Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Epidemiological Insights into Disease and Risk Factor Interactions: A Web-Based Visualization Approach

Julien Colot

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :
Prof. dr. ir. Jan AERTS

SUPERVISOR :
Karim DOUIEB

2023
2024

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Epidemiological Insights into Disease and Risk Factor Interactions: A Web-Based Visualization Approach

Julien Colot

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :

Prof. dr. ir. Jan AERTS

SUPERVISOR :

Karim DOUIEB

Democratizing Epidemiology: Interactive Web Visualizations for Public Health Awareness

A Thesis

Submitted For The Degree Of

Master in Statistics and Data Science

by

Julien Colot

Internal Supervisor: Professor Jan Aerts

External Supervisor: Karim Douïeb

Faculty of Sciences

University of Hasselt

August 2024

This thesis is dedicated to
Nour and Hanène

Acknowledgements

I would like to express my gratitude to my internal supervisor, Professor Jan Aerts, for his continuous guidance and support throughout the course of this research. His expertise and encouragement have been instrumental in the completion of this thesis. I am also deeply thankful to my colleague, Karim Douieb, for his collaboration and insights. Additionally, I extend my sincere appreciation to Emmanuel André's team, including Caspar Geenen, Jonathan Thibaut, and Sarah Gorissen, for their contributions and assistance. My special thanks go to my family, Hanène and Nour, for their continuous support and understanding, which allowed me to dedicate my free time to studying the Master of Statistics and Data Science from University of Hasselt, and particularly, to work through the completion of this thesis. I would also like to extend my gratitude to the open-source community of visualization tool creators, without whose contributions this thesis would not have been possible. This includes Mike Bostock and Philippe Rivière from `D3` and `Observable`, Volodymyr Agafonkin, the author of `Leaflet` and `Supercluster`, Nick Rabinowitz and Isaac Brodsky, key developers of `H3`, as well as the many lesser-known authors and contributors of open-source projects in the data visualization field.

Contents

1 Abstract

Epidemiological data often involves both location and time and frequently requires the simultaneous display of multiple indicators—such as the number of cases and deaths, incidence rates and population, or various disease-carrying species—to understand co-occurrence patterns. This thesis focuses on the design, creation, and evaluation of a web-based application that allows users to explore and visualize this type of data interactively. The application places particular focus on aggregation techniques to prevent clutter in the visual space while preserving the ability to visualize patterns, comparing different techniques to reach that goal. It also emphasizes cross-filtering interactions between spatial and temporal components to allow dynamic exploration across these dimensions at different resolutions. While the application is designed for broad public use, its implications for enhancing engagement in citizen science initiatives are an important consideration in this research.

2 Introduction

Epidemiological data often includes both geographical and temporal components, providing public health officials, researchers, and the media with tools to communicate the spread and impact of diseases through precise cartographic and temporal visualizations. These visualizations help inform the public about potential risks, increase awareness of disease dynamics, and promote effective countermeasures. When clear and accurate, these visual representations guide public behavior and support disease prevention and control.

The availability of recent libraries, such as `DuckDB`, `D3`, and `DeckGL`, which allow querying and visualizing large datasets directly in the browser on relatively low-end computers, offers new opportunities to bring the essence of epidemiological datasets closer to a broad audience. These advancements enable the creation of dynamic, user-friendly interfaces that allow users to explore data in real-time, providing insights that static visualizations cannot. By incorporating interactive features like zooming, panning, and filtering, modern visualization techniques enhance user engagement and deepen understanding of the data.

This research outlines the design, development, and evaluation of a novel application created to compare various visualization techniques for moderately large epidemiological datasets. The application initially focused exclusively on the user reports dataset from MosquitoAlert, a key stakeholder in this research and a citizen science project that tracks invasive mosquitoes through public participation via a mobile application. To demonstrate the generalizability of the techniques used, a second section was added to visualize a dataset containing COVID-19 indicators in France during the pandemic. The common objective of both sections is to enhance the communication of epidemiological data to the general public, making the information more accessible and comprehensible.

The development of the application adhered to several key principles, most notably the Visual Information-Seeking Mantra articulated by Shneiderman [39]: "overview first, zoom and filter, then details on demand." This approach begins by presenting a broad overview to identify patterns and trends, followed by more detailed exploration through zooming, filtering, and accessing specific details as needed. To create a system for exploring datasets in both space and time, a custom-made interactive visualization system incorporating a zoomable timeline and a zoomable map was designed and implemented. The system was then evaluated for its ability to allow users to intuitively browse the datasets.

As datasets grow larger, the gap between the vast data space and the limited visual space becomes more challenging to manage. This difference makes it harder to visualize and analyze the data effectively, requiring data aggregation techniques that are both fast and accurate. These techniques help ensure that visual representations stay clear and understandable at different levels of detail while allowing for smooth exploration. Because of this, extra attention has been given to data clustering and visualization by aggregation.

The application's aggregation techniques were evaluated and compared with existing methods, considering both algorithmic efficiency and visual effectiveness. The evaluation process included benchmark comparison for algorithmic speed and feedback from a panel of users, collected through structured questionnaires and in-depth interviews.

3 Societal Relevance and Ethics

The COVID-19 crisis highlighted how health emergencies can generate vast amounts of information, leading to an "infodemic" [48], where misinformation spreads, particularly on social media. Griffin notes that "anyone with access to the Internet and a basic computer can now make maps to serve their own interests" [18]. This emphasizes the need for data visualization creators to follow guidelines to avoid misleading the public.

When done properly, data visualization of epidemiological data can restore trust by providing clear and reliable information, raising public awareness, guiding policy decisions, and improving disease control measures. Accurate and clear visuals help the public make informed decisions, supporting public health. However, misleading visuals can cause confusion, fear, or a false sense of security. Privacy concerns also arise when handling sensitive health data, requiring proper anonymization and aggregation to maintain trust.

Data visualization may also enhance public participation in citizen science initiatives like MosquitoAlert, allowing participants to explore the datasets they help create and supporting epidemiological research.

This research addresses these ethical considerations by providing useful visualizations while protecting privacy and avoiding misleading interpretations. By following best practices in data visualization and ethics, this work aims to contribute positively to public health while ensuring accuracy and privacy.

4 Historical Background

The origins of data visualization in epidemiology can be traced back to the late 18th century when Valentine Seaman, a physicist, plotted cases of Yellow fever in his *Inquiry into the cause of the prevalence of yellow fever in New York*, in which he plotted fatal cases of Yellow fever, already noting the problem of overplotting and the problem of the inaccuracy of data: "The cases that have occurred being too numerous to attempt to get an accurate history of them all, and the want for proper marks would, at best, leave but an objectionable result."

While the method was innovative, Seaman was mistaken in his conclusions, attributing Yellow Fever to a *tertium quid*, a combination of emanations from a sick person and putrid miasmata. The debate over the cause of diseases—miasma versus germs—continued until another famous map brought decisive progress. In 1854, a severe cholera outbreak struck Soho in London, killing 616 people. John Snow, a British physician, mapped the cholera cases, identifying a cluster around a contaminated water pump on Broad Street. This visualization was decisive in demonstrating the waterborne nature of cholera, supporting the germ theory, and disproving the miasma theory.

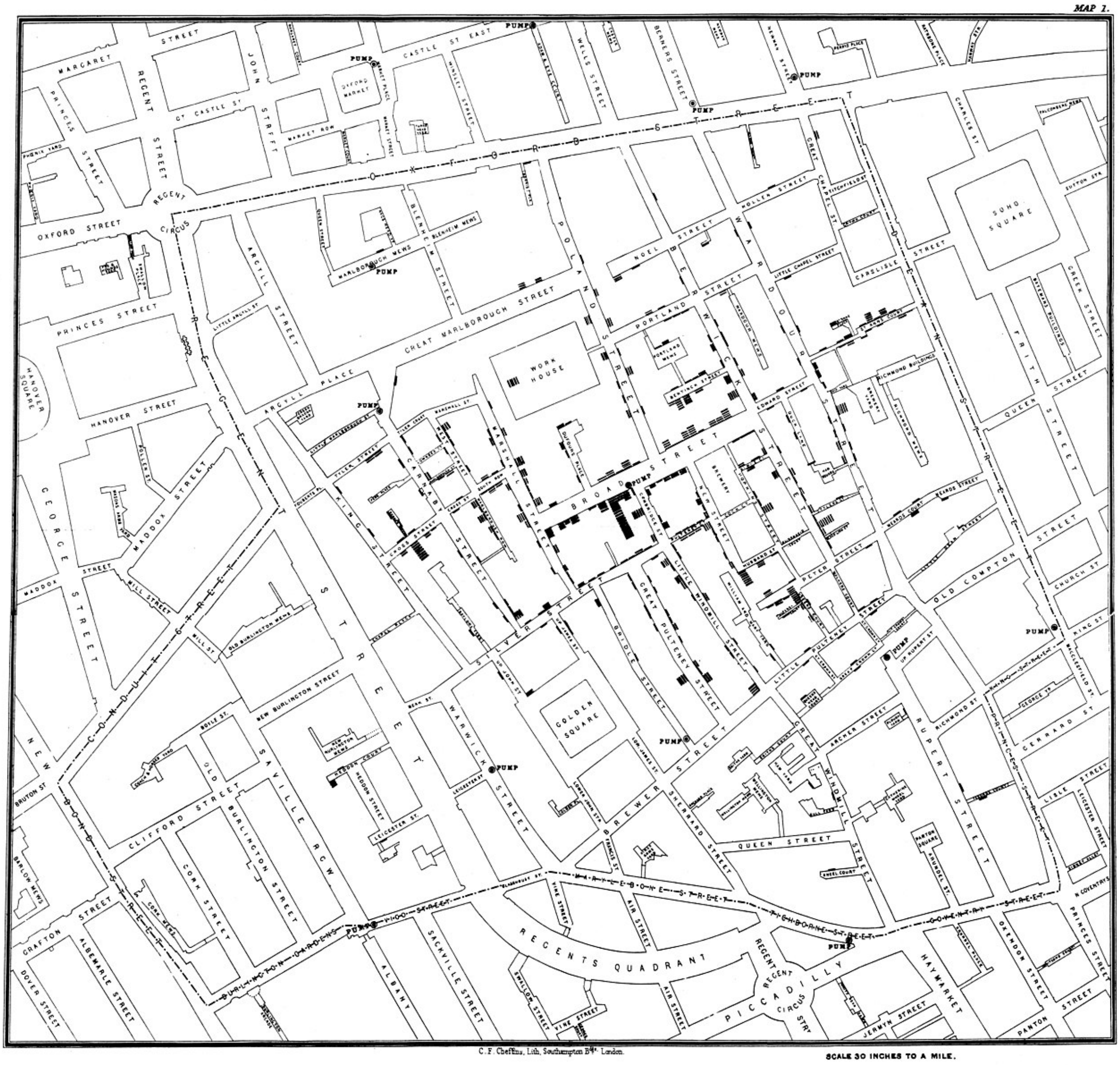


Figure 1: John Snow's map of the Soho's 1854 cholera outbreak, which was key in demonstrating the waterborne nature of Cholera

While cartographic tools were used for visualization, the plotting along the time-axis of disease indicators and potential co-factors also knew its early development at about the same period with a famous example by another physicist during the same cholera outbreak in London in an attempt to strengthen the miasma theory, where he plotted cases along time, along with meteorological condition at the same time points 2.

A few years later, in 1858, Florence Nightingale, a nurse who served during the Crimean War where she witnessed numerous deaths caused by avoidable diseases due to the lack of hygiene, created one of the most famous visualization in history. She made a polar area chart, also known as "Coxcomb" chart, using a radial time axis. This chart showed the causes of death,

