

Visual Analytics of Spatial Time Series Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Marija Priselac, BSc ing.comp.

Matrikelnummer 11735131

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: Privatdoz. Dipl.-Ing. Dr.techn. Krešimir Matković

Wien, 12. Oktober 2021

Marija Priselac

Krešimir Matković

Visual Analytics of Spatial Time Series Data

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Marija Priselac, BSc ing.comp.

Registration Number 11735131

to the Faculty of Informatics

at the TU Wien

Advisor: Privatdoz. Dipl.-Ing. Dr.techn. Krešimir Matković

Vienna, 12th October, 2021

Marija Priselac

Krešimir Matković

Erklärung zur Verfassung der Arbeit

Marija Priselac, BSc ing.comp.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. Oktober 2021

Marija Priselac

Kurzfassung

Unabhängig davon, welche Algorithmen und Technologien entwickelt werden, bleiben der menschliche Verstand und das logische Denken wichtige Werkzeuge für die Analyse, Modellierung und Lösung von Problemen. Die visuelle Darstellung von Daten gilt als der effektivste Weg, um Informationen an das menschliche Gehirn zu übermitteln und analytisches Denken zu fördern. Visual Analytics umfasst eine Reihe von Techniken, Methoden und Werkzeugen, die analytisches Denken durch visuelle Darstellungen verschiedener Datentypen unterstützen. Aufgrund ihrer Komplexität und Größe eignen sich räumliche Zeitreihendaten für die Implementierung solcher Techniken, da ihre Analyse nach wie vor eine Herausforderung darstellt. Viele ökologische, soziale und wirtschaftliche Prozesse der modernen Zivilisation werden durch räumliche Zeitreihen dargestellt, was die Notwendigkeit interaktiver visueller Darstellungen für ihre effizientere Analyse unterstreicht.

Ein klares Beispiel für solch komplexe Prozesse ist die wirtschaftliche Rezession, ein Rückgang der Wirtschaftstätigkeit, für den es keine einheitliche formale Definition gibt. Sie wird jedoch häufig mit Rezessionsfaktoren wie BIP, Gini-Index oder Inflation beschrieben, die allesamt Beispiele für räumliche Zeitreihendaten sind und deren Veränderung ein klarer Indikator für die Wirtschaftslage sein kann. Da die Rezessionsanalyse ein sehr komplexes Thema ist und nicht ganz klar ist, welche wirtschaftlichen Faktoren den größten Einfluss haben, sind rein automatisierte Verfahren nicht geeignet und es gibt Spielraum für Fortschritte bei den analytischen Ansätzen.

Diese Arbeit schlägt eine Anwendung “Recession Explorer”: Visual Analytics der wirtschaftlichen Rezession und ihre Vorhersage als Beispiel für ein ganzheitliches System vor, das räumliche Zeitreihendaten anzeigt und Muster und Erkenntnisse in den Daten untersucht. Eine solche Kombination von Ansätzen bietet eine einzigartige Perspektive auf wirtschaftliche Rezessionsstudien, indem sie sowohl das menschliche Denken auf hoher Ebene als auch die Verwendung fortschrittlicher mathematischer Algorithmen erleichtert. Das Ziel der Anwendung besteht darin, zu zeigen, dass der Einsatz von Visual Analytics ein nützlicher Ansatz ist, um die Herausforderungen einer wirtschaftlichen Rezession zu bewältigen und Benutzern im Allgemeinen mit interaktiven Visualisierungen bei der Analyse und dem Umgang mit räumlichen Zeitreihendaten zu helfen.

Abstract

Regardless of what algorithms and technologies are developed, the human mind and logical reasoning remain important tools for analysing, modelling, and solving problems. Visual representation of data is considered the most effective way to convey information to the human brain and promote analytical thinking. Visual analytics encompasses a set of techniques, methods, and tools that support analytical thinking through visual representations of various types of data. Due to their complexity and size, spatial time series data are suitable for implementation of such techniques, as their analysis remains challenging. Many environmental, social, and economic processes of modern civilization are represented by spatial time series, which emphasises the need for interactive visual representations for their more efficient analysis.

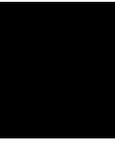
One clear example of such complex processes is economic recession, a decline in economic activity for which there is no single formal definition. However, it is often described in terms of recession factors such as GDP, the Gini index, or inflation, all of which are examples of spatial time series data, and whose change can be a clear indicator of the state of the economy. As recession analysis is a very complex topic and it is not entirely clear which economic factors have the greatest impact, purely automated techniques are not appropriate and there is scope for advances in analytical approaches.

This thesis proposes an application “Recession Explorer”: visual analytics of economic recession and its forecasting as an example of a holistic system that displays spatial time series data and explores patterns and insights in the data. Such a combination of approaches provides a unique perspective on economic recession studies by facilitating both high-level human reasoning and the use of advanced mathematical algorithms. The goal of the application is to demonstrate that the use of visual analytics is a beneficial approach to address the challenges of economic recession and, more generally, to assist users with interactive visualisations when dealing with and analysing spatial time series data.

Contents

Kurzfassung	vii
Abstract	ix
Contents	xi
1 Introduction	1
2 Recession	5
2.1 What is a recession?	5
2.2 Causes of a recession	6
2.3 How to predict a recession?	7
3 Analysis of Spatial Time Series	9
3.1 Characteristics and Challenges	9
3.2 Data Types	10
4 Visual Analytics and Spatial Time Series Data	19
4.1 Spatial Time Series in Visual Analytics	22
5 Prediction Methods	25
5.1 Definition of Linear Regression	26
6 Design of Visual Analytics Process	29
6.1 Implementation Technologies	31
6.2 Machine Learning Task	32
6.3 Visual Analysis Task	36
7 Evaluation	43
7.1 Use Case Scenarios of The Recession Explorer Application	43
8 Conclusion	49
8.1 Discussion and Limitations	50
8.2 Future Work	51
	xi

