p. 20210311.

[185] I. H. Sarker, *Machine learning: Algorithms, real-world applications and research directions*, SN computer science, 2 (2021), p. 160.

[186] R. Scheepens, N. Willems, H. Van de Wetering, G. Andrienko, N. Andrienko, and J. J. Van Wijk, *Composite density maps for multivariate trajectories*, IEEE Transactions on Visualization and Computer Graphics, 17 (2011), pp. 2518–2527.

[187] C. Schulz, A. Zeyfang, M. van Garderen, H. B. Lahmar, M. Herschel, and D. Weiskopf, *Simultaneous visual analysis of multiple software hierarchies*, in 2018 IEEE Working Conference on Software Visualization (VISSOFT), IEEE, 2018, pp. 87–95.

[188] H.-J. Schulz, *Treevis.net: A tree visualization reference*, IEEE Computer Graphics and Applications, 31 (2011), pp. 11–15.

[189] B. R. Shapiro and F. A. Pearman, *Using the interaction geography slicer to visualize new york city stop & frisk*, in 2017 IEEE VIS Arts Program (VISAP), IEEE, 2017, pp. 1–8.

[190] Y. Shi, S. Chen, P. Liu, J. Long, and N. Cao, *Colorcook: Augmenting color design for dashboarding with domain-associated palettes*, Proceedings of the ACM on Human-Computer Interaction, 6 (2022), pp. 1–25.

[191] B. Shneiderman, *The eyes have it: A task by data type taxonomy for information visualizations*, in The craft of information visualization, Elsevier, 2003, pp. 364–371.

[192] B. W. Silverman, *Density estimation for statistics and data analysis*, Routledge, 2018.

[193] A. Soriano-Vargas, B. Hamann, and M. C. F de Oliveira, *Tv-mv analytics: A visual analytics framework to explore time-varying multivariate data*, Information visualization, 19 (2020), pp. 3–23.

[194] S. Srabanti, G. E. Marai, and F. Miranda, *Covid-19 ensemblevis: Visual analysis of county-level ensemble forecast models*, in 2021 IEEE Workshop on Visual Analytics in Healthcare (VAHC), IEEE, 2021, pp. 1–5.

[195] J. Stasko and E. Zhang, *Focus+ context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations*, in IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, IEEE, 2000, pp. 57–65.

[196] T. Steinger, H. Gilliand, and T. Hebeisen, *Epidemiological analysis of risk factors for the spread of potato viruses in switzerland*, Annals of applied biology, 164 (2014), pp. 200–207.

[197] C. D. Stolper, A. Perer, and D. Gotz, *Progressive visual analytics: User-driven visual exploration of in-progress analytics*, IEEE Transactions on Visualization and Computer Graphics, 20 (2014), pp. 1653–1662.

[198] D. Sun, R. Huang, Y. Chen, Y. Wang, J. Zeng, M. Yuan, T.-C. Pong, and H. Qu, *Planningvis: A visual analytics approach to production planning in smart factories*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 579–589.

[199] A. Sweet and J. Davies, *Fear down, job-seeking up as australians feel the financial impact of covid-19*.
CoreData, March 2021.

[200] The NSW Government , *Covid-19 data and statistics*.
`https://www.nsw.gov.au/covid-19/data-and-statistics`.

[201] The NSW Government, *Covid-19 data and statistics*.
`https://www.nsw.gov.au/covid-19/stay-safe/data-and-statistics`, 2022.

[202] M. Thöny, R. Schnürer, R. Sieber, L. Hurni, and R. Pajarola, *Storytelling in interactive 3d geographic visualization systems*, ISPRS International Journal of Geo-Information, 7 (2018), p. 123.

[203] M. Trajkova, F. Cafaro, S. Vedak, R. Mallappa, S. R. Kankara, et al., *Exploring casual covid-19 data visualizations on twitter: Topics and challenges*, in Informatics, vol. 7(3), Multidisciplinary Digital Publishing Institute, 2020, p. 35.

[204] J. P. Ulahannan, N. Narayanan, N. Thalhath, P. Prabhakaran, S. Chaliyeduth, S. P. Suresh, M. Mohammed, E. Rajeevan, S. Joseph, A. Balakrishnan, et al., *A citizen science initiative for open data and visualization of covid-19 outbreak in kerala, india*, Journal of the American Medical Informatics Association, 27 (2020), pp. 1913–1920.