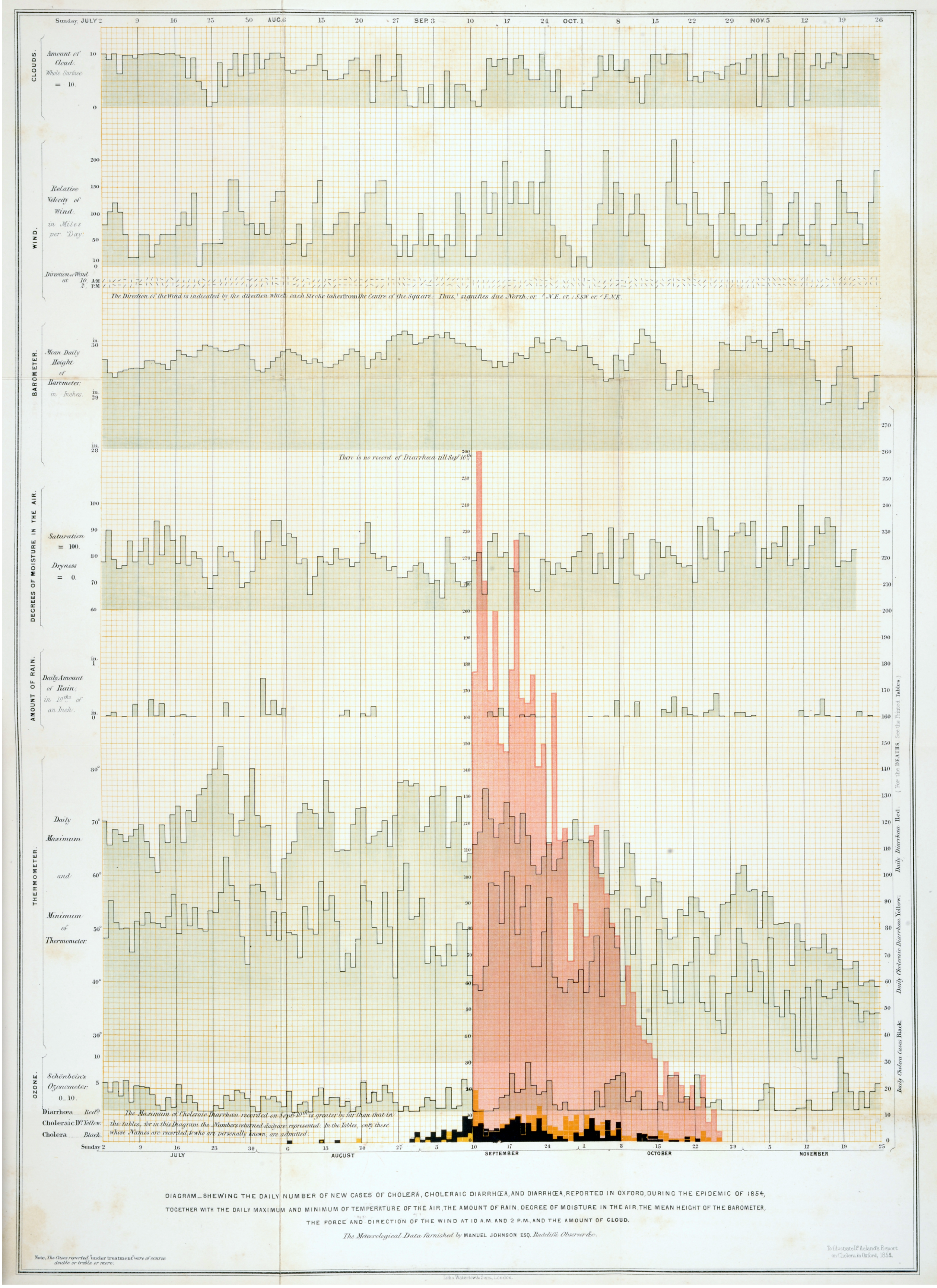
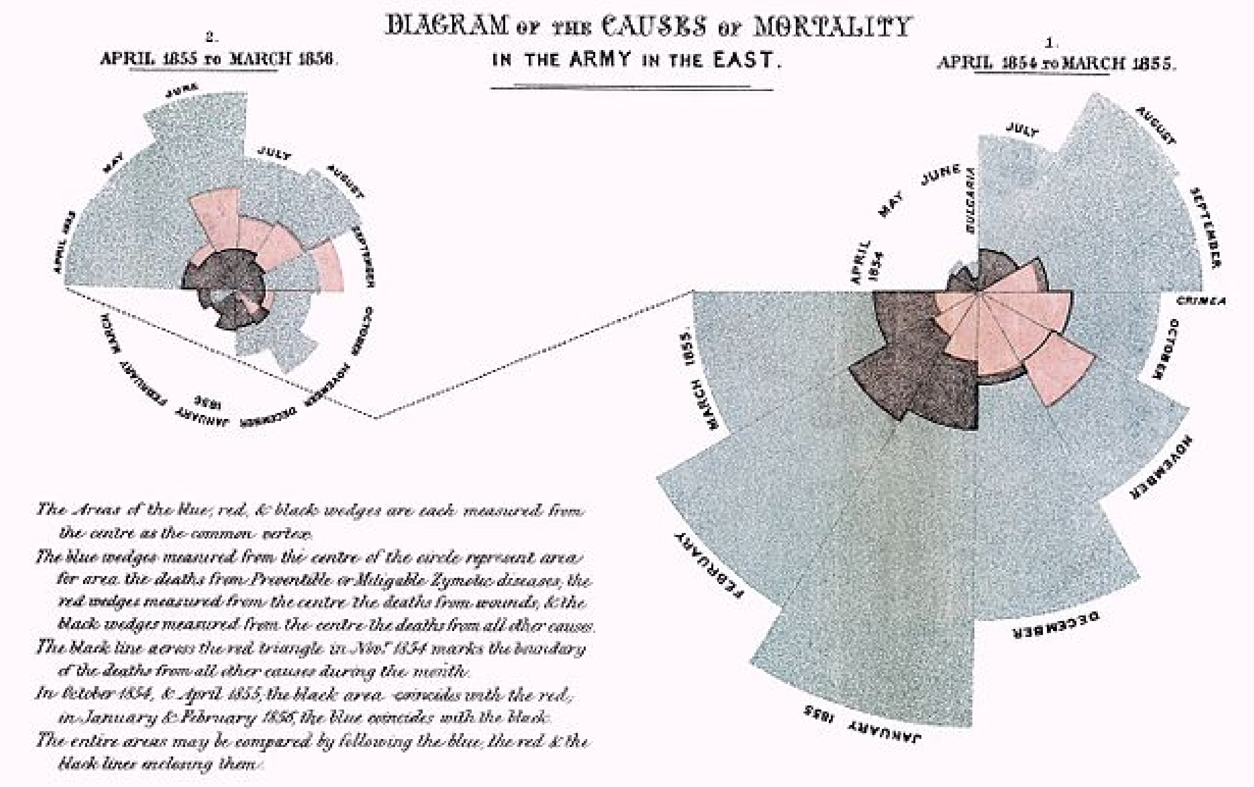


Figure 2: Acland's chart in his attempt to attribute cholera to miasma during Soho's 1854 cholera outbreak, plotting case numbers along with meteorological conditions

with stacked areas of each section representing the number of deaths per month, for each category of cause of death. It demonstrated that many deaths in the Crimean War were due to preventable diseases rather than battle wounds, highlighting the need for better hygiene and sanitation 3.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Figure 3: Nightingale's "Diagram of the causes of mortality in the army in the East" showing the causes of deaths on a radial time axis, highlighting preventable ones in blue

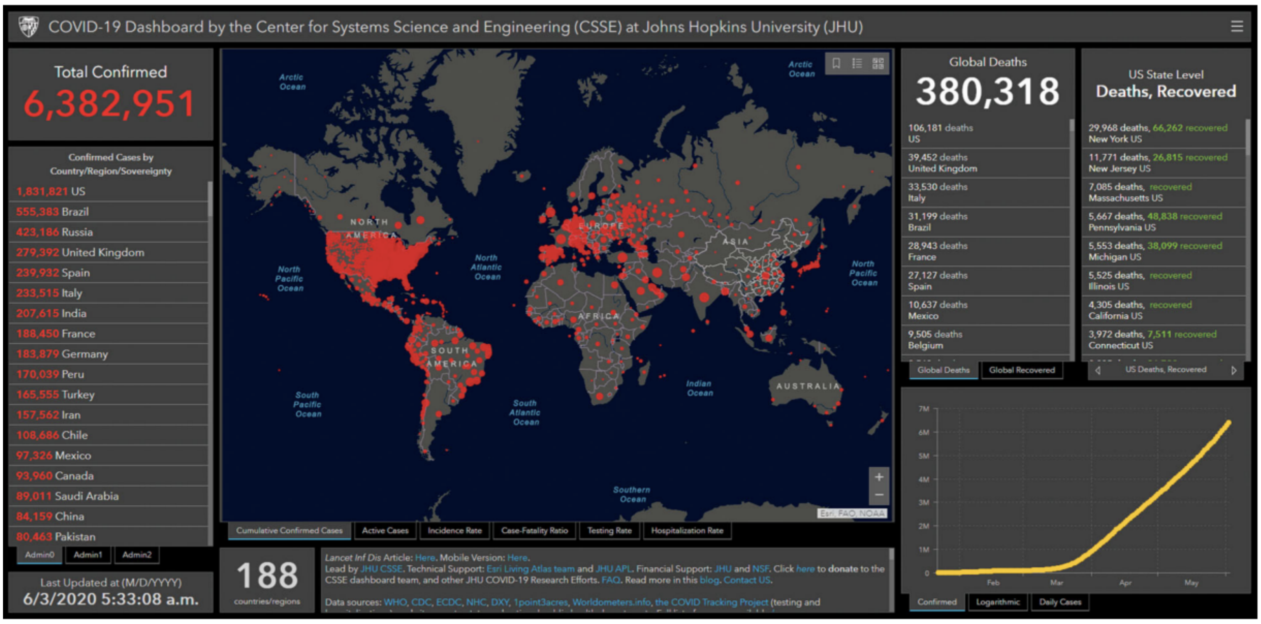
More recently, the COVID-19 crisis has probably been a defining moment for the visualization of epidemiological data [24]. Dynamic mapping methods, access to multi-scale digital maps, big data processing, and online dissemination have allowed real-time tracking of the disease's spread. The COVID-19 dashboard from Johns Hopkins University (JHU) is a prime example 4, providing an easy-to-use tool for researchers, public health authorities, and the public to monitor the outbreak. This dashboard has become a global reference for situational awareness and has been widely imitated.

5 Research Questions

The research questions for this thesis are framed within the context of disseminating information to the general public. The goal is to use widely accessible tools to reach a broad audience inclusively, while also enabling and fostering citizen science in projects like MosquitoAlert. These questions are organized into two main categories:

5.1 Design

- **Interactive Visualization:** How can we design visualizations that are effective for interactively exploring large epidemiological datasets?



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Total Confirmed
6,382,951

Confirmed Cases by Country/Region/Sovereignty

1,831,821	US
555,383	Brazil
423,186	Russia
279,392	United Kingdom
239,932	Spain
233,515	Italy
207,615	India
188,450	France
183,879	Germany
170,039	Peru
165,555	Turkey
157,562	Iran
108,686	Chile
97,326	Mexico
93,960	Canada
89,011	Saudi Arabia
84,159	China
80,463	Pakistan

Last Updated at (M/D/YYYY)
6/3/2020 5:33:08 a.m.

188
countries/regions

Lancet Inf Dis Article: Here. Mobile Version: Here.
Lead by JHU CSSE. Technical Support: Esri Living Atlas team and JHU APL. Financial Support: JHU and NSF. Click here to donate to the CSSE dashboard team, and other JHU COVID-19 Research Efforts. FAQ. Read more in this blog. Contact US.

Data sources: WHO, CDC, ECDC, NHC, DXY, 1point3acres, Worldometers.info, the COVID Tracking Project (testing and

Global Deaths
380,318

106,181 deaths	US
39,452 deaths	United Kingdom
33,530 deaths	Italy
31,199 deaths	Brazil
28,943 deaths	France
27,127 deaths	Spain
10,637 deaths	Mexico
9,505 deaths	Belgium

US State Level
Deaths, Recovered

29,968 deaths, 66,262 recovered	New York US
11,771 deaths, 26,815 recovered	New Jersey US
7,085 deaths, recovered	Massachusetts US
5,667 deaths, 48,838 recovered	Pennsylvania US
5,553 deaths, 38,099 recovered	Michigan US
5,525 deaths, recovered	Illinois US
4,305 deaths, recovered	California US
3,972 deaths, 7,511 recovered	Connecticut US

Figure 4: Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

- **Pattern Identification:** How can we identify and highlight geospatial and temporal patterns at different stages of epidemiological events? Specifically, how can emerging threats and weak signals be effectively visualized alongside patterns that occur when those threats evolve into events at a much larger scale?

5.2 Technologies

- **Responsive Interaction:** What technologies are most appropriate for providing fluid interaction with epidemiological data, ensuring quick response times to user inputs?

- **Scalability:** How can these technologies be scaled to handle increasingly large datasets without compromising performance or user experience?

- **Integration:** How can these technologies be integrated into a web-based platform to allow fluid interaction and exploration of geospatial and temporal data related to the field of epidemiology by a wide audience?

6 Methodology

This section provides an overview of the datasets, the visual design of the application, the key tools and techniques employed in its development, and the methods used for its evaluation. The following subsections place the application within the context of research on interactive data visualization, with a particular focus on epidemiological data.

6.1 Datasets

Two main datasets were used in this study based on their availability, relevance to the stakeholders of this study, and representativeness for the techniques proposed in the thesis. Other datasets were used to provide additional context and are described as well.

6.1.1 MosquitoAlert Reports Dataset

MosquitoAlert [23] is a citizen science project and mobile application designed to monitor the progression of invasive mosquito species, particularly those that are vectors for diseases such as dengue, Zika, and chikungunya. Launched in 2014, this initiative is a collaboration between several research institutions: CREAF (Centre for Ecological Research and Forest Applications), the Pompeu Fabra University (UPF), ICREA (Catalan Institution for Research and Advanced Studies), and the CEAB-CSIC (Center for Advanced Studies of Blanes) and public health organizations. The primary goals of *MosquitoAlert* are to engage the public in scientific research to enhance mosquito surveillance, and support public health efforts in controlling mosquito-borne diseases.

The *MosquitoAlert* mobile application allows users to report sightings and bites of mosquitoes, as well as breeding sites such as stagnant water sources. Users can upload photos and provide information about the location and time of their observations. These reports are then verified by a team of experts, and the validated data is used to create real-time maps and models of mosquito distribution and activity [33].

6.1.2 COVID-19 France Dataset

The dataset contains daily data for several COVID-19 indicators in France gathered by *Santé publique France*, the French public health agency. It contains several columns, of which the ones used in the application are:

Contextual Data

- **date**: Date
- **dep**: Department

Data Related to Hospital Situation

- **hosp**: Number of patients currently hospitalized for COVID-19.
- **incid_hosp**: Number of new patients hospitalized in the last 24 hours.
- **rea**: Number of patients currently in intensive care units (ICU).
- **incid_rea**: Number of new patients admitted to ICU in the last 24 hours.