List of Figures	53
Bibliography	55



Introduction

Applications of spatial time series data are rapidly increasing and so is research in this area. The diversity of data types makes the analysis of spatial time series more complex and raises many challenges, which are addressed by newly introduced methods and techniques. Visual analytics represents one such method that aims to combine automated analysis techniques with interactive visualisations [1]. The progressive intertwining of information technology with various disciplines such as politics, economics, scientific research, and others has resulted in a huge amount of data. At the same time, the ability to collect and store data is growing faster than the ability to analyse it. Since raw data as such mostly has no value, data analysis techniques must be applied to extract the information contained in the data. These processes often lead to information overload, which can result in irrelevant, misrepresented, or incorrectly processed data. Therefore, appropriate methods are needed to exploit hidden opportunities and knowledge that lie in unexplored data resources [1]. Nowadays, it is widely accepted that complex real-world data cannot be adequately and efficiently analysed using only automated computational methods or interactive visualisations. Research in visual analytics aims to multiply the analytical power of humans and computers by finding effective ways to combine interactive visual techniques with algorithms for computational data analysis, and to develop new methods and procedures in which visualisation and computation interact and complement each other [2]. Combining interactive visual analysis with e.g., machine learning approaches such as linear regression to create simplistic models for predictions leads to unique prospects by supporting both the advancement of human reasoning and the use of advanced mathematical algorithms. The main principle of data visualisation is to allow an analyst to see patterns in the data. Similarly, researchers in the field of visual analytics strive to develop techniques that effectively visualise various patterns for the analyst's visual perception [2]. However, each analyst tends to interpret the patterns differently. In order to use these insights for further analysis, the patterns should be presented in an unambiguous form [2]. Various tools allow visual

representations to be supplemented by text, audio, or drawings, but they do not support additional computerised analysis of the patterns. Therefore, patterns must take the form of computer-processable models. Visual analytics tools and methods can support the process of creating such models.

The goal of this thesis is to implement a visual analytics process for spatial time series data, in particular for economic recession data. The interactive visual analysis is combined with a machine learning based model for forecasting of the economic recession for further extending the analytics process. The process was conducted on the dataset taken from Gapminder [3] which contains various recession factors such as GDP, inflation rate, employment rate, Gini index, investment, trade, etc. from 86 countries for the period from 1992 to 2016. As a result, an application Recession Explorer: visual analytics of economic recession and its forecasting was designed as an example of a holistic system that displays spatial time series data. The development of this application is divided into two tasks, which are explained in the following sections.

The first task is machine learning. It requires initial data preprocessing, followed by selection of the "best" model for the given spatial time series data. Preprocessing of spatial time series data is usually done to correct noise, errors and missing data, and to perform exploratory space-time analysis to understand the basic distributions of the collected data. Since data were not available for some countries in some years of the collected dataset, data cleaning was performed. Considering that the dataset contains a large volume of data, the records containing a missing values were removed. After data cleaning, the next step is data transformation. The method performed was normalization by arithmetic mean and standard deviation. Given the nature of the problem, an important feature of the model is the explanation of the predicted variables. Complex machine learning forecasting models, such as neural networks or the more recently developed class of recurrent neural networks, were not studied because of their lack of interpretability. Therefore, linear regression as an example of a white-box approach represents a suitable compromise to achieve interpretability. Since recession factors need to be predicted before recession can be predicted, this task was divided into two parts. First, predicting recession factors using linear regression with elastic net regularization, and second, predicting recessions using K-Means classification.

The second task is designing an application for visual analytics of economic recession and its forecasting. Visualisation of spatial time series data has long been researched and remains an important topic. Although visual analytics techniques have not yet been applied to this particular case, there are solutions that are somewhat related. Given the dataset, an interactive world map is a great way to visualise the results. Not only are maps great visual representations that engage users, but they also combine information from spatial data with data over time, leading to better understanding. Maps are also intuitive and allow for easy comparison of activity in different locations at a glance.

The design of the Recession Explorer application represents a possible answer to a research question: "What is an appropriate design of an effective exploratory analysis solution for spatial time series data?" Since this research question is very broad, there

is probably no single answer. Therefore, this thesis uses an example from the field of economic recession as one potential answer to the research question. The contributions of this thesis are summarized in the following list:

- Identification of the target problem and specification of the research question in the problem area of spatial time series analysis
- Exploration of relevant visual analytics methods and techniques in the field of spatial time series analysis
- Development of linear models for the prediction of economic recession factors and recession classification
- Implementation of the Recession Explorer application as an example of a holistic analytics system

The thesis is divided into eight main chapters, each of which has a different focus. The structure is as follows:

1. Chapter 1 discusses the motivation, problem statement, and research question.
2. Chapter 2 provides a definition of recession, discusses the causes of recession, and explains how business cycle analysts and data scientists predict recessions.
3. Chapter 3 introduces the analysis of spatial time series. It gives a general introduction to the topic, highlights the challenges, and describes different data types and the visualisation techniques for these data types.
4. Chapter 4 describes the field of visual analytics, followed by the definition of visual analytics, the process, the challenges, and the integration of spatial time series and visual analytics.
5. Chapter 5 provides the machine learning theory needed to predict recession factors and future recessions.
6. Chapter 6 presents the application Recession Explorer: visual analytics of economic recession and its forecasting. All the basic steps from implementation technologies to the implementation itself are elaborated and explained in detail.
7. Chapter 7 presents the evaluation of the application Recession Explorer described in the previous chapter.
8. The final chapter summarizes the main results, highlights unanswered questions and those beyond the scope of this thesis, points out known limitations, and finally provides recommendations for future work.

Recession

The world has experienced many periods of recession throughout history. In modern economies, recessions usually occur once a decade, followed by a period of strong economic growth [4]. This accurately explains the business cycle as an alternation of periods of expansion and contraction (recession) [5]. However, there is no single definition of a recession in the literature, and interpretations differ depending on the relevant economic concepts considered. The following sections provide some definitions of recession, discuss the causes of recession, and describe the techniques analysts use to predict recessions.

2.1 What is a recession?

Business cycle analysts use the term recession to refer to weak economic periods, whose duration, depth, and diffusion exceed the usual limits; thus, there are three key dimensions of a recession, known as the “three Ds”: Duration, Depth, and Diffusion [6]. This definition states that in a recession, economic activity declines, the decline has widespread effects on the economy, and lasts for a period of time. In addition, the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER) in the United States defines a recession as a significant decline in economic activity throughout the economy that lasts more than a few months and is usually visible in real GDP, real income, employment, industrial production, and sales [4]. These definitions show that the term is not very specific and provides a degree of freedom in determining whether or not an economy is in recession. Determining a recession based on earlier definitions requires extensive data analysis. In the 1970s, for example, Julius Shiskin’s simpler definition became popular: “recession is a decline in seasonally and calendar-adjusted real gross domestic product (GDP) for at least two consecutive quarters” [7]. Since GDP is indeed the most representative indicator of economic activity and the minimum period for defining a recession is at least six months, this definition takes into account two dimensions of the “three Ds”—duration and diffusion. The main drawback of the Shiskin

definition is that GDP fluctuates due to successive data revisions, which can easily turn a negative GDP into a positive one. If this definition were used to confirm a recession, many countries would have many more recessions than actually counted. Another definition of recession comes from the Austrian Business Cycle Theory [8]. This theory shows how central banking and monetary policy are related to economic events and the psychology and incentives that investors, producers, and consumers face in the economy. The main cause of recession is the production of new money in the form of loans and deposits by the banking system. Central banks do this because it is their basic business model. This leads to a decline in savings and the temporary illusion of a strong economy as prices and spending rise throughout the economy. However, once investors, producers, and consumers try to increase consumption while decreasing savings, the economy begins to fail. This leads to business failures, rising unemployment, debt deflation, and other signs of recession. As long as central banks produce new money and continue with this business model, recessionary periods are inevitable.