[205] UOM, *Coronavirus 10-day forecast*.
University of Melbourne, March 2021.

[206] M. Usman, H. Zhou, S. Moon, X. Zhang, P. Faloutsos, and M. Kapadi-aMember, *A multi-scale geospatial dataset and an interactive visualization dashboard for computational epidemiology and open scientific research*, IEEE Computer Graphics and Applications, (2022).

[207] USYD, *Nsw covid-19 cases and community profile by the university of sydney dashboard*.
The University of Sydney, March 2021.

[208] T. Venturini, M. Jacomy, and P. Jensen, *What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts*, Big Data & Society, 8 (2021), p. 20539517211018488.

[209] J. J. C. Vergara and G. V. Boerner, *Visualizing temporal patterns in multidimensional data*, Information Visualization, 10 (2011), pp. 35–46.

[210] VSG, *Victorian coronavirus (covid-19) data*.
Department of Health and Human Services Victoria, March 2021.

[211] G. Wang, J. Guo, M. Tang, J. F. de Queiroz Neto, C. Yau, A. Daghistani, M. Karimzadeh, W. G. Aref, and D. S. Ebert, *Stull: Unbiased online sampling for visual exploration of large spatiotemporal data*, in 2020 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2020, pp. 72–83.

[212] J. Wang, K. Zhao, D. Deng, A. Cao, X. Xie, Z. Zhou, H. Zhang, and Y. Wu, *Tac-simur: Tactic-based simulative visual analytics of table tennis*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 407–417.

[213] C. Ware, *Information visualization: perception for design*, Morgan Kaufmann, 2019.

[214] L. L. Y. Wei, A. A. A. Ibrahim, K. Nisar, Z. I. A. Ismail, and I. Welch, *Survey on geographic visual display techniques in epidemiology: Taxonomy and characterization*, Journal of Industrial Information Integration, 18 (2020), p. 100139.

[215] J. W. Wells et al., *Visualizing the covid-19 pandemic*, The Lancet Digital Health, 2 (2020), pp. e1–e4.

[216] D. Weng, C. Zheng, Z. Deng, M. Ma, J. Bao, Y. Zheng, M. Xu, and Y. Wu, *Towards better bus networks: a visual analytics approach*, IEEE transactions on visualization and computer graphics, 27 (2020), pp. 817–827.

[217] E. P. White, S. M. Ernest, P. B. Adler, A. H. Hurlbert, and S. K. Lyons, *Integrating spatial and temporal approaches to understanding species richness*, Philosophical Transactions of the Royal Society B: Biological Sciences, 365 (2010), pp. 3633–3643.

[218] WHO, *Who coronavirus (covid-19) dashboard.*
https://covid19.who.int/info/, March 2021.

[219] E. Wilde and D. German, *Merge-tree: Visualizing the integration of commits into linux*, Journal of Software: Evolution and Process, 30 (2018), p. e1936.

[220] P. C. Wong and R. D. Bergeron, *30 years of multidimensional multivariate visualization.*, Scientific Visualization, 2 (1994), pp. 3–33.

[221] L. Woodburn, Y. Yang, and K. Marriott, *Interactive visualisation of hierarchical quantitative data: an evaluation*, in 2019 IEEE Visualization Conference (VIS), IEEE, 2019, pp. 96–100.

[222] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, and M.-M. Cheng, *Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation*, IEEE Transactions on Image Processing, 30 (2021), pp. 3113–3126.

[223] X. Xie, J. Wang, H. Liang, D. Deng, S. Cheng, H. Zhang, W. Chen, and Y. Wu, *Passvizor: Toward better understanding of the dynamics of soccer passes*, IEEE Transactions on Visualization and Computer Graphics, 27 (2020), pp. 1322–1331.

[224] H. Xu, A. Berres, G. Thakur, J. Sanyal, and S. Chinthavali, *Episemblevis: A geo-visual analysis and comparison of the prediction ensembles of multiple covid-19 models*, Journal of Biomedical Informatics, 124 (2021), p. 103941.

[225] K. Xu, J. Yuan, Y. Wang, C. Silva, and E. Bertini, *mtseer: Interactive visual exploration of models on multivariate time-series forecast*, in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–15.

[226] A. S. Yadaw, Y.-c. Li, S. Bose, R. Iyengar, S. Bunyavanich, and G. Pandey, *Clinical features of covid-19 mortality: development and validation of a clinical prediction model*, The Lancet Digital Health, 2 (2020), pp. e516–e525.