Data Related to COVID-19 Deaths

- **dchosp**: Hospital deaths.
- **incid_dchosp**: New hospital deaths in the last 24 hours.
- **esms_dc**: Deaths in ESMS (medico-social establishments).
- **dc_tot**: Cumulative deaths (total deaths recorded in hospitals and EMS).

6.1.3 Kontur Population Dataset

The original *Kontur* Population dataset uses H3 hexagons (cf. 6.3.1) at resolution 9, each with a radius of approximately 400 meters, to represent population counts across the globe. These population estimates are primarily based on the Global Human Settlement Layer (GHSL) [2], with additional input from Facebook's High-Resolution Settlement Layer (HRSL) [1] where available. To improve the accuracy of the distribution, the dataset incorporates additional sources like Microsoft Building Footprint, Land Information New Zealand, and Copernicus Global Land Service. Dasymetric techniques are used to correct known artifacts in the GHSL and HRSL datasets by marking unpopulated areas such as quarries, major roads, lakes, and forests using OpenStreetMap data. Extremely populated cells are adjusted by redistributing their populations to neighboring cells, and population counts are rounded to the nearest whole number for precision. The dataset is updated each year.

Kontur also offers separate datasets per country. The dataset used in this application is for France, aggregated from H3 resolution 9 to 7, reducing the size by a factor of approximately $7 \times 7 \equiv 49$ for manageability in the front end. The processed dataset contains 111,179 rows, fully covering continental France and Corsica with hexagons and associated population estimates.

To allow joining with COVID-19 data at the department level, the GeoJSON shapes of the French departments were used to link the hexagons to their containing department through the use of the `geojson2h3` utility maintained by Uber [42].

6.2 Visual Design

6.2.1 General Design

The application is designed with a large map that takes up most of the screen space. This decision is informed by research showing that users prefer large maps because they can see more elements at once [40]. A zoomable timeline is also displayed at the bottom, allowing the user to explore the datasets in both time and space.

The design was inspired by `Timemap.js`, a project by Nick Rabinowitz, who is also a key developer of H3, a geospatial indexing system used in this research and described in further detail in 6.3.1.

`Timemap.js` combined a map with a timeline from the Semantic Interoperability of Metadata and Information in Unlike Environments (SIMILE) [21], a project initiated in 2005 by the Massachusetts Institute of Technology. The project aimed to provide browser-based tools for the exploration and visualization of datasets. The timeline allowed users to pan indefinitely and click on data points represented as circle symbols to view detailed information. Although both projects are no longer maintained, `Timemap.js` introduced a powerful method of interacting with data by integrating a timeline and a map.

This combination of a timeline and a map offers a simple and effective way to explore geospatial data over time, enabling users to cross-filter and view data at different levels of detail. While literature is sparse about such a solution, one maintained project that offers similar functionality is *xyzt.ai* [46], a platform for visualizing large datasets in space and time developed by a startup company coincidentally based in Hasselt.

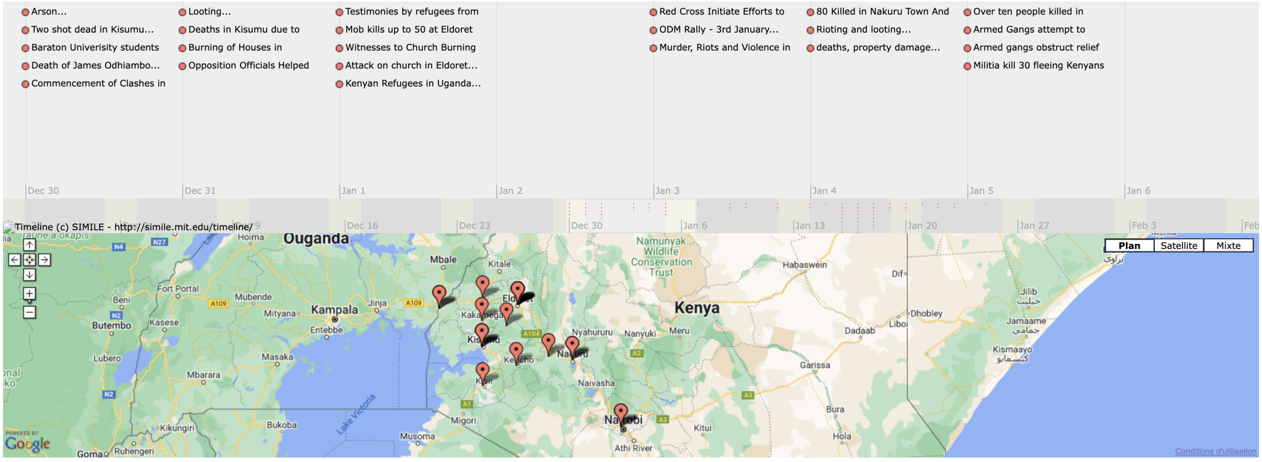

Arson...
Two shot dead in Kisumu...
Baraton Univerisity students
Death of James Odhiambo...
Commencement of Clashes in
Looting...
Deaths in Kisumu due to
Burning of Houses in
Opposition Officials Helped
Testimonies by refugees from
Mob kills up to 50 at Eldoret
Witnesses to Church Burning
Attack on church in Eldoret...
Kenyan Refugees in Uganda...
Red Cross Initiate Efforts to
ODM Rally - 3rd January...
Murder, Riots and Violence in
80 Killed in Nakuru Town And
Rioting and looting...
deaths, property damage...
Over ten people killed in
Armed Gangs attempt to
Armed gangs obstruct relief
Militia kill 30 fleeing Kenyans
Dec 30
Dec 31
Jan 1
Jan 2
Jan 3
Jan 4
Jan 5
Jan 6
Timeline (c) SIMILE - http://simile.mit.edu/timeline/
Dec 16
Dec 23
Dec 30
Jan 6
Jan 13
Jan 20
Jan 27
Feb 3
Plan
Satellite
Mixte
Ouganda
Kenya
Kampala
Hoima
Mbale
Kitale
Jinja
Entebbe
Fort Portal
Mubende
Mityana
Beni
Butembo
Kasese
Lubero
Masaka
Mbarara
Bukoba
Kikungiri
Goma
Ruhengeri
Musoma
Migori
Kisii
Kericho
Kakamega
Nakuru
Naivasha
Nyahururu
Nanyuki
Meru
Nairobi
Athi River
Kitui
Namunyak Wildlife Conservation Trust
Habaswein
Dadaab
Garissa
Bura
Hola
Dhobley
Jilib
Jamaame
Kismaayo
Lake Victoria
Google
Conditions d'utilisation

Figure 5: Timemap.js, a project from 2008 by Nick Rabinowitz [36], which inspired the layout and interactive features of this application.

6.2.2 Visual Encoding

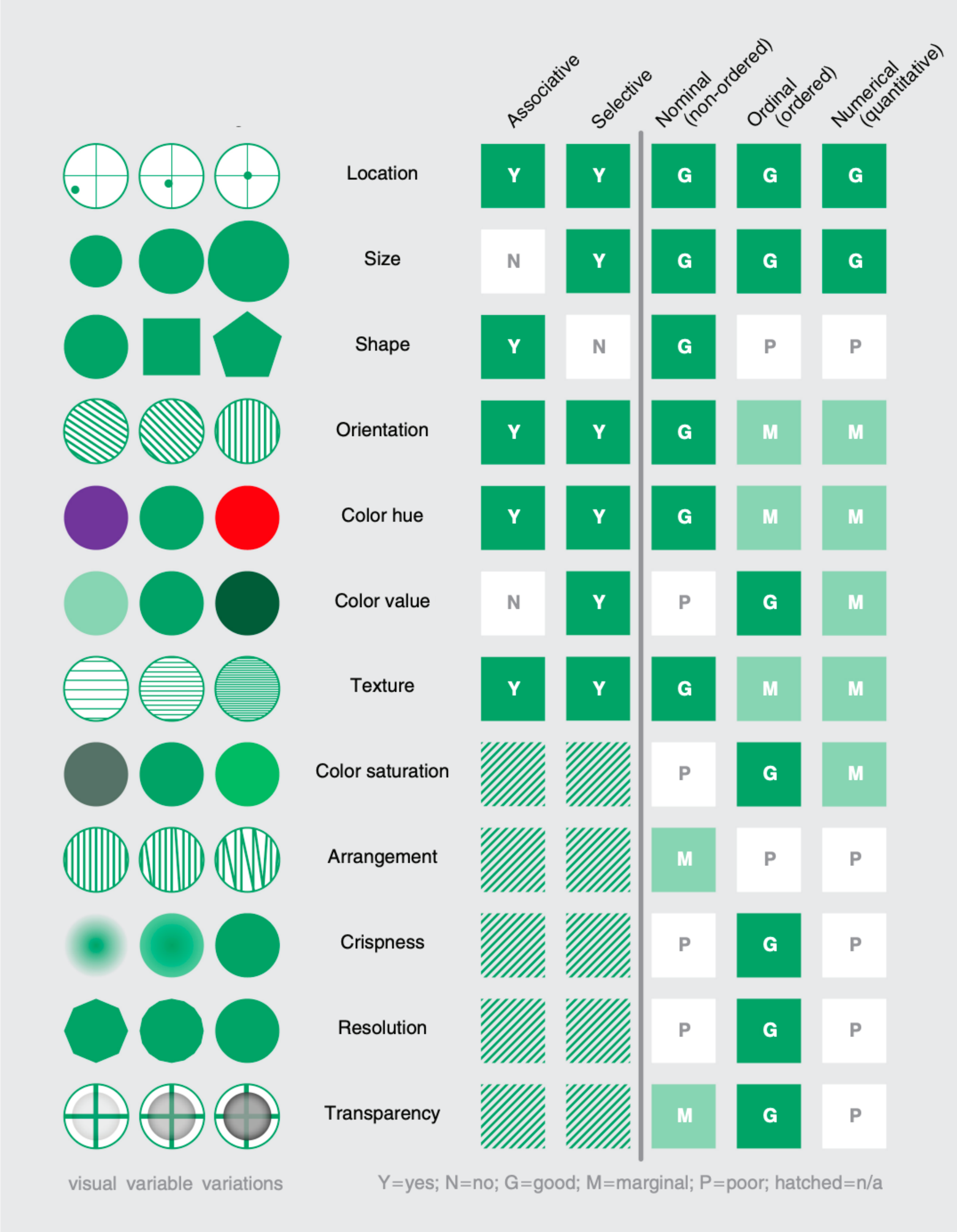
When creating a visualization, the data visualization creator faces numerous options to encode the data in a meaningful and expressive way. The visual variables describe the graphic dimensions across which the symbols or glyphs of a visualization can be varied to encode information. These visual variables were originally described by French cartographer and professor Jacques Bertin (1918–2010) in his 1967 book *Sémiologie Graphique*. The English translation, *Semiology of Graphics*, was released in 1983 and is now recognized as a seminal theoretical work in both cartography and the broader field of information visualization [5].

Bertin's work was inspired by semiotics, the study of sign systems, which seeks to understand how one object comes to stand for another. This framework data visualization creators to think about how a symbol (the signifier) representing a phenomenon (the signified) communicates meaning to the visualization user (the interpretant). By breaking down visualizations into their basic graphic elements, unclear or ambiguous symbols can be identified and improved, enhancing communication and comprehension.

Bertin initially identified seven visual variables, termed *variables rétiniennes* in French (retinal variables): location, size, shape, orientation, color hue, color value, and texture. This initial set was later expanded by Morrison in 1974 [32] to include color saturation and arrangement. Finally MacEachren extended the set further to include crispness, resolution, and transparency, as described in his work [28], to accommodate new methods of encoding information enabled by digital displays.

Bertin defined variables that can be perceived as belonging to the same group when they share the same variation as associative and variables that allow the eye to focus individually upon each variation of the visual variable as selective. A summary of the visual variables and their associative and selective qualities, along with their ability to encode different types of variables (nominal, ordinal, and numerical) is shown in 6.

Each variable plays a distinct role in how information is encoded and interpreted by the viewer. Below is a brief overview of their usage in the application:



	Associative	Selective	Nominal (non-ordered)	Ordinal (ordered)	Numerical (quantitative)
Location	Y	Y	G	G	G
Size	N	Y	G	G	G
Shape	Y	N	G	P	P
Orientation	Y	Y	G	M	M
Color hue	Y	Y	G	M	M
Color value	N	Y	P	G	M
Texture	Y	Y	G	M	M
Color saturation	▨	▨	P	G	M
Arrangement	▨	▨	M	P	P
Crispness	▨	▨	P	G	P
Resolution	▨	▨	P	G	P
Transparency	▨	▨	M	G	P

visual variable variations

Y=yes; N=no; G=good; M=marginal; P=poor; hatched=n/a

Figure 6: Visual variables, adapted from an article on visual variables by Roth [37]

Location Location indicates the visualization symbol's position relative to a coordinate frame. In clustered visualizations (cf. 6.3.5), the location can be defined based on a centroid calculation based on the underlying clustered variables.