2.2 Causes of a recession

In business cycle theories [9] the factors that can cause a recession can be divided into two main groups, exogenous and endogenous, depending on how they explain business cycle fluctuations [10]. Endogenous factors arise within the economic system [11]. These factors can cause changes in the phases of the business cycle. The most important changes [12]:

1. Changes in Demand — Changes in demand lead to a change in economic activity. When demand increases, businesses produce more goods to meet the increased demand. Profit is higher and there is more income, employment and production. This leads to economic expansion. On the other hand, if demand falls, firms stop producing, there is less income, employment and output. This leads to an economic slump and eventually a contraction (recession).
2. Changes in Investment — Relatively small changes in investment can cause large changes in the economy triggering changes in output, income, and consumption [13]. A decline in investment reduces the production of capital goods, which means fewer payments to the firms that produce capital goods. As a result, these firms buy fewer consumption goods, which decreases the income of other producers. As the income of these producers falls, they cut payments to their resources, so they have less income, which reduces their consumption. As they consume less, production goes down, payments and income go down, and consumption is lower. The process continues. On the other hand, an increase in investment increases the production of capital goods, which leads to more consumption, an increase in production, payments, and income.
3. Changes in Macroeconomic Policy — Macroeconomic policy is a plan and action taken by the government to affect the economy as a whole [14]. The main macroeconomic policies of the government are fiscal policy, monetary policy and

supply-side policies. Fiscal policy refers to changes in government spending and taxation. Monetary policy consists of changes in the money supply, interest rates and the exchange rate. Supply-side policies aim to improve the competitiveness and efficiency of the free market. Monetary and fiscal policies ensure that the economy operates close to its potential. When economic growth is weak, the government lowers tax rates and interest rates and increases government spending and the money supply. In this way aggregate demand increases and stimulates economic growth. In other conditions, when inflation is high, the performance of the economy is above its potential level. The government increases tax rates and interest rates and decreases government spending and the money supply. These actions reduce aggregate demand, real GDP growth, and inflation.

Exogenous factors, on the other hand, are unforeseen situations in the outside world that affect the business cycle. Some of the most accurate are:

1. War — War is one of the most important exogenous factors [15]. It has a great influence on the economic situation. In war, large economic and monetary resources are spent on the production of weapons. Since war is the main concern, the need for consumer goods is neglected and the economy experiences a downturn. After the war, economic and monetary resources are needed to rebuild the places destroyed in the war and people start working for a living. Slowly, demand increases and the economy begins to recover.
2. Technology — In the modern world, new technologies are developed every day [16]. Investment and development of technology leads to new investment. This increases demand and employment, thus boosting the economy.
3. Natural disaster — The economy can experience a recession in the event of a natural disaster [17]. Hurricanes, earthquakes, droughts, and floods can cause large costs to governments, businesses, and individuals.

2.3 How to predict a recession?

Over the years, economists have searched for a way to predict future phases of a business cycle. After the Great Recession of 2008, the ability to predict potential financial crises has become more important than ever. The most commonly used tool for predicting recessions is the yield curve. The yield curve has a solid theoretical foundation and is a well-known concept in economics [18]. It is easy to use and significantly outperforms other macroeconomic indicators in predicting recessions [19]. The curve plots the yield on a bond against time to maturity [19]. Analysts carefully watch how the slope of the curve changes in the context of a recession. Usually the curve is upward sloping and convex, but when the curve flattens or is downward sloping, it means that GDP has declined and a recession is imminent in the near future.

Many analysts use special econometric models to predict the onset of a recession [20]. As mentioned earlier, the business cycle consists of phases of expansion and contraction. However, the challenge lies in the transitions between expansion and contraction when these special models are useful. During expansion, for example, when there is no threat of recession, analysts turn to large-scale econometric models for their forecasts. Nevertheless, these models can lead to large forecast errors at turning points in the business cycle. Therefore, small-scale forecasting models can play an important role in predicting future economic activity when an expansion is coming to an end [20]. The best known small-scale business cycle models are: simple rules of thumb using the Conference Board's composite index of leading indicators (CLI), Neftçi's probability model of imminent recession using the CLI, a regression based model of the probability of recession called the Probit model, a GDP forecasting model, and a recession prediction model proposed by Stock and Watson [20]. All five small-scale business cycle models can provide reliable information about future recessions. Of course, the models are not error-free, so it is important to monitor their performance to determine when they are likely to predict correctly and when they are not.

In recent years, artificial intelligence has been successfully applied in various economic studies [18]. Many machine learning techniques are used for modeling and predicting economic activities and recessions in particular [21]. James et al. [22], for example, used Support Vector Machines (SVM) to predict the onset and end of recessions in real time. Nowcasting, the prediction of the present, near future, and recent past of an economic indicator, is crucial for recessions that are detected months after they occur. The authors have shown that Support Vector Machines have excellent predictive performance for this task. They captured all six recessions from 1973 to 2018 and provided the signal with minimal delay. In addition, Vrontos et al. [21] applied various machine learning techniques to predict recessions and compared the results with other econometric techniques, such as multiple Logit and Probit models. The techniques used included regularization techniques such as Ridge, Least Absolute Shrinkage and Selection Operator (LASSO) as well as Elastic Net, Discriminant Analysis Classifiers, Bayesian Classifiers and Classification and Regression Trees (CART) such as Bagging, Random Forests, and Boosting. The results confirmed that the application of machine learning techniques represents a further advance in recession forecasting compared to traditional econometric techniques. In the vast majority of cases, forecasts produced using machine learning models significantly improve predictive ability and accuracy compared to Logit and Probit models.

Analysis of Spatial Time Series

Spatial time series analysis involves the analysis of data with temporal attributes and absolute and relative positions in 3D space, as well as the process and methodology of analysing these data. The following sections provide general characteristics of spatial time series data as well as some challenges, followed by representations of different types of data that exist in the spatial time series literature.

3.1 Characteristics and Challenges

Spatial time series analysis is an expanding area of research that has accompanied the development of powerful computer processors. Spatial time series data consist of spatially and temporally recorded data describing a phenomenon at a specific location and time period [23]. Analysis in statistical/econometric models describing spatial time series data began in the mid-seventies and has increased substantially in the last decade [24]. In 1975, Cliff and Ord [25] first created a model for the relationship between two variables in space and time. Since then, much research has been done and various techniques have been developed.