[227] L. Yan, Y. Wang, E. Munch, E. Gasparovic, and B. Wang, *A structural average of labeled merge trees for uncertainty visualization*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 832–842.

[228] C. Yang, Z. Zhang, Z. Fan, R. Jiang, Q. Chen, X. Song, and R. Shibasaki, *Epimob: Interactive visual analytics of citywidehuman mobility restrictions for epidemic control*, IEEE Transactions on Visualization and Computer Graphics, (2022).

[229] J. Yang, M. O. Ward, and E. A. Rundensteiner, *Interring: An interactive tool for visually navigating and manipulating hierarchical structures*, in IEEE Symposium on Information Visualization, 2002. INFOVIS 2002., IEEE, 2002, pp. 77–84.

[230] T. Yang, Y. Zhou, D. Feng, and H. Hu, *Cvas: An interactive visual analytics system for exploring covid-19 information on the web*, in 2022 The 6th International Conference on Compute and Data Analysis, 2022, pp. 116–121.

[231] X. Yang, L. Shi, M. Daianu, H. Tong, Q. Liu, and P. Thompson, *Blockwise human brain network visual comparison using nodetrix representation*, IEEE transactions on visualization and computer graphics, 23 (2016), pp. 181–190.

[232] A. Yoshizumi, M. M. Coffer, E. L. Collins, M. D. Gaines, X. Gao, K. Jones, I. R. McGregor, K. A. McQuillan, V. Perin, L. M. Tomkins, et al., *A review of geospatial content in ieee visualization publications*, in 2020 IEEE Visualization Conference (VIS), IEEE, 2020, pp. 51–55.

[233] D. Yu, O. Ian, L. Jie, Y. Xiaoru, and N. Q. Vinh, *User-centered visual explorer of in-process comparison in spatiotemporal space*, Journal of Visualization, (2022), pp. 1–19.

[234] X. Yu, M. D. Ferreira, and F. V. Paulovich, *Senti-covid19: An interactive visual analytics system for detecting public sentiment and insights regarding covid-19 from social media*, IEEE Access, 9 (2021), pp. 126684–126697.

[235] C. Zhang, X. Wang, C. Zhao, Y. Ren, T. Zhang, Z. Peng, X. Fan, X. Ma, and Q. Li, *Promotionlens: Inspecting promotion strategies of online e-commerce via visual analytics*, IEEE Transactions on Visualization and Computer Graphics, 29 (2022), pp. 767–777.

[236] W. Zhang, J. K. Wong, X. Wang, Y. Gong, R. Zhu, K. Liu, Z. Yan, S. Tan, H. Qu, S. Chen, et al., *Cohortva: A visual analytic system for interactive exploration of cohorts based on historical data*, IEEE Transactions on Visualization and Computer Graphics, 29 (2022), pp. 756–766.

[237] Y. Zhang, Y. Sun, J. D. Gaggiano, N. Kumar, C. Andris, and A. G. Parker, *Visualization design practices in a crisis: Behind the scenes with covid-19 dashboard creators*, IEEE Transactions on Visualization and Computer Graphics, (2022).

[238] Y. Zhang, Y. Sun, L. Padilla, S. Barua, E. Bertini, and A. G. Parker, *Mapping the landscape of covid-19 crisis visualizations*, in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–23.

[239] J. Zhao, F. Chevalier, C. Collins, and R. Balakrishnan, *Facilitating discourse analysis with interactive visualization*, IEEE Transactions on Visualization and Computer Graphics, 18 (2012), pp. 2639–2648.

[240] B. Zheng and F. Sadlo, *On the visualization of hierarchical multivariate data*, in 2021 IEEE 14th Pacific Visualization Symposium (PacificVis), IEEE, 2021, pp. 136–145.

[241] Y. Zhou, H. He, J. Rong, Y. Cheng, Y. Li, W. Zhong, and F. Jiang, *Visual analysis and exploration of covid-19 based on multi-source heterogeneous data*, in 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), IEEE, 2020, pp. 62–69.

[242] Z. Zhou, L. Meng, C. Tang, Y. Zhao, Z. Guo, M. Hu, and W. Chen, *Visual abstraction of large scale geospatial origin-destination movement data*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 43–53.

[243] Z. Zhou, X. Zhang, X. Zhou, and Y. Liu, *Semantic-aware visual abstraction of large-scale social media data with geo-tags*, IEEE Access, 7 (2019), pp. 114851–114861.