Size Size refers to the space occupied by a visualization symbol. Larger symbols or a high density of symbols can result in clutter and occlusion, potentially diminishing the clarity of the visualization. When symbols are used on a map, their size may be distorted due to the need to project the Earth's spherical surface onto a plane. In the application described in this thesis, a Mercator projection is used to display both the symbols and the map, which can affect the perceived size of symbols relative to each other, especially at lower zoom levels.

Shape Shape defines the external form or outline of a symbol. It is important for qualitative symbols that can then be combined with another visual variable (for instance color). Shapes can be combined. In the web application, we experiment with two types of primary shapes due to the ease of representing them at scale with the GPU-based visualization framework used in the application (`Deck.gl`): rectangle shapes combined in mini square stacked bars and packed circles using the front chain packing algorithm described by Wang et al. in [44].

Orientation Orientation describes the direction or rotation of a symbol. It is used in multivariate glyph symbols and flow maps to indicate directionality but was not used in this research.

Color Hue Color hue refers to the dominant wavelength of a color (e.g., blue, green, red) and is commonly used in choropleth maps and categorical data representations. However, color hue should be used with care to ensure inclusiveness, as a percentage of the population is affected by color blindness and may not perceive certain colors. The color schemes used in this application were drawn from ColorBrewer [19]. In the MosquitoAlert section, color was employed to encode categorical variables, specifically different mosquito species.

Color Value Color value refers to the lightness or darkness of a color, and it is particularly useful for representing a range of ordinal or numerical values in choropleth maps or in combination with other visual variables. In the application, color value is used to encode COVID-19 indicators in both choropleth and proportional symbols maps. However, using color to represent quantitative indicators in an interactive application presents significant challenges. The scale of these indicators, such as incidence and prevalence, can vary dramatically between the peaks and troughs of an epidemic, spanning multiple orders of magnitude. At the same time, detecting weak signals is important for identifying the early stages of an outbreak.

Texture Texture describes the fill pattern's coarseness within a symbol. Previously used in halftone techniques, we show later in 6.3.5 that grid clustering can be used to produce texture patterns similar to halftone to encode multivariate data as described by Bertin [5].

Color Saturation Color saturation indicates the intensity or purity of a color. It wasn't used in the application.

Arrangement Arrangement describes the layout of graphic symbols. It varies from regular to irregular patterns. When data is clustered in the data space and displayed in the visual space, arrangement will depend on the clustering algorithm used and may appear more random when traditional clustering algorithms are used or more organized when the centroids of cells in a grid are used as will be discussed in 6.3.5.

Crispness Crispness indicates the sharpness of a symbol's boundary, with a gradient in opacity around the borders. Crispness can be used to highlight uncertainty about the displayed data. MacEachren et al. [29] show crispness to be a very effective visual variable for representing uncertainty. It is not used in the application.

Resolution Resolution describes the spatial precision of a symbol's display. As is detailed in 6.3.5, the resolution of the symbols can be defined by parametrization of the clustering algorithm.

Transparency Transparency refers to the degree of blending between a symbol and its background. While transparency can help alleviate occlusion by allowing overlapping symbols to remain partially visible, it is not a scalable solution as the number of symbols increases—occlusion will still occur. Additionally, when symbols of different colors overlap, transparency necessitates color blending. This can create challenges in interpreting categories encoded with different color hues, as the overlapping semi-transparent symbols may blend into entirely new colors, potentially obscuring the distinctions between different categories in multicategorical data.

6.2.3 Thematic Maps

Thematic maps show how a particular subject or theme is distributed across a geographic area. Since there are many types of thematic maps, this research doesn't cover all of them. Instead, this section focuses on the specific types of thematic maps used in this study for comparison. One key feature of the maps used is that they maintain the true shapes and sizes of the areas they represent, except for the necessary projection (Mercator) from a sphere to the 2-D space of a user's screen. This is different from cartograms, which are not considered here because they distort areas to show variables, making them incompatible with popular map services like `Mapbox`, which is used in this research.

Choropleth Choropleth maps are a type of thematic map where geographic regions are colored (with uniform colors) or patterned in proportion to a statistical variable that represents an aggregate summary of a geographic characteristic within each region. The term "choropleth" comes from the Greek words "choros" (area/region) and "plethos" (multitude). Choropleth maps have a long history, with early examples dating back to the 19th century, such as Charles Dupin's 1826 map of literacy in France 7. In data science, choropleth maps are widely used for visualizing spatial data to reveal patterns and insights related to demographics, economics, public health, and environmental data.

Despite their widespread use and utility, choropleth maps have several limitations. As Ward et al. state, ”A problem of choropleth maps is that the most interesting values are often concentrated in densely populated areas with small and barely visible polygons, and less interesting values are spread out over sparsely populated areas with large and visually dominating polygons” [45]. Choropleth maps shade entire regions based on aggregate data, which in the context of epidemiological data can obscure the actual distribution of the population within those regions. This is known as the modifiable areal unit, as exemplified in 8

These critiques highlight the importance of careful design and complementary use of other visualization techniques, such as dot density maps or heat maps, to provide a more comprehensive understanding of spatial data.

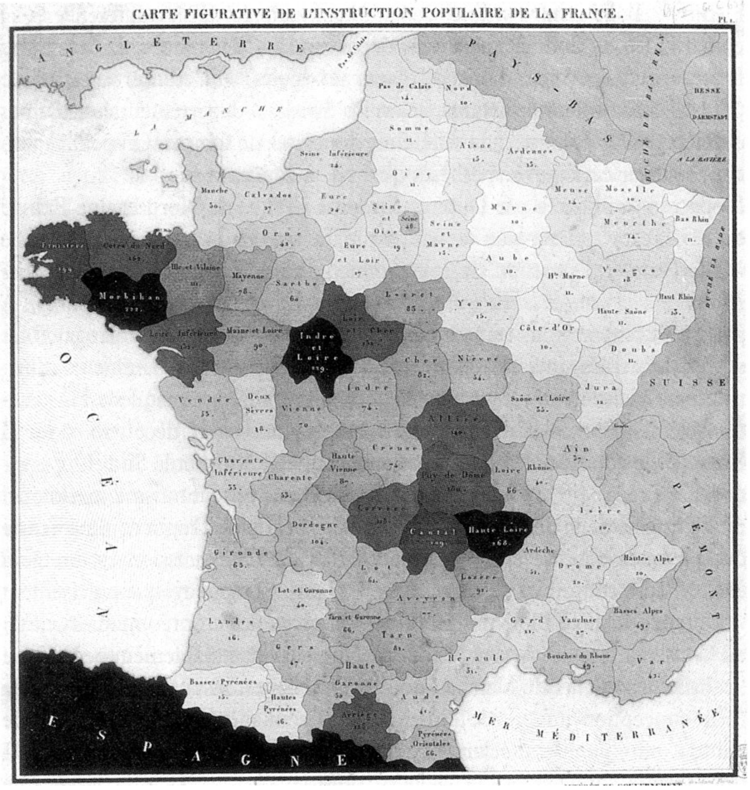


Figure 7: Charles Dupin's 1826 map of literacy in France, an early example of a choropleth map.

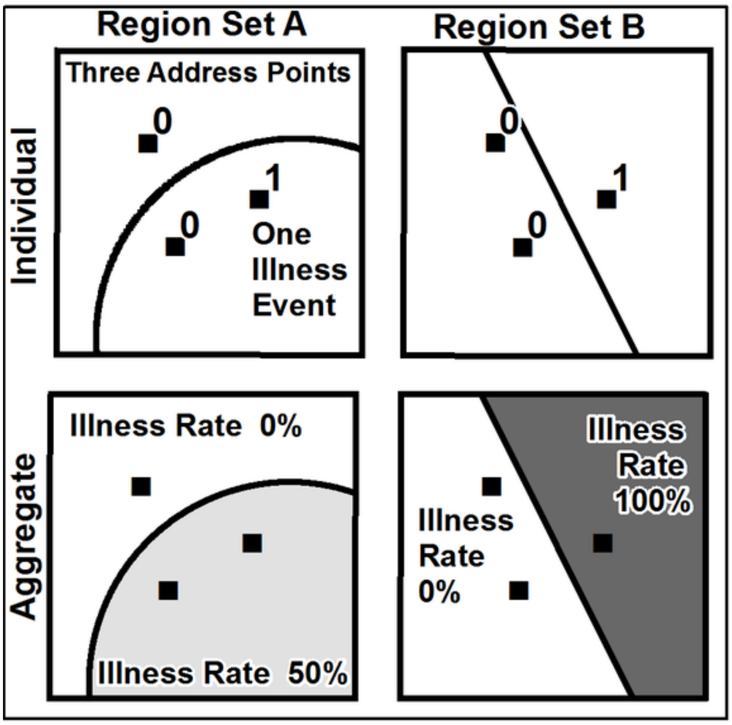


Figure 8: Modifiable areal unit problem, image from Wikipedia.

Dot Distribution Map A dot distribution map can be used when precise locations, such as latitude and longitude, are available in the dataset, as with the *MosquitoAlert* dataset. In this type of map, dots representing individual observations are plotted as an additional layer, visually indicating the specific locations of these observations.

When the number of dots increases, occlusion becomes a concern, as dots can overlap and hide each other, obscuring the data. This issue can be mitigated by using smaller dots, applying transparency to the dots, or introducing some jitter. Jittering involves slightly adjusting the position of each dot to reduce overlap while maintaining the overall spatial pattern. This technique ensures that each data point remains visible, enhancing the map's readability and effectiveness in conveying information.

With interactive maps, the radius of the dots can be dynamically adjusted based on the zoom level. This can be done in several ways:

- Scaling with Zoom Level: The radius can be modified according to the zoom level, so dots appear larger or smaller as the user zooms in or out. This helps maintain visibility and proportional representation at different scales.

- Constant in Screen Space: The radius can be kept constant in screen space, measured in pixels. This means the dots will appear the same size on the screen regardless of zoom level, which is useful for maintaining consistent visual prominence.

- Constant in Physical Space: The radius can be kept constant in physical space, measured in meters. This ensures that the dots represent the same physical area on the ground, which is useful for geographic accuracy and consistency.

Proportional Symbol Map A proportional symbol map is a map that uses symbols, often circles or simple regular shapes, to represent a specific variable. The size of each symbol changes according to the value it represents; for example, larger circles might indicate cities with bigger populations. Typically, the symbol size is calculated to match the data accurately, often using a square transformation to determine the side length or radius of the shape.

One challenge with proportional symbol maps is managing the range of scales as noted by Mocnick et al. [31]. When some areas have very large values, their symbols can overlap or dwarf those in areas with smaller values. However, representing weak signals is also important.

Multi-categorical data can be visualized by using different colors for the symbols, allowing multiple variables to be shown at the same location. This approach is most effective when the variables are of similar magnitudes, as using different scales for different indicators can cause confusion, as highlighted by Mocnik et al. in their COVID-19 visualizations [31].

In the MosquitoAlert section of the application, the issue of scale was addressed by adjusting the symbol sizes based on the context, with the largest value within the visible map area determining the maximum symbol size to prevent overlap. This allows to reveal patterns in areas with low numbers through zoom interaction. Additionally, the mere presence of data with very low values was indicated by semi-transparent shapes, allowing weaker signals to be highlighted.

Supplementation of Chloropleth with Proportional Symbol Map In this subsection, we present a visual experiment where the traditional choropleth map is replaced or supplemented by a proportional symbol map layer to represent the French population distribution along with COVID-19 indicators. This approach is intended to mitigate the limitations of choropleth maps in accurately representing population densities.

We use proportional circles within the H3 grid system described in section ?? to highlight population density. This method visualizes population distribution by varying the size of the circles in proportion to the population within each H3 cell. Color hue can then be used to encode an epidemiological indicator.

To create the proportional symbol map, we first obtained detailed population data for France, from the open dataset from Kontur.

Choosing an appropriate resolution for the grid is important. Higher resolutions provide more detail but may increase visual complexity, and add an additional cost in terms of performance when joining multiple datasets.

We calculate the radius of the circles based on the population data, ensuring that the circles are proportional to the population size in each H3 cell. The proportional symbols must be scaled appropriately to be visually distinct while avoiding excessive overlap.

Due to the significant diversity in the French population, with vast rural areas of low population density and a few urban centers of very high population density, a zoomed-out view of the map displays very small circles for rural areas, while circles in urban centers tend to overlap and obscure each other. To mitigate this issue, a variable radius is applied to the circles: as the user zooms in, the circles are rescaled, reducing overlap and revealing detailed population distribution patterns. This dynamic adjustment ensures that population densities in both rural and urban areas are accurately and clearly represented at different zoom levels. This approach was evaluated and compared to the traditional choropleth visualization.

6.2.4 Representation of Time

Time is presented at different resolutions in the two main datasets used in this research. In the MosquitoAlert dataset, time is recorded with high precision as a timestamp, whereas in the COVID-19 dataset, time is grouped into 24-hour periods (dates). To facilitate navigation through these datasets at varying time resolutions and allow for easy adjustment of the time range, a zoomable timeline was created in D3. This timeline dynamically aggregates time-based on standard divisions (year, month, week, day, hour), adjusting the time buckets according to the zoom level.

In an interactive application, time adds another layer of complexity when visually encoding variables on a map. This is especially true for the COVID-19 indicators datasets for France, where disease indicators span multiple orders of magnitude through peaks and troughs. A uniform color scheme across the entire dataset would fail to provide sufficient contrast during the wave's low points and early stages, making it difficult to identify where a wave begins.