Nowadays, spatial time series data are widely used in economics, public health, military, environment, geology, geography, ecology, etc. These data are easily accessible and abundant. The analysis of these increasingly widespread data enables the advancement of the state of the art in various scientific disciplines. The main difference between spatial time series data and other data already found in the classical data mining literature [26], is the existence of dependencies between measurements derived from spatial and temporal dimensions. Two major challenges arise in the analysis of spatial time series data: first, the continuous and discrete changes in spatial and non-spatial properties of spatial time objects, and second, the influence of collocated neighboring spatial time objects on each other. Another challenge in interpreting the results of spatial time series analysis is that space is two-dimensional and can move in an unlimited direction, while time can only

move forward, i.e., it is one-dimensional. Many widely used data mining techniques rely on the assumption that data instances are independent and identically distributed, but this assumption is violated in spatial time series analysis. Spatial time series data are structurally related to each other in the context of space and time, and have different characteristics in certain spatial domains and time periods [27]. If these characteristics are ignored in data analysis, it would lead to poor accuracy and interpretability of the results. Considering that data instances can be defined as objects and features in different ways when analyzing spatial time series, there are several approaches to analyzing this type of data. One approach is to treat spatial locations as objects and use the measurements collected at a spatial location over time to define the features [27]. An example of this can be found in climate science, where one of the goals is to group places that experience similar climate over time. In this case, places are treated as instances/objects and characteristics are defined based on climate variables measured over time [28]. Another approach treats time points as objects and uses measurements collected at all spatial locations under consideration to define features [27]. For example, in the study of the human brain, the goal is to identify the time points at which similar brain activity is detected in the brain. Thus, the time points are treated as instances/objects and the features are defined using the observed spatial map of activity [29]. There are some other approaches that treat events as objects, and the features are defined based on the spatial and temporal (time series) information of the events. An example of these approaches is the detection of crimes committed in close spatial and temporal proximity, where the occurrence of a crime is treated as an object and the location and time of the crime are treated as features, including other features such as the type of crime and the number of victims involved [27]. As can be seen, the coexistence of spatial and temporal components in spatial time series data brings some new challenges, but also some new opportunities for the analysis of spatial time series data.

3.2 Data Types

There are a large number of spatial time series data types that differ in the way space and time are used in the process of data collection and representation [27]. To choose the most effective data mining and visualisation technique, it is important to define the type of spatial time series data. There are usually four categories of spatial time series data: event data, trajectory data, point reference data, and raster data. Each category is described in the following subsections.

3.2.1 Event Data

In general, a spatial time event can be described by a point, a place, and a time that indicate where and when the event occurred [27]. For example, a crime may be described by the location of the crime and the time at which the crime occurred. Similarly, an outbreak of disease can be defined by the location and time when the patient first became infected [27]. Other real-world examples of spatial time series data include

traffic, social media, earth sciences, etc. A collection of spatial time point events is called a spatial point pattern [30]. Figure 3.1 shows an example of a spatial point pattern in a two-dimensional Euclidean coordinate system, where (l_i, t_i) denotes the location and time point of an event. In addition, each spatial time series event can also consist of non-spatial time series variables, called marked variables, which provide additional information about each spatial time series event. In the Figure 3.1, the marked variable can take three categorical values: A, B, and C [27]. Existing data mining techniques for event data consider points of spatial time events occurring in Euclidean space. In some cases, however, the distance between two events is measured not by the Euclidean distance but by the shortest distance, and events are not always point objects in space but can also be lines or polygons [27].

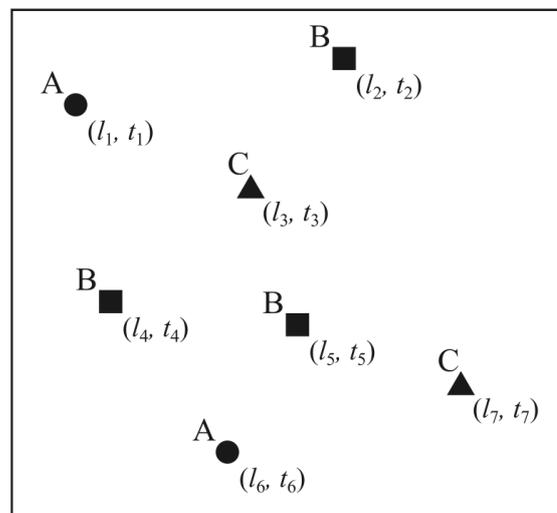


Figure 3.1: Events belonging to three types: A(circles), B(squares) and C(triangles) [27]

Visualizing Event Data

The most popular visualization techniques for event data are event maps and 3D views. These techniques are briefly discussed in the following sections.

- Event maps — Spatial events can be displayed as points in space and time and visualized as icons on a map. In event maps, the icons representing an event are distributed on the map. Although the event map is often the main view, it can be combined with various other representations to show the temporal components of the data. Examples of such representations include interactive timelines, temporal distribution displays, and temporal brushes. These allow the user to visualize time based on a specific value, see the distribution of events over a period of time, and specify the time interval for the visualization. A common type of interaction in this type of event map is the ability to highlight icons to get more information

about the event. Other interactive elements include time-related activities such as advancing the time and adjusting the time interval ("brushing"). Filtering the data by attributes, areas, or times is a common interaction, as is zooming and panning the map [31]. An example of an event map is shown in the Figure 3.2.



Figure 3.2: An example of an event map using icons to represent events, in this case crimes, on a map. Different icons and colors represent different crimes [32]

- Visualizing spatial time series events in three dimensions — Three-dimensional views are used to represent the temporal aspects of data, and can be important for visualizing the motion of objects in three dimensions. In three-dimensional visualizations, the temporal component can be represented by an axis so that the visualization can show both the temporal and spatial components of the data in a single overview. The most commonly used technique is a space-time cube (event cube) shown in the Figure 3.3, where the X and Y axes represent the spatial components and the Z axis represents the temporal components [31].

3.2.2 Trajectory Data

Trajectories are the paths taken over time by bodies moving in space [27]. For example, the route taken by a taxi from its pick-up location to its drop-off location, or animals migrating from their current location to another [34]. Trajectory data are usually collected by attaching sensors to moving bodies that periodically transmit information about the body's location over time. In the Figure 3.4, the trajectories of three moving bodies: A, B, and C are shown. Like event data, trajectory data can also contain labelled variables, such as the heart rate of a runner running on a circuit or the sequence of messages exchanged by a patrol car [27].

Visualizing Trajectory Data

- The most popular visualization technique for trajectory data is visualization in three dimensions — As with event data, three dimensional visualizations can be used for

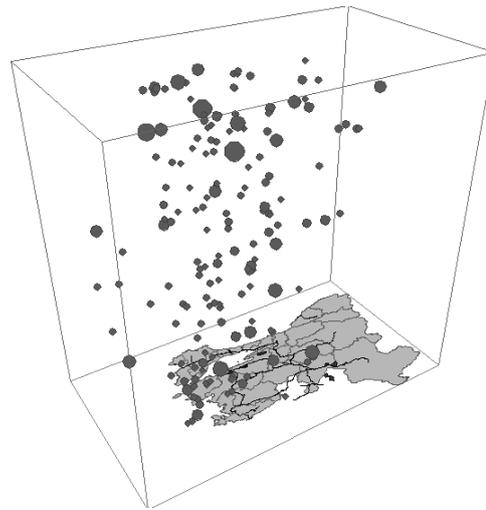


Figure 3.3: Data about events are presented in a space–time cube. The vertical positions of the circles correspond to the times when the events occurred [33]

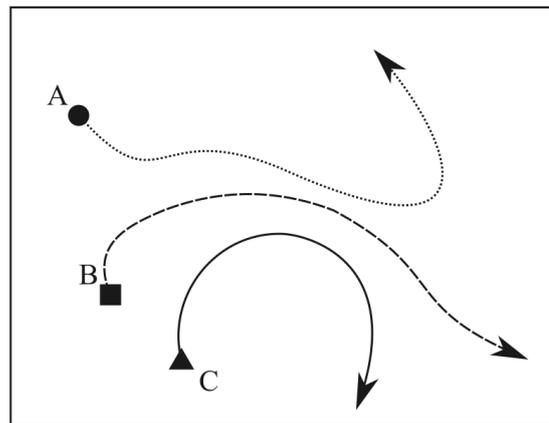


Figure 3.4: Trajectories of three moving bodies, A, B and C [27]

trajectory data. Flight data is an example, since aircraft move in a 3D space, it is sometimes useful to view the 3D trajectory. The space–time cube shown in the Figure 3.5, is most commonly used for trajectory data. However, there are some other techniques such as tubes and ribbons or stack based trajectories. In the tubes and ribbons method, the user gets a clearer indication of the elevation movement over time and the color of the trajectories denotes other thematic attributes of the data such as velocity or inclination. The stack based approach shown in the Figure 3.6, uses color-coded bands to create a series of trajectories by stacking the bands along a path. A two-dimensional map forms the basis for the visualization, and the trajectory data are mapped along the Z–axis. This approach helps to

prevent overplotting [31].

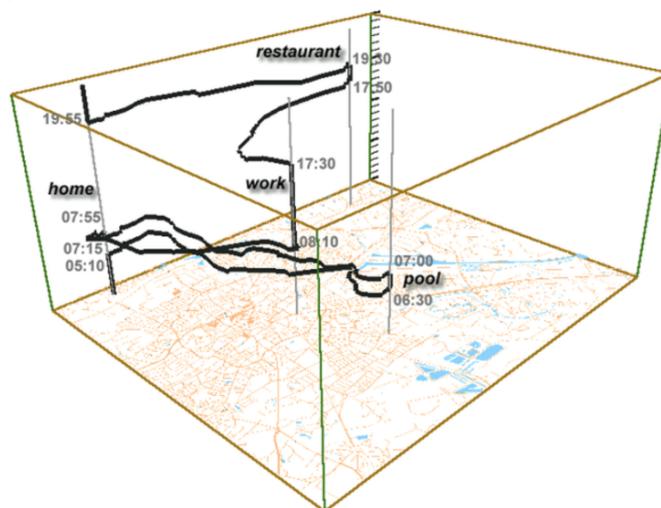


Figure 3.5: Space time trajectory cube represents the travel of one person on a typical work day [35]

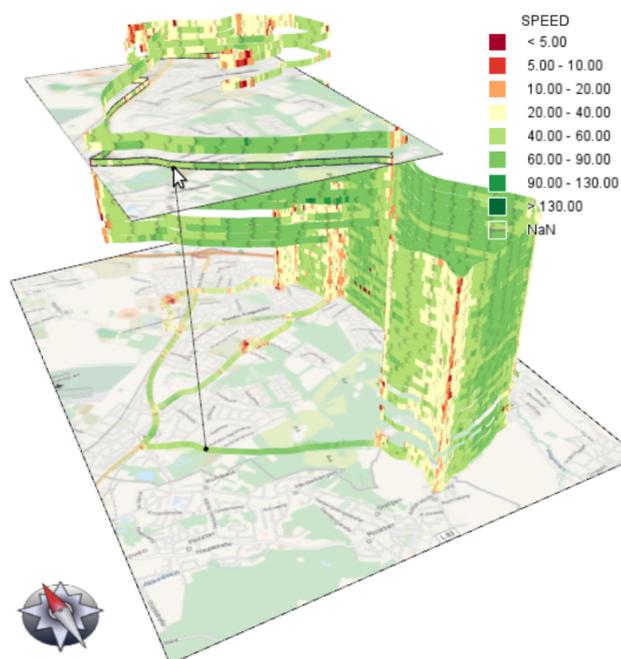


Figure 3.6: Visualization of trajectories as colored 3D bands and 2D paths [36]

3.2.3 Point Reference Data

Point reference data contain continuous spatial time series measurements such as temperature or population over a set of moving reference points in space and time [27]. Examples include measuring temperature and humidity using weather balloons floating in space, or measuring sea surface temperature at different locations using buoy sensors. Figure 3.7 shows an example of the spatial distribution of a continuous spatial time series field at two different timestamps measured at reference points (shown as dots) on the two timestamps. The observations at these reference points can be used to reconstruct the spatial time series field at an arbitrary location and time using data-driven methods (e.g., smoothing methods) or physics-based methods (e.g., meteorological reanalysis approaches) [27].

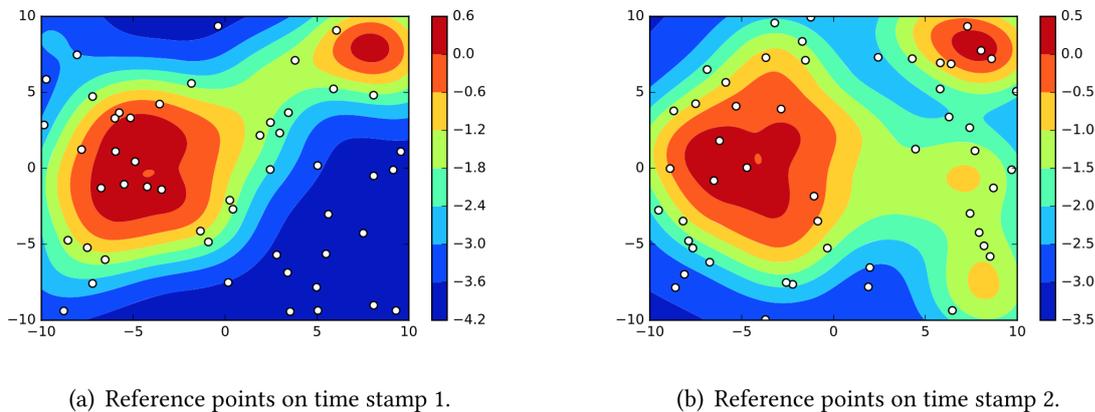


Figure 3.7: An example of spatial time series reference points (shown as dots) on two different times stamps [27]

Visualising Point Reference Data

It is difficult to determine visualizations that are relevant only to this type of data. For this reason, most visualizations can be represented by adding animations or dividing the display into multiples, where each multiple represents a point in time [31].