6.2.5 Interactivity and Animation

In the early days of the personal computer revolution, Bertin emphasized the importance of interaction in graphics: "This is a fundamental point, because it is the internal mobility of the image which characterizes modern graphics. A graphic is no longer 'drawn' once and for all; it is 'constructed' and reconstructed (manipulated) until all the relationships which lie within it have been perceived" [6, p. 5].

In the application described here, zooming plays an important role in the visual exploration of the datasets both spatially and temporally. Google Maps has popularized the zoom interaction with scrolling, centering the zoom at the position of the cursor. It has since become a very intuitive way of navigating online maps as studied by You et al. [47], later imitated by other Web Map Services (WMS). To propose a similar experience in the temporal dimension, a zoomable timeline was built with D3 and synchronized with the map allowing the user to cross-filter the dataset in space and time in the MosquitoAlert section. In the COVID-19 section, clicking on specific periods on a temporal x-axis bar chart on the timeline displays the geographical distribution of the indicator on the map for the given period.

Checkboxes are also present in both sections, to allow filtering by species in the MosquitoAlert section, and to select an indicator in the COVID-19 France section

Speed is another important consideration in interactive applications. As Heer noted, "To be most effective, visual analytics tools must support the fluent and flexible use of visualizations at rates resonant with the pace of human thought" [20]. This research addressed performance concerns by selecting optimal libraries and clustering algorithms to ensure responsive interactions.

An experiment was also conducted, enabling the visualization of the distribution of COVID-19 epidemiological indicators over time on a map of France, with each day or week displayed in sequence, producing a movie-like effect that reveals temporal and geographical patterns.

Both interaction and animation aspects of the application were evaluated with a panel of users.

6.3 Tools and Techniques

6.3.1 H3 Geospatial Indexing System

The `H3` geospatial indexing system is a discrete global grid system that features a multi-resolution hexagonal tiling of the Earth sphere with hierarchical indexes. The `H3` grid is created by mapping a regular hexagon grid onto the faces of an icosahedron and then using inverse gnomonic projection to project these faces onto the Earth's spherical surface [38]. It is impossible to grid the icosahedron only with hexagons, so 12 pentagons are also present, centered on the vertices of the icosahedron. At each resolution except the finest, the 12 pentagons are subdivided into 5 hexagons around a smaller pentagon at the center and the hexagons are subdivided into 7 smaller hexagons to achieve the next resolution level (aperture 7). Due to the effect of the inverse gnomonic projection and the presence of pentagons, the size of the cells slightly differ across the grid, the 12 pentagons are smaller than the hexagons and the hexagons increase in size as a function of the distance from the pentagons [26].

There are only three regular tessellations of the Euclidean plane 10, which are ways of tiling the plane with regular convex shapes of the same size. The shapes are the triangle, the square, and the hexagon. Among these three shapes, the hexagon possesses unique advantages when compared to the two alternatives:

1. The distance is uniform between a hexagon's center and the centers of its six neighboring hexagons.

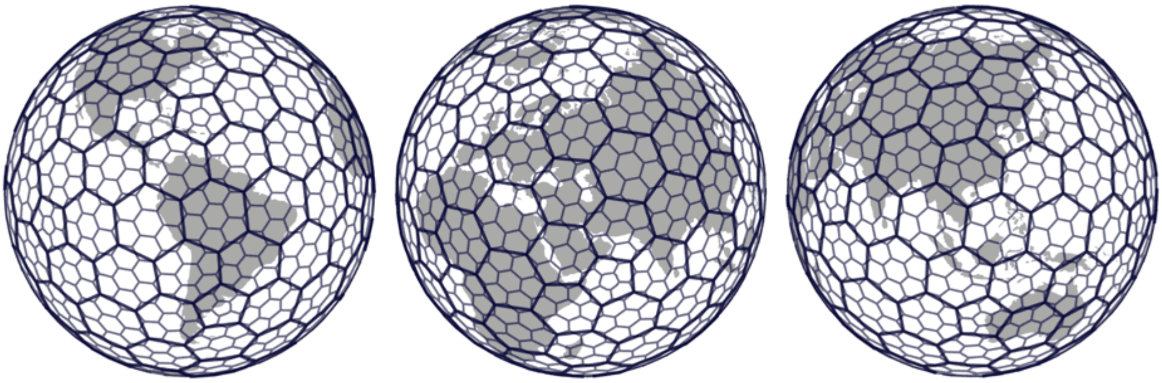


Figure 9: H3, developed by Uber, divides the Earth into hexagons and 12 pentagons. Image: Uber.

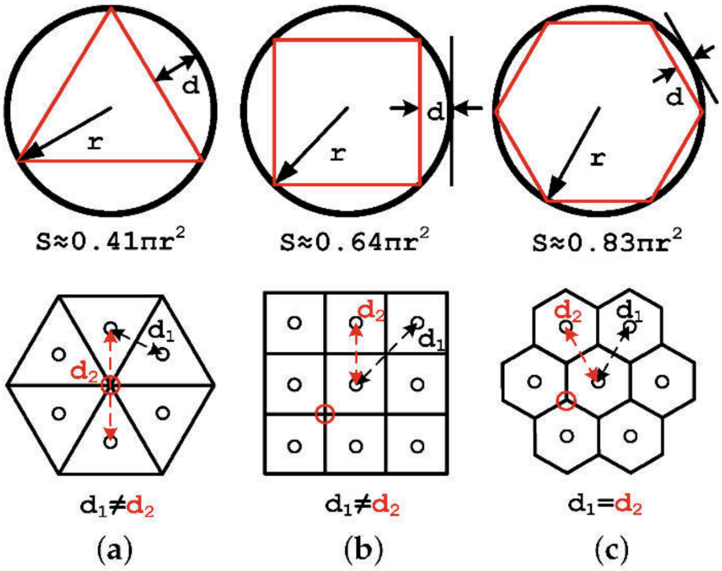


Figure 10: The three regular tessellations of the Euclidean plane and their properties, image from [4].

2. The ratio between a hexagon and its circumscribed circle is the largest among the three shapes.

The hexagonal grid is also the densest packing of circles in Euclidean space, a property which can be exploited by the use of circles as map symbols to represent the values of data points or clustered data points in a way similar to halftone patterns 11.

The H3 hexagonal hierarchical grid system does not cleanly divide cells into seven smaller hexagons between levels of the hierarchy and containment of children hexagons is not perfect but provides a good approximation 12. This enables truncation of precision within a fixed margin of error through bitwise operation for an H3 index and allows the determination of all child indexes from a parent index.

Approximate containment occurs only when truncating the precision of an index, but exact

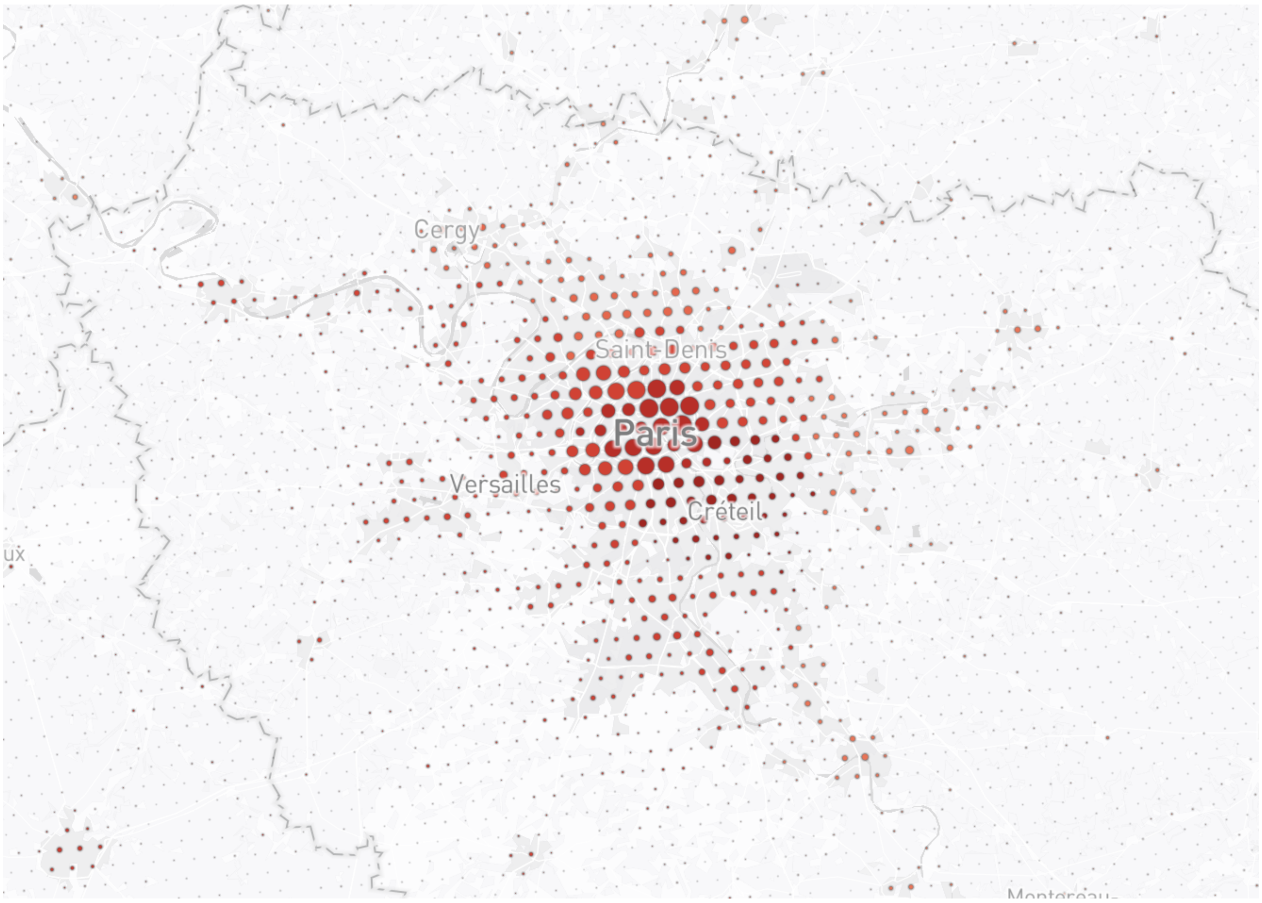


Figure 11: Hexagonal packing of circle symbols encoding values (population) with `H3`, producing an arrangement reminiscent of haltone patterns

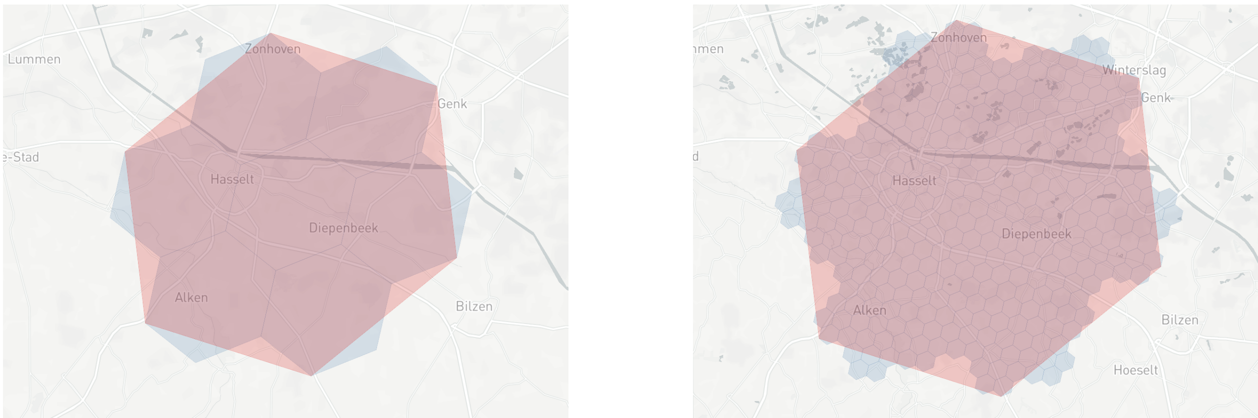


Figure 12: `H3` cell non-containment. At > 2 resolutions finer than the parent, a few entire cells lie outside the parent's boundary.

containment can also be obtained by re-indexing data points from their longitude and latitude at a given resolution.

Functions for changing precision (`h3ToParent`, `h3ToChildren`) are quick due to simple bitwise operations (masking of least significant bits), and geographically close locations typically have numerically close indexes.

In the application described in this thesis, `H3` was used to index the geolocated datasets at the lowest resolution, with about one-meter precision, replacing the traditional longitude and latitude coordinates. This approach enabled the aggregation and joining of multiple datasets using a unified indexing system for geospatial data, thereby allowing the visualization of these datasets simultaneously through various layers.

One particular use case of H3 investigated in this research is its use for aggregation, taking advantage of the hierarchical structure of the index as will be detailed in 6.3.5.

6.3.2 DuckDB-Wasm

DuckDB-Wasm is an in-process analytical SQL database running in the browser, a port to WebAssembly of the original implementation in C++. WebAssembly (Wasm) is a binary instruction format for a stack-based virtual machine, designed as a portable target for compilation of high-level languages like C, C++, and Rust. It is intended to execute code at near-native speed within the browser.