3.2.4 Raster Data

Raster data, unlike point reference data, consists of measurements of a continuous or discrete spatial time series field recorded at fixed locations in space and at fixed times [27]. Raster is described as a set of fixed locations, $S = s_1, s_2, \dots, s_m$, either regularly distributed in space with constant distance between adjacent locations, or distributed in an irregular spatial pattern, as shown in Figure 3.8. For each location, there are record observations at a fixed set of timestamps, $T = t_1, t_2, \dots, t_n$, which in turn may be regularly spaced or irregularly spaced (Figure 3.9). The Cartesian product of S and T gives the complete spatial time grid, $S \times T$, in which each vertex of the ST grid, (s_i, t_j) , has a particular measurement [27].

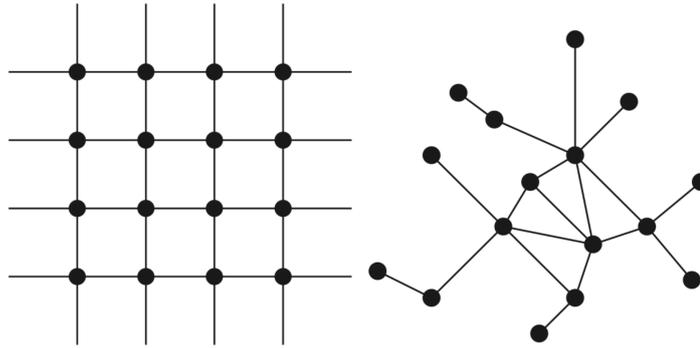


Figure 3.8: The set of locations in the spatial time series raster can either be located regularly or irregularly in space [27]

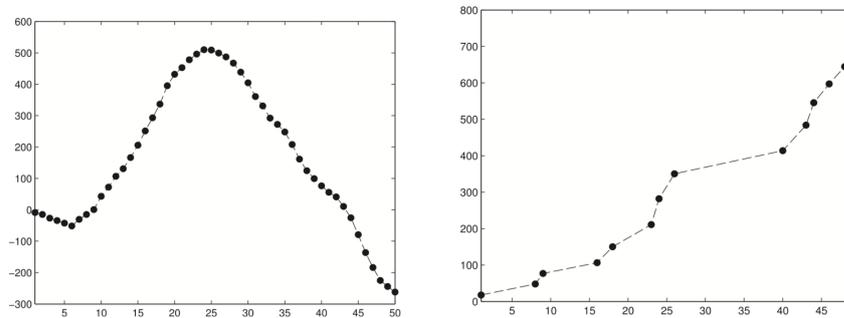


Figure 3.9: The set of time stamps can also be either spaced regularly or irregularly [27]

Visualising Raster Data

The most popular visualization techniques for raster data are raster maps, isarithmic maps, and heat maps. These maps are briefly explained in the following sections.

- Raster map — Raster maps are represented by a grid system that indicates the pattern of spatial time series data shown in the Figure 3.10. In addition, raster maps can be presented in three dimensions, with some attributes of the data displayed along the Z axis [31].
- Isarithmic map — Isarithmic maps, also known as contour maps, can represent smooth, continuous data. There are many ways to represent these types of maps, but two popular representations are the isopleth map, shown in the Figure 3.11, and the isometric map. These maps are suitable for displaying the surface temperature in an area because the temperature is continuous and does not have abrupt changes in value or show a continuous and gradual change across space [31].
- Heat map — A heat map is a graphical representation of data where values are indicated by colors shown in Figure 3.12. There are many types of heat maps

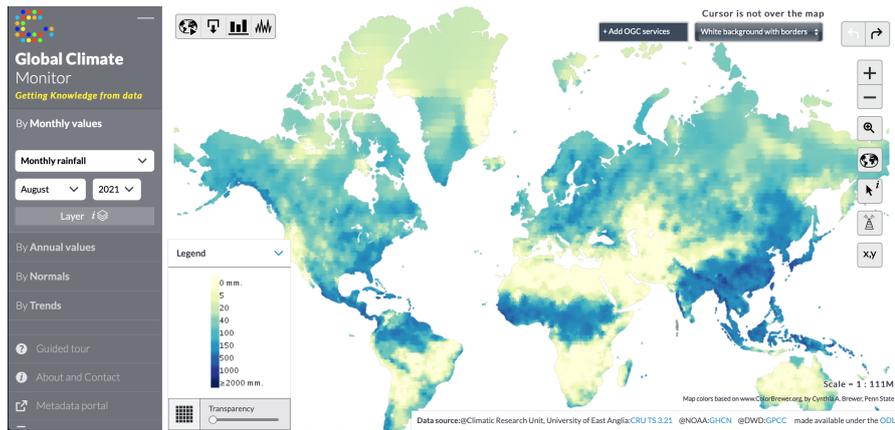


Figure 3.10: Raster map showing the monthly rainfall in August 2021 [37]

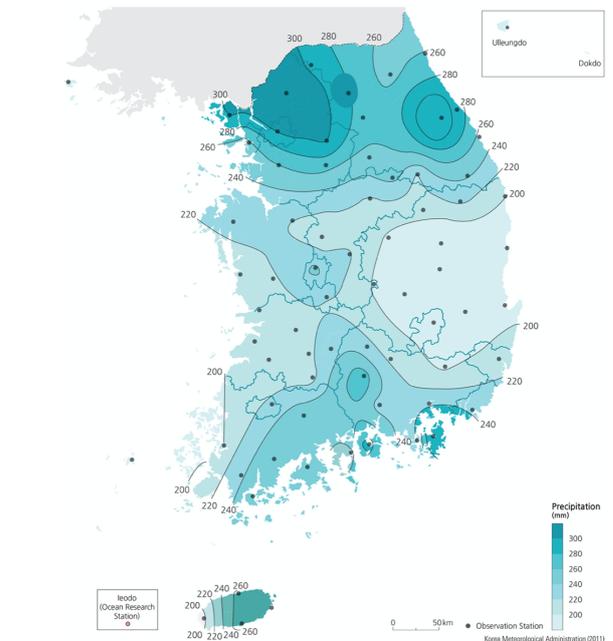


Figure 3.11: Isopleth map showing annual precipitation averaged over 30 years (1981–2010) and homogeneous rainfall zones for annual total precipitation in South Korea [38]

and they are often represented as a matrix where each individual value in the matrix is indicated by a different color. Heat maps can also be represented by a 2D square mosaic raster plot, either alone or over a geographic map to represent spatial information, similar to a raster map [31].

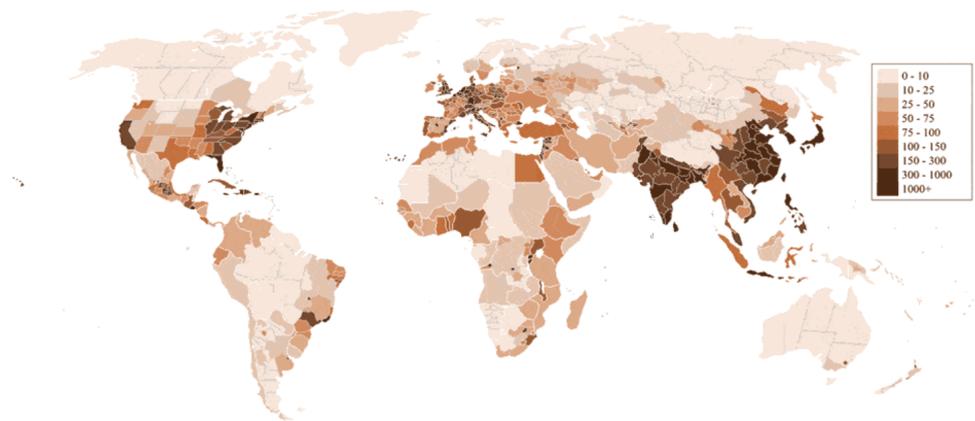


Figure 3.12: Heat map showcasing population density [39]

3.2.5 Converting Data Types

In order to analyse data using the appropriate data mining technique, it is sometimes beneficial to transform one data type into another. Possible transformations between spatial time series data types [27]:

- An event data type can be transformed into a raster data type by aggregating the counts of events in each cell of a ST grid
- Raster data can be converted to event data by using algorithms for event extraction
- Raster data can also be converted into point reference data by considering each vertex of the ST grid as a ST reference point.
- Point reference data can be transformed into a raster data by interpolation or aggregation over a ST grid

Visual Analytics and Spatial Time Series Data

Due to advances in computer technology in recent decades, data is being produced at an incredible rate. As the amount of this data keeps increasing and it is available to everyone, it has become a challenge to use it efficiently. Visual analytics seeks to find better and more efficient ways to understand and analyse large amounts of data while being able to act on the results immediately and in real time. By combining the analytical capabilities of the computer with the skills of the human analyst, visual analytics offers new insights and empowers individuals to take control of the analytics process [1]. Although there are many powerful tools for automated data analysis, they only work accurately for well-defined and well-understood problems. Automated data analysis tools represent the knowledge of the person who created them, but they lack the ability to communicate that knowledge. And that ability is essential. That is why there is a complex path from raw data to a decision. If that decision turns out to be wrong, it is important to be able to analyse the process that led to that decision. The goal of Visual Analytics is to make the path of data and information processing transparent for analytical discussion. Visual Analytics promotes constructive evaluation, correction, and rapid improvement of processes and models, ultimately improving knowledge and decisions [1]. Visual analytics is deeply interdisciplinary, integrating several complementary disciplines (shown in Figure 4.1) such as visualisation, data mining, data management, data fusion, statistics, and cognitive science. The core idea of visual analytics is to integrate all these areas into one scientific discipline. Experts in the application areas know that visualisation is useful and valuable, but the solutions used are often ad-hoc, and hardly reflect the state of the art in interactive visualisation and the full complexity of the problems for which visual analytics seeks answers [1].

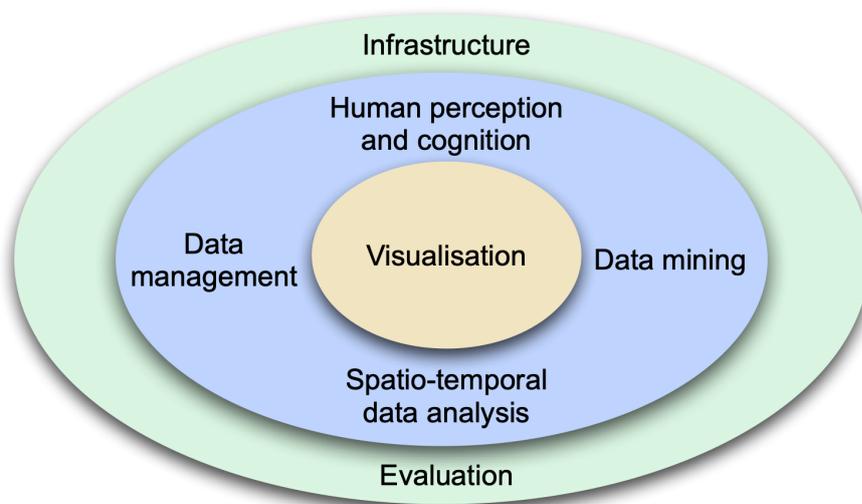


Figure 4.1: Disciplines which contribute towards visual analytics [1]

The first steps toward visual analytics were taken in the 1970s, when the focus shifted from confirmatory data analysis, in which only charts and other visual representations were used to present results, to exploratory data analysis, where the user began to interact with the data and results [40]. As graphical user interfaces and interaction devices advanced, information visualisation became more interesting. When the potential to engage the user in the knowledge discovery and data mining process through beneficial visualisation techniques was discovered, it led to numerous researches in the field of visual data analysis and visual data mining. Nowadays, the term visual analytics is used in a broader context, describing a new multidisciplinary field that combines different areas of study, including visualisation, human–computer interaction, data analysis, geospatial and temporal data processing, spatial decision support, and statistics [1].

As already mentioned, visual analytics combines different research fields, so it is not easy to define it. Nevertheless, the currently accepted definition is “Visual analytics combines automated analysis techniques with interactive visualisations for effective understanding, reasoning, and decision–making based on very large and complex datasets” [1]. Visual Analytics, then, develops tools and techniques that enable people to [1]:

- Combine information and gain insights from large, dynamic, confusing, and often contradictory data
- Discover the expected and the unexpected
- Produce timely, valid, and understandable assessments
- Communicate these assessments effectively for action

To achieve the goal of visual analytics, an overarching process for visual analytics must be defined. The visual analytics process combines automated and visual analysis methods with a human interaction to extract knowledge from data [1]. Figure 4.2 shows an abstract overview of the stages (shown as ovals) and their transitions (arrows) in the visual analytics process.