Powered by WebAssembly, DuckDB-Wasm efficiently handles SQL queries within the browser environment, offering compatibility with various data formats such as Parquet, CSV, and JSON through Filesystem APIs or HTTP requests. This can eliminate or supplement the use of a server-based API, enabling direct interaction with data files from the client side, thus reducing latency and enhancing performance for data-intensive applications relying on the client for computations. A special extension of DuckDB for the manipulation of H3 indexes [8] is used in the application, allowing aggregation and conversion of H3 indexes to latitude and longitude coordinates as a preparation step for the visualization of H3 indexed data points on a map.

At the time of this writing, an official build of DuckDB-Wasm allowing the loading of unsigned extensions was not available from the maintainer's repository. A custom build was made as part of this thesis research, allowing loading of the DuckDB-H3 extension [14].

6.3.3 D3.js

D3.js or simply D3 (Data-Driven Documents) is a JavaScript library developed by Mike Bostock [7] for producing dynamic, interactive data visualizations in web browsers. It leverages modern web standards such as SVG or Canvas, allowing developers to bind data to a Document Object Model (DOM) and apply data-driven transformations to the document. D3's powerful data manipulation and visualization capabilities make it a popular choice for creating custom visualization in the browser. D3 also incorporate functionalities to introduce interactivity such as zoom and pan operations.

In the context of the application described here, D3 was used to develop a zoomable timeline, also using Observable Plot, an abstraction layer over D3 facilitating the creation of plots.

6.3.4 Deck.gl

Deck.gl is an advanced WebGL-powered framework for visualizing large-scale datasets directly in the browser, especially useful for thematic map layers. Developed by Uber, Deck.gl is designed to work with modern map rendering libraries like Mapbox GL and Google Maps, providing a rich set of layers for geographic data visualization. In the app, Mapbox was chosen as map base rendering library.

When working with `H3` indexes, `Deck.gl`'s `H3Layer` can be used to render hexagonal grids efficiently, enabling the visualization of large-scale geospatial data in an interactive manner. `Deck.gl`'s performance optimizations and integration with `WebGL` make it particularly suited for applications that require smooth rendering of complex datasets, even on lower-end hardware.

6.3.5 Clustering

As previously discussed, overplotting becomes a significant problem in visualizations as dataset size increases, due to the limited visual space available for display. To mitigate this, the generation of proportional symbol maps can be automated using clustering algorithms that group observations and prevent overlap. The clustering approaches proposed in this thesis are guided by the principle that "perceptual and interactive scalability should be constrained by the chosen resolution of the visualized data, rather than by the number of records" [27], and they adhere to most of the guidelines established by Elmqvist et al. [15]:

Entity Budget Maintain a limit on the number of entities.

Visual Summary Aggregates should effectively convey underlying data information.

Visual Simplicity Aggregates should remain clear and straightforward.

Discriminability Aggregates should be distinguishable from individual data items.

Fidelity Be cautious that abstractions do not mislead.

Interpretability Aggregation should only go as far as ensuring the result remains correctly interpretable within the visual mapping.

The guideline on discriminability was not followed, as clusters were treated as a means to represent observation density rather than individual observations.

Two different algorithms are discussed and compared.

Hierarchical Greedy Clustering Hierarchical greedy clustering is an algorithm designed to group points into clusters at a specified zoom level. The process involves the creation of a k-d tree. A k-d tree (k-dimensional tree) 13 is a binary search tree used for organizing points in a k-dimensional space. It recursively partitions the space into two halves at each node, alternating between dimensions, making it efficient for range searches and nearest neighbor queries in multidimensional data.

The hierarchical greedy clustering algorithm then proceeds with the following steps:

1. **Initialize with a Point**: Start with any point from the dataset.

2. **Search for Nearby Points**: Identify all points within a specified radius using the k-d tree.

3. **Form a Cluster**: Create a new cluster with these nearby points.

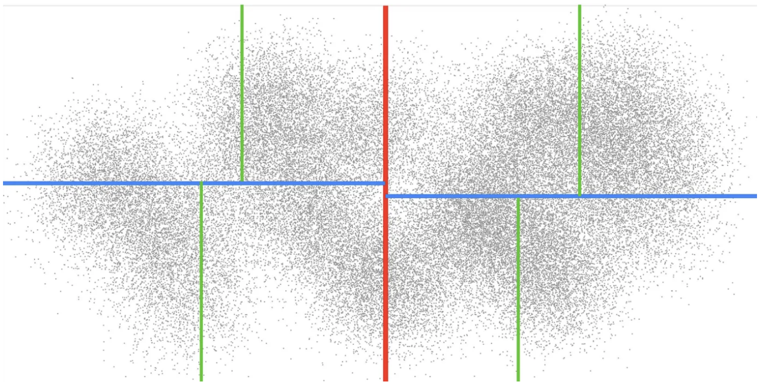


Figure 13: First three levels of a k-d tree.

4. **Repeat for Remaining Points**: Continue selecting new, unclustered points, repeating the process until all points are clustered.

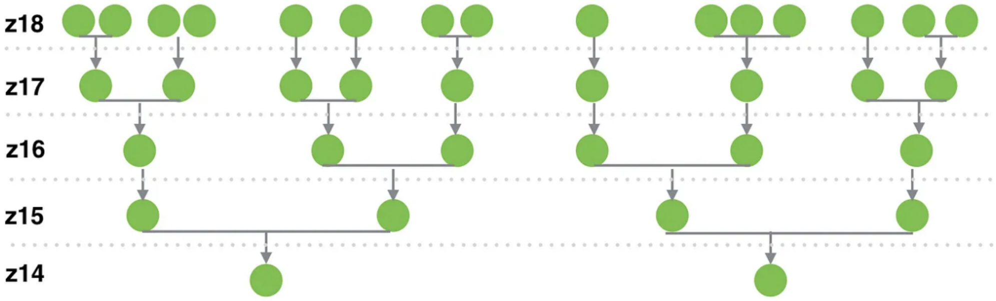


Figure 14: Structure of clusters produced by the hierarchical greedy algorithm, organized in a tree as a function of the zoom level.

At each zoom level, from the highest to the lowest, this process is repeated iteratively, starting with the weighted centroids from the previous level (bottom-up), forming a hierarchical tree structure. The number of elements decreases exponentially with each level.

For dynamic datasets, clusters must be recalculated after each dataset filtering, which is computationally expensive. The k-d tree must be rebuilt with a time complexity of $\mathcal{O}(n \log n)$, and clusters are then found greedily based on radius lookups at the lowest zoom level, also in $\mathcal{O}(n \log n)$. This process is repeated for all zoom levels, leading to an overall time complexity of $\mathcal{O}(n \log^2 n)$. Once the tree is built, fetching the clusters for a given zoom level is almost instantaneous, in $\mathcal{O}(1)$.

The library implementing this algorithm used in the application for clustering the dataset is called `Supercluster` [3], developed by Volodymyr Agafonkin, the creator of `Leaflet`. The lookup radius can be adjusted through parametrization of the clustering function to offer clusters of variable resolution.

Hexagonal Grid Clustering Grid clustering proceeds by binning the data into regular shapes in the dimensions of the data to calculate aggregated values in each bin. The aggregated values in the data space can then be represented in visual space. Using a hierarchical grid, different levels of aggregation can be obtained depending on the zoom level.

Other grid systems are available such as the square grid system `S2` from Google [17] but hexagonal grids, such as H3 offer a significant advantage due to the uniform distance between a hexagon and its six neighbors, leading to consistent and balanced clustering across the grid. The H3 system can create aesthetically arranged clusters. With small clusters, this approach enables proportional symbol maps that closely resemble exhaustive maps, as proposed by Bertin (see Figure 15) in [5].

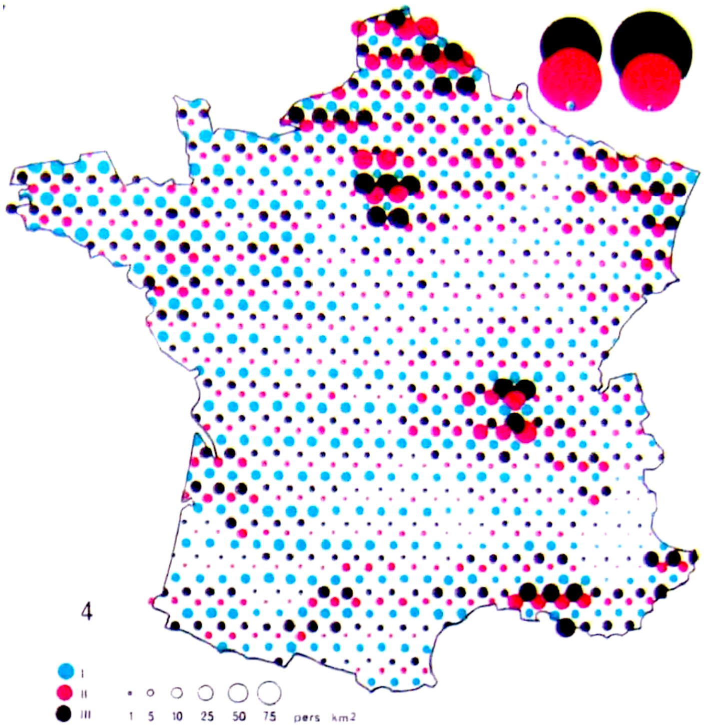


Figure 15: Bertin exhaustive map. The map shows the number of French workers per geographic department in 1954, per economic sectors: the primary (agriculture), the secondary (industry), and the tertiary (commerce, transports, services).

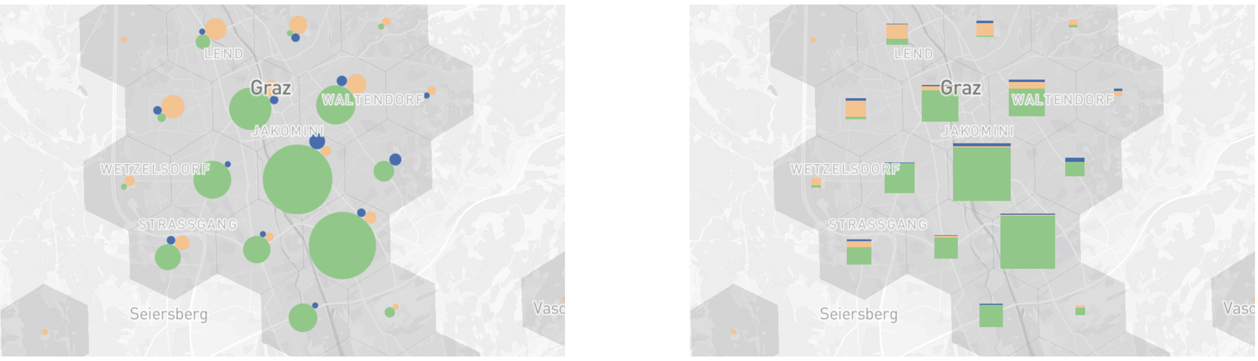
An important constraint when using H3 for clustering is that its 16 hierarchical levels with aperture 7 do not perfectly align with the zoom levels used in popular Web Map Services, where tiles are divided by four at each level. As a result, a mapping must be made using logarithmic transforms to reconcile these differences.

The algorithmic complexity of clustering a dataset with H3 is $\mathcal{O}(n \log n)$ for an unsorted dataset, equivalent to the complexity of sorting. This is because finding the parent of a given cell relies on bitwise manipulation (masking of the least significant bits) of the index.

Comparison of the Two Clustering Methods This thesis proposes to compare hexagonal grid clustering, using the H3 framework, with hierarchical greedy clustering. The comparison will evaluate the efficiency of grid-based clustering implemented with H3 as compared with Supercluster, focusing on computational performance, clustering accuracy, and the visual arrangement of clusters.

This research complements the research by Beresnev et al. on hexagonal clustering [4] in which aggregation is done server-side, by proposing an implementation running fully in the browser, which makes comparison with Supercluster in terms of performance possible.

Clustering of Multi-Categorical Data Clustering can be helpful to mitigate overplotting, but the proportion of different categories within a cluster can also reveal important patterns. There are several ways to represent these within-cluster categories, but two methods were chosen for their ease of implementation and scalability using `Deck.gl`. The first method breaks down the cluster into smaller circles for each category, with the radius of each circle scaled according to the category's proportion. The circles are then packed together using the front chain algorithm by Wang et al. described in [44]. Each circle is colored with a different hue to represent its category. The second method uses stacked colored rectangles of variable sizes to show both the proportion of each category and the total value for the cluster 16. Both methods were evaluated based on preferences from a panel of users.



(a) Packed circle representation

(b) Stacked bar representation

Figure 16: Two alternative ways used in the application to represent the within-cluster proportion. The figures show the proportion of observations of mosquitoes from different species.

6.4 Evaluation and Testing

Evaluating the quality and effectiveness of a visualization presents significant challenges, as emphasized by various authors [35, 25, 43]. The literature on this subject typically focuses on two primary aspects [9]:

- **Quantitative Metrics:**

- **Quantitative Metrics:** This includes metrics like the time taken to perform tasks with the visualization, the computational complexity of the interaction algorithms, and how smoothly users can interact with the system.

- **Qualitative Aspects:** These aspects involve assessing the usability and insight-generation capabilities of the visualization, often through observation techniques and interviews.

Carpendale [9], drawing on similar challenges highlighted by McGrath in social science research [30], notes that a key difficulty in evaluating data visualizations is achieving results that balance generalizability (the ability to extend results beyond the experimental context), precision (confidence in the results and control over extraneous variables), and realism (the extent to which the evaluation setting mirrors the real-world context of use).