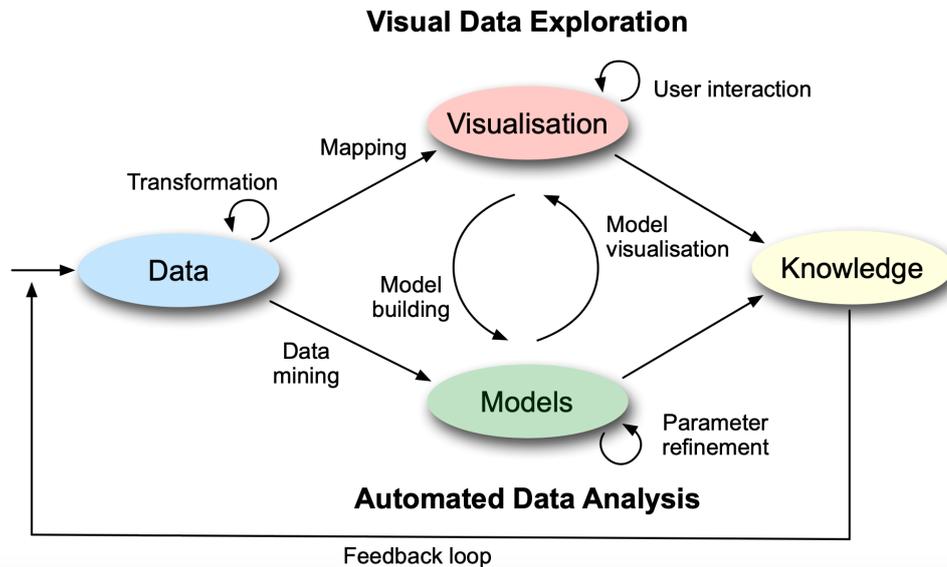


Figure 4.2: The visual analytics process is described by interaction between data, visualisations, models about the data, and the users in order to discover knowledge [1]

The first step in the visual analytics process is preprocessing and transforming the data to derive different representations for further analysis (indicated as Transformation arrow in Figure 4.2). Other common preprocessing tasks include data cleaning, normalisation, grouping, or integration of different data sources. In the next step, the analyst decides whether to use visual or automatic analysis methods. Assuming automated analysis is chosen, data mining methods are used to create models. Once a model is created, the analyst needs to validate and possibly improve the model, which is best done by interacting with the data using visualisations. In this way, analysts can interact with the automated methods by changing parameters or choosing different algorithms. This helps identify inaccurate results early on, which improves results and leads to higher confidence. In the other case, when visual data exploration is chosen, the analyst has to confirm the developed hypothesis. The insights from the visualisation can be used to automatically create analytical models. In summary, the insights gained in the visual analytics process come from either one or a combination of visualisation, automated analysis, models, and the human analyst [1].

4.1 Spatial Time Series in Visual Analytics

In daily life, people deal with space and time and most decisions are made depending on space and time. To deal with the current problems of modern society, people need to understand the global processes that affect demography, economy, environment, energy, epidemics, food, international relations, and other factors [1]. It is necessary to know how typical features of these processes change, evolve, and correlate in space and time. Advanced computer techniques for data analysis and sophisticated maps are needed to address these questions. In today's world, space-time analysis is no longer reserved for scientists and experts, but for anyone who wants to plan a trip, look for a job, or find a new place to live. Analysts should find a way to create accessible and usable analysis tools for the wide range of potential users to help solve a variety of problems.

Spatial time series describe the change in attribute values over time at numerous spatial locations [41]. For a given location, the data define the values of one or more attributes at different time fragments or intervals called "time steps" [42]. Prior to any analysis, it is important to ensure that the data has adequate coverage and quality. While there are fairly detailed tools for inspecting features of tabular data, these do not address the specifics of spatial and temporal data, such as time cycles, spatial and temporal autocorrelation, or the subtleties of many trends in space and time, such as the presence of geographic boundaries and barriers. Considering the task and design conditions of experts in this field for a visual analytics tool, the following section presents some research on visual analysis of spatial time series data.

4.1.1 Modeling of Spatial Time Series

Kamariankis and Prastacos [24] have conducted a study on the techniques proposed for modelling spatial time series, in particular STARIMA (spatial temporal autoregressive integrated moving average). STARIMA specifies each observation at time t and location i as a weighted linear combination of past observations and neighbouring observations that are both spatially and temporally lagged [2]. This involves the prior specification of a set of weight matrices, where the weights characterise the effect between locations for different time lags. The analyst specifies the weights. One of the drawbacks of the STARIMA model is precisely its dependence on the analyst to select the correct weights, since the quality of the model depends strictly on how well the between-location effects are described by numerical weights. Another drawback is that the model is a kind of "black-box" and the user cannot easily understand the behaviour of the model [2]. Ziegler et al. [43], used clustering to facilitate visual exploration of a huge dataset of financial time series. Before applying clustering, the time series data were generalised and compressed, which increased the effectiveness of clustering and extracted temporal changes from inconsistencies. Guo [44] and Andrienko et al. [45], applied self-organising maps to spatial time series, determined colours of clusters, and used these colours to paint areas in map representations and lines in time graphs. Another development in the field of visual analytics is Time Mask [46]. Time Mask is a technique for time-based data retrieval

and filtering. It works with all forms of time-based data, selects entire time intervals where certain conditions of interest are met. Then, selects fragments from all available time-related data whose time references match the selected intervals. The advantage of the Time Mask filtering tool is that analysts can select time intervals based on the condition of one event and examine how the other event evolved in those intervals [47]. Since social media produces a huge amount of data, much research has been done in this area as well. Luo and MacEachren [48] have developed a theoretical and scientific infrastructure for geo-social visual analytics. Subsequently, more research has been conducted on this topic: analysis of people's reactions to meaningful events [49], proposed visualizations and analytical workflows for the combined analysis of the temporal, spatial, social and thematic aspect of social media data [50] [51]. The research field of visual analytics also includes the development of predictive models [47]. Several concepts have been proposed in relation to space and time in visual analytics. Some of them are predicting areas with higher probability of epidemic outbreak [52], modeling traffic flows under normal and exceptional circumstances [53], predicting the distribution of crime incidents, and decision support for resource allocation [54]. These concepts allow analysts to combine their domain knowledge and observations from the model. Consequently, they assist analysts in verifying their findings and help them in decision making.

As can be seen, the current direction of visual analytics in relation to space and time focuses on advancing the state of conceptualization of spatial time series data and new systematic methods for data analysis, comprehensive analysis of data properties and semantic interpretation of data, advancing techniques and tools for task-oriented data