While data visualization creators often seek to evaluate how effectively their designs generate insights, this proves to be an even more complex challenge than data visualization evaluation in a broader sense [35, 9, 41]. Plaisant [35] describes data visualizations as tools that help in "answering questions you didn't know you had," which ties directly to the concept of insight—a quality that is notoriously difficult to measure. This challenge arises because insights are subjective and may only surface long after the initial interaction with the visualization.

To address these evaluation challenges and integrate both quantitative and qualitative approaches, Stasko proposed a value-driven evaluation framework [41]. He argues that:

> "Visualization should ideally provide broader, more holistic benefits to a person about a data set, giving a 'bigger picture' understanding of the data and spurring insights beyond specific data case values."

Stasko formalized this approach through the introduction of the value equation:

$$V = T + I + E + C$$

where:

- **T** - The ability of a visualization to minimize the time required to answer a broad range of questions about the data.
- **I** - The capability of a visualization to foster insights and provoke insightful questions about the data.
- **E** - The effectiveness of a visualization in conveying an overall essence or key takeaway from the data.
- **C** - The ability of a visualization to build confidence, knowledge, and trust regarding the data, its domain, and context.

While this value-driven approach is compelling, it initially lacked detailed guidelines for quantifying the resulting value. To enhance this framework, Wall et al. [43] introduced a methodology involving questionnaires designed to evaluate each component of the value equation systematically. Although the evaluation of the application discussed here drew inspiration from this enhanced approach, it was not implemented in full. Instead, the key concepts served as guiding principles for the evaluation process.

In a comprehensive survey of visualization literature, Lam et al. identified seven key scenarios for visualization evaluation [25]. For the evaluation of the application, which targets a general public audience and makes use of a new algorithm for data aggregation, three scenarios were particularly relevant:

- Evaluating Communication Through Visualization (CTV)
- Evaluating User Experience (UE)
- Evaluating Visualization Algorithms (VA)

The paper by Lam et al. outlines typical tools suited for each scenario. Based on the recommended methods for these scenarios, and to ensure complementary approaches, three evaluation techniques were selected:

- Questionnaires with a series of open-ended, multiple choice and Likert-scale questions.
- Observations of users interacting with the application while being requested to accomplish a series of tasks, followed by interviews.
- Comparison of algorithms used for data clustering.

7 Results

7.1 Application

The application, in the version tested during questionnaires and interviews at the following URL `https://d27werzm43zim8.cloudfront.net/`.

7.2 Questionnaire Answers

Thirteen results were collected from users having watched an explanatory video about the application. The sample is small but already shows some trends regarding preferences in the options provided within the application. A summary of the answers for each question is presented here, and the full answers can be consulted as supplementary digital material in the annexes.

Do you have experience in epidemiology, public health, or medical science?

Five of the thirteen respondents reported having professional or academic experience in epidemiology, public health, or medical science. This experience may introduce bias, limiting the generalizability of the results to the broader population. However, it also offers valuable insights into aspects that may not have been considered due to the researcher's lack of formal training and professional experience in these fields.

Do you have experience with data visualization, as a consumer or creator?

Eight of the thirteen respondents indicated experience with data visualization, either through professional exposure or as creators. This high proportion results from selection bias, as the questionnaire was distributed mainly among study stakeholders, colleagues of the researcher, and via the professional network LinkedIn, where many of the researcher's connections work in data science. While this bias may affect the generalizability of the results, it also provides an advantage, as respondents' familiarity with data visualization could lead to more informed and relevant feedback.

What do you like in the application?

The analysis of the answers showed that users particularly enjoyed the speed at which the dataset was loaded and browsed, with six out of the thirteen mentioning adjectives such as "fast," "quick," or "fluid." The interactivity offered by cross-filtering between the map and the timeline was also appreciated, with six out of the thirteen mentioning this aspect. Some answers or parts of answers were less specific, using adjectives such as "nice" or "cool." One respondent expressed appreciation for the autoplay feature, which allows the COVID-19 dataset to be played automatically in a movie-like fashion.

What do you dislike in the application?

This question elicited a variety of responses, making it difficult to prioritize the feedback by importance. Two respondents found the amount of information and options to be overwhelming at first, citing a lack of clarity in functionality. Additionally, two respondents had issues with the contrast of colors. Another respondent criticized the lack of contextualization and storytelling. Overall, this section was very instructive and provided valuable insights for iterating and refining the application.

What additional features or improvements would you suggest for the application?

Similarly to the previous question, this one also elicited a variety of responses. Two respondents expressed a desire to import their own datasets, highlighting the potential of the application as a broader online data exploration tool, possibly extending beyond epidemiological data.

A respondent working in the field of epidemiology pointed out that in the COVID-19 section, having the population density circles colored homogeneously for an entire department could be misleading: while the indicators are collected at the department level, there can be disparities within the department. This respondent also identified an issue with circles overlapping the boundaries between two or more departments, suggesting that an average should be calculated in such cases or proposing the use of hollow circles with only a stroke to still highlight the population while maintaining the choropleth as a background.

Two other themes raised were the need to improve the clarity of labels and incorporate more storytelling elements to contextualize the visualization. Additionally, other user interface (UI) issues were mentioned, such as the suggestion to move the parameters drawer from the right to the left side of the screen to adhere to UI best practices and conventions.

Is the application insightful?

This was a Likert-scaled question with responses on a range of 7 steps from 1 to 7, from "Not at all" to "Very much". This resulted in an average rating of 6.31 (std. err. 0.75) which is encouraging, but might highly suffer from *courtesy bias*, which is a type of bias that occurs when the respondent tends to not fully state their opinion for fear of being perceived as overly critical.

How would you rate the application in terms of aesthetics?

This question also provided good ratings, 6.31 out of 7 (std. err. 0.75) with an average rating with the same caveats as with the previous question. Aesthetics was considered in the questionnaire as aesthetics can improve the attractiveness and popularity of an application [10], which is an important aspect to consider to increase reach and engagement.

Did you encounter performance issues? (at loading the application and when browsing the datasets)

Most of the users didn't encounter performance issues worth mentioning. Four respondents mentioned occasional short delays of one second or less when adjusting zoom or changing parameters.

What type of hardware did you use, in particular, which screen size and CPU, if you know?

All but one respondent was using high-end laptops, with a majority of recent MacBooks, highlighting the need to test the application with less powerful machines. The user using a lower-end machine considered the performance as acceptable.

In the MosquitoAlert section, is the filtering of data based on the currently visible part of the timeline and map intuitive?

The filtering based on both the visible part of the timeline and the visible part of the map was considered intuitive for most users, with an average score on the Likert scale of 5.53 out of 7 (std. err. 1.33). This highlights the potential of complementing a map with a timeline for filtering spatio-temporal datasets.

Depending on zoom levels, the scale of the symbols (circles or squares) is rescaled. Do you find it problematic or counterintuitive?

Some users found it unintuitive that the circle scale adjusted depending on the context. With a Likert scale where 1 meant "Not problematic at all" and 7 meant "Highly problematic", the question received an average score of 3.07 (std. err. 2.1), this indicates that this aspect should receive further attention. The initial intention behind this rescaling was to make observations with low values more visible when the viewport contains only low values, emphasizing proportions rather than absolute numbers. This feedback highlights a broader challenge in data visualization: balancing the need to represent phenomena that span multiple orders of magnitude while also drawing attention to weaker signals.

In the MosquitoAlert section, do you think that the clustering adds value compared to the no-aggregation option (all observations visible individually)?

Most users find it useful to have aggregation to group observations and prevent clutter, with an average score of 5.53 over 7 (std. err. 1.45) on the Likert scale.

Do you see a difference, and do you have a preference between the "H3 grid" and the "Supercluster" ways of aggregating the observations?

Overall, nine out of thirteen respondents expressed a preference for the H3 grid. One respondent specifically appreciated the symmetrical arrangement provided by H3. However, caution is advised as there may be bias in these responses: The breaks at which H3 clustering occurs are determined by the aperture of the hierarchy, which doesn't always align with the cluster radii obtained using Supercluster. This often results in H3 clusters appearing larger in the application. This was noted by two respondents: one who preferred the H3 grid for this reason and another who favored Supercluster. This aspect could have been better controlled in this small-scale evaluation and may require further testing on a larger scale.

Do you prefer the "Packed Circles" or the "Stacked Bar" option to visualize clusters?

Most users (84.6%) preferred the "Packed Circles" option in this two-choice question. Although the exact reasons for this preference aren't fully clear, the result, when considered alongside the preference for H3 grid versus Supercluster visualizations, suggests that Bertin's exhaustive maps might be a strong method for showing multi-categorical data in interactive applications.

In the COVID-19 section, do you prefer the choropleth (uniform colors) view, the visualization highlighting population density with circles, or do you think they complement each other?

Most users (84.6%) thought that the choropleth map and population density circles either worked well together (61.5%) or preferred the population density circles on their own (23.1%). Only 15.4% preferred the choropleth map by itself. This suggests that combining choropleth maps with proportional symbol maps could be effective for better data interpretation. However, it's important to ensure that the values displayed are clear and easily distinguishable.

In the COVID-19 section, do you think highlighting the population brings additional insight or value?

All respondents agreed that highlighting the population adds insight or value, with different levels of certainty on a Likert scale, averaging 6.0 out of 7 (std. err. 1.08). This agreement, combined with the responses to the previous question, shows that adding population density circles can provide a valuable extra layer of information to choropleth maps when additional data is available, making the data more informative.

Do you think that the animation of indicators over time adds value and insights in the COVID-19 section?

All but one respondent agreed that animating indicators over time in the COVID-19 section adds value and insight, with an average of 5.92 out of 7 (std. err. 1.84) on the Likert scale. This strong agreement highlights the importance of temporal animations in making COVID-19 data more understandable and engaging, allowing users to see trends and changes over time in a clearer way.

7.3 User Observations and Interviews

7.3.1 Interview A

The first interviewee is an adult female with a background in biochemical science. The interview was conducted in French, a language in which the respondent has near-native fluency. She had watched the same explanatory video prepared for the respondents of the questionnaire but did not complete the questionnaire herself. The interview and observation were recorded. During the interview, she was asked similar questions to those asked in the questionnaire, regarding her preferences between options available in the application. Additionally, she was asked to perform simple tasks, such as zooming in on a specific date and observing the temporal patterns of mosquito reporting on a 24-hour scale, or selecting a date in the COVID-19 section for a given indicator. The interviewee was asked to describe what she observed in different parts of the app and how she interpreted the information.

MosquitoAlert The interviewee completed all the requested tasks with ease. She noticed a shift in the center of the proportional symbols representing counts at very low zoom levels when using the `H3` clustering, with some circles appearing misplaced, over the sea instead of land. This artifact and its cause had to be explained, but as she zoomed in, she realized that greater precision could be achieved through interaction, and the artifact at lower zoom levels no longer seemed problematic. Interestingly, she preferred the absence of clustering, which contrasts with the preferences of most questionnaire respondents.

COVID-19 France The interviewee was asked to explain her understanding of the visualization after a brief introduction. The role of the visual variables wasn't immediately clear to her; she initially interpreted the radii as representing the number of cases when they actually represented population density. This underscores the importance of clear legends and contextual explanations to ensure a proper understanding of the visualization. The interviewee particularly appreciated the animation of the dataset over time in a movie-like fashion, which sparked insights and hypotheses about the geographical spread of the disease.

7.3.2 Interview B

The second interviewee holds a doctorate in particle physics and has successfully transitioned into a career in data science, where she has accumulated several years of experience. Both the observation and the interview were conducted online, with the session being recorded for future reference. The interviewee's computer was relatively old, manufactured in 2017. The interview was conducted in French, a language in which the interviewee demonstrates near-native proficiency.

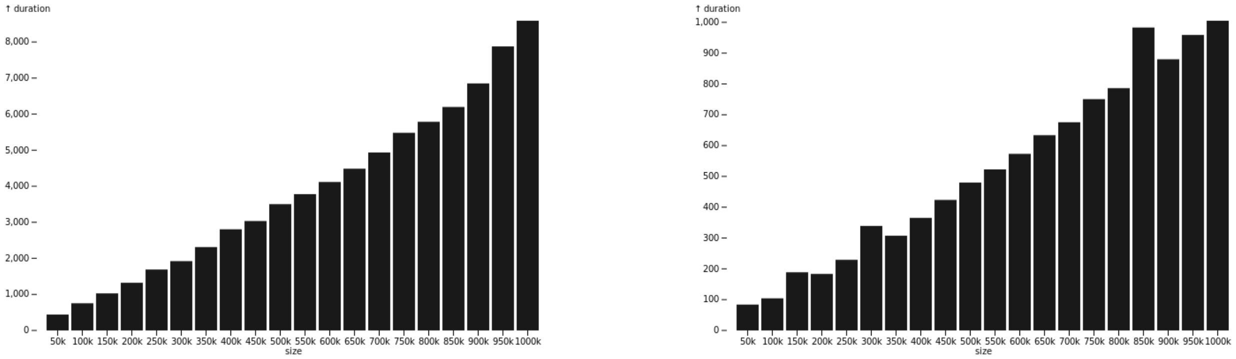
MosquitoAlert The interviewee was asked to navigate the dataset across various time periods and geographical areas to observe patterns. She was able to identify observational patterns with ease and did not require additional explanations, indicating that the task could be performed intuitively. The interviewee confirmed the user-friendliness of the exploration method, which involved navigating through space and time using the timeline and the map. She expressed a preference for the `H3` grid clustering over both `Supercluster` and the absence of clustering. Although some lag was noted during dataset updates in response to user interactions, likely due to the lower performance of the computer used, the interviewee considered the overall performance acceptable.

COVID-19 France The interviewee was introduced to the two different thematic maps: choropleth and density circles. She expressed a preference for the density circles, while also acknowledging the utility and complementarity of both visualizations. The animation of the dataset through time and space, presented in a movie-like fashion, was found to be interesting. However, the interviewee raised questions regarding the intended audience, noting that the value of such visualization would depend on the target users. After clarifying the target audience, the interviewee mentioned that she would use such a tool as a citizen if it were made publicly available. When asked about potential improvements, the interviewee could not identify any changes or enhancements she would suggest at the moment.

7.4 Computational Comparison of Clustering Algorithms

While the application allowed the selection of both the `Supercluster` and `H3` algorithms and the ability to test them inside the application from a user perspective in terms of the visual arrangement of clusters, the comparison from a computational point of view would have been difficult within the application. Both implementations provided fast results due to the small size of the MosquitoAlert dataset, with only about 40,000 data points. The `Supercluster` implementation also suffers from a handicap as it is written in pure `JavaScript`, whereas the `H3` implementation used in the application is written in `C++`, a potentially much faster language, compiled to `Wasm` as a target to run in the browser.

Therefore, to provide a fair comparison, an Observable notebook [12] was created where one million random points were generated with latitude ranging from 85 to -85 and longitude ranging from 180 to -180 degrees. The points were indexed with `H3` at the lowest resolution (about one-meter precision) and clustered with both methods at various resolutions. This comparison specifically addresses the context of dynamic datasets, i.e., datasets that can be modified based on user interactions. The speed did not seem to depend on the chosen resolution in both `Supercluster` and with `H3` as this step is $\mathcal{O}(1)$ in `Supercluster` once the k-d tree has been built with `Supercluster`, and the `H3` algorithm remains $\mathcal{O}(n\ log\ n)$ for all resolutions. The random dataset was sliced and increased in size by increments of 50,000 points to observe the speed of clustering for variable dataset sizes. The results favor the H3 algorithm by a factor of 10 17.



(a) Clustering performance in ms as a function dataset size with Supercluster

(b) Clustering performance in ms as a function dataset size with H3

Figure 17: Comparison of performance Supercluster vs. H3

It should be noted that clustering with `H3` comes with some caveats: the clusters will be centered on the centroid of the parent element, not the center of mass of the children. Non-containment can also occur, with some elements displayed in a neighboring cluster, and finally, the radii are constrained by the grid resolution. Despite these caveats, `H3` can be considered when a faster alternative is required or when hexagonal grid arrangement is desired. Future research could also lead to solutions where the two methods complement each other, as discussed in 10. For speed-critical, computationally intensive algorithms, another notebook prepared for this research shows that the use of `DuckDB` with the `H3` extension can improve performance further by a factor of about two [13].

8 Discussion and Conclusion

The findings from the questionnaires and observations offered valuable insights into user interaction with and perception of the application. Despite the small sample size, it was sufficient to identify distinct preferences and areas needing improvement. With further refinements, the application could be prepared for release to a wider audience. The MosquitoAlert team proposed incorporating the interactive visualization into the "labs" section of their website [34], which would enhance the visibility of this thesis's results and potentially stimulate further development of the application.

A key consideration in data visualization is the quality of the underlying data, as the effectiveness of any visualization is limited by the accuracy and reliability of the data it represents. This highlights the importance of providing context for visualizations, including warnings about potential issues and biases that may stem not only from the visualization itself but also from the data collection processes and the data used.

Web-based interactive data visualization can serve as a powerful communication tool for reaching a wide audience, especially in public health and epidemiology. However, it should be integrated into a broader, holistic strategy and used alongside other communication methods.

9 Summary of Findings

The application demonstrates several promising techniques, but also areas for improvement:

- The integration of a zoomable timeline with a zoomable map for exploring spatiotemporal data is an innovative and underutilized approach that received positive feedback from users.

- Data clustering using `H3` proves to be an efficient and quick method to reduce visual clutter. It can be effectively combined with circle packing or other techniques to display multicategorical data.

- Choropleth maps have limitations and may introduce biased perceptions of distributional patterns for epidemiological indicators such as incidence and prevalence rates. Complementing or replacing choropleth maps with proportional symbols indicating population density can add value. This can be achieved by incorporating additional datasets.

- Animation of spatiotemporal datasets over time, while displaying geographical distribution, was well-received by users.

- `H3` is a promising technique for indexing geographical areas, enabling the integration of multiple datasets through a shared indexing system.

- A technical ecosystem composed of `DuckDB`, `H3` and files in `Parquet` format, combined with visualization libraries such as `D3` and `Deck.gl` offers a powerful solution for interactive visualization entirely in the browser.

- Data scaling over multiple orders of magnitude poses significant challenges to be visualized in a way that conveys the full picture without dwarfing weak signals.

- Data visualization should be integrated into a holistic approach, working alongside other techniques to convey the essence of the data, especially when intended to be used by a large audience.

10 Implications for Future Research

As was shown, hexagonal grid aggregation with `H3` offers a fast method for creating visually appealing clusters on the map in $\mathcal{O}(n\ log\ n)$ in general and $\mathcal{O}(n$ in a properly sorted dataset. The clusters are positioned in the center of the parent cell, which can be a desirable arrangement for regularly spaced symbols but could also introduce some bias if the centroid of data points in a cluster and the centroid of the parent cell are far away. A proposal for future research would be to use hexagonal grid aggregation with H3 as a preprocessing step, one or two resolutions finer than the desired cluster radius, before using a more costly algorithm.

A sketch of the algorithm would be as follows:

1. Pre-aggregate with H3 using:

   ```
   h3.getHexagonEdgeLengthAvgM(resolution) < targetRadius
   ```

   (with `targetRadius` in meters).

2. For each cluster of this pre-aggregation, calculate the centroid based on ij distance[1].

3. Use a more costly algorithm, such as the greedy clustering algorithm, to merge overlapping clusters.

Distance lookups for the last step to merge overlapping clusters could also take advantage of k-ring calculation (`h3.gridDisk()`), which offers a good approximation of Haversine distance.

Another area for further research is finding better ways to represent epidemiological data spanning over multiple orders of magnitude, in particular in the context of interactive visualizations. Some techniques were investigated to use an ordinal scale for colors by clustering the scale in 1-D based on clustering of the data in a given dimension with techniques such as Jenks breaks [22] or quantiles but a challenge is to provide smooth transitions when the dataset can also be explored along time, while it was observed that those techniques introduce jumps in the produced ordinal scale when the data is refiltered, and this could confuse the user.

References

[1] Facebook's high resolution population density maps demographic estimates. https://dataforgood.facebook.com/dfg/tools/high-resolution-population-density-maps, 2024. Accessed: 2024-08-22.

[2] Global human settlement layer. https://human-settlement.emergency.copernicus.eu/about.php, 2024. Accessed: 2024-08-22.

[3] Vladimir Agafonkin. Clustering millions of points on a map with supercluster. https://blog.mapbox.com/clustering-millions-of-points-on-a-map-with-supercluster-272046ec5c97, 2016. Accessed: 2024-08-14.

[4] A Beresnev, A Semenov, and E Panidi. Hexagonal grids applied to clustering locations in web maps. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:435–440, 2022.

[5] Jacques Bertin. *Semiology of graphics.* University of Wisconsin press, 1983.

[6] Jacques Bertin. *Graphics and graphic information processing.* Walter de Gruyter, 2011.

[1] The ij distance refers to the distance calculated between cells in the `H3` grid system [16]. An Observable notebook prepared for this research illustrates this approach [11].

[7] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[8] Isaac Brodsky. h3-duckdb: Duckdb extension for h3 geospatial indexing. `https://github.com/isaacbrodsky/h3-duckdb`, 2023. Accessed: 2024-08-12.

[9] Sheelagh Carpendale. Evaluating information visualizations. In *Information visualization: Human-centered issues and perspectives*, pages 19–45. Springer, 2008.

[10] Nick Cawthon and Andrew Vande Moere. The effect of aesthetic on the usability of data visualization. In *2007 11th International Conference Information Visualization (IV '07)*, pages 637–648, 2007.

[11] Julien Colot. H3 centroid with ij coordinates. `https://observablehq.com/@jcolot/h3-centroid-with-ij-coordinates`, 2023. Accessed: 2024-08-14.

[12] Julien Colot. Clustering performances: H3 vs supercluster. `https://observablehq.com/@jcolot/clustering-performances-h3-vs-supercluster`, 2024. Accessed: 2024-08-14.

[13] Julien Colot. Duckdb + h3: Clustering performance. `https://observablehq.com/@jcolot/duckdb-h3-clustering-performance`, 2024. Accessed: August 14, 2024.

[14] Julien Colot. Duckdb-wasm with unsigned loadable extensions. `https://observablehq.com/@jcolot/duckdb-in-webassembly-with-loadable-extensions`, 2024. Accessed: 2024-08-14.

[15] Niklas Elmqvist and Jean-Daniel Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE transactions on visualization and computer graphics*, 16(3):439–454, 2009.

[16] H3 Geo. Coordinate systems - h3 documentation, 2024. Accessed: 2024-08-13.

[17] Google. S2 geometry library. `http://s2geometry.io/`. Accessed: 2024-08-14.

[18] Amy L Griffin. Trustworthy maps. *Journal of Spatial Information Science*, 2020(20):5–19, 2020.

[19] Mark Harrower and Cynthia A Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[20] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue*, 10(2):30–55, 2012.

[21] David François Huynh, Stefano Mazzocchi, and Lee Feigenbaum. Simile timeline. `http://simile.mit.edu/timeline/`, 2005. Accessed: 2024-08-04.

[22] George F Jenks. The data model concept in statistical mapping. *International yearbook of cartography*, 7:186–190, 1967.

[23] Živko Južnič-Zonta, Isis Sanpera-Calbet, Roger Eritja, John RB Palmer, Agustí Escobar, Joan Garriga, Aitana Oltra, Alex Richter-Boix, Francis Schaffner, Alessandra Della Torre, et al. Mosquito alert: leveraging citizen science to create a gbif mosquito occurrence dataset. *Gigabyte*, 2022, 2022.

[24] Alexander J Kent. Mapping and counter-mapping covid-19: From crisis to cartocracy, 2020.

[25] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2011.

[26] H3 Core Library. Tables of cell statistics across resolutions, 2024. Accessed: 2024-08-14.

[27] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. immens: Real-time visual querying of big data. In *Computer graphics forum*, volume 32, pages 421–430. Wiley Online Library, 2013.

[28] Alan M MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 2004.

[29] Alan M MacEachren, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE transactions on visualization and computer graphics*, 18(12):2496–2505, 2012.

[30] Joseph E McGrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in human–computer interaction*, pages 152–169. Elsevier, 1995.

[31] Franz-Benjamin Mocnik, Paulo Raposo, Wim Feringa, Menno-Jan Kraak, and Barend Köbben. Epidemics and pandemics in maps–the case of covid-19. *Journal of Maps*, 16(1):144–152, 2020.

[32] Joel L Morrison. A theoretical framework for cartographic generalization with the emphasis on the process of symbolization. *International Yearbook of Cartography*, 14(1974):115–27, 1974.

[33] Mosquito Alert. Explora. `https://www.mosquitoalert.com/en/project/explora/`, n.d. Accessed: 2024-08-13.

[34] MosquitoAlert. Mosquitoalert labs, 2024. Accessed: 2024-08-16.

[35] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pages 109–116, 2004.

[36] Nick Rabinowitz. Timemap.js. `https://github.com/nrabinowitz/timemap`, 2008. MIT License.

[37] Robert E Roth. Visual variables. *International encyclopedia of geography: People, the earth, environment and technology*, pages 1–11, 2017.

[38] Kevin Sahr, Denis White, and A Jon Kimerling. Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003.

[39] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier, 2003.

[40] Artemis Skarlatidou, Muki Haklay, and Tao Cheng. Trust in web gis: the role of the trustee attributes in the design of trustworthy web gis applications. *International Journal of Geographical Information Science*, 25(12):1913–1930, 2011.

[41] John Stasko. Value-driven evaluation of visualizations. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 46–53, 2014.

[42] Inc. Uber Technologies. geojson2h3, 2023.

[43] Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):491–500, 2018.

[44] Weixin Wang, Hui Wang, Guozhong Dai, and Hongan Wang. Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 517–520, 2006.

[45] Matthew O Ward, Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications.* AK Peters/CRC Press, 2010.

[46] xyzt.ai. xyzt.ai: Spatio-temporal visual analytics platform, 2024. Accessed: 2024-08-11.

[47] Manlai You, Chun-wen Chen, Hantsai Liu, and Hsuan Lin. A usability evaluation of web map zoom and pan functions. *International Journal of Design*, 1(1), 2007.

[48] John Zarocostas. How to fight an infodemic. *The lancet*, 395(10225):676, 2020.

A Appendices

A.1 Digital Supplementary Material

The complete set of questionnaire answers is provided as supplementary material and can be accessed online at `https://github.com/jcolot/master-thesis-materials/tree/main`. The names and contact details of the respondents have been censored.

The code of the application is available at `https://github.com/jcolot/master-thesis-public/tree/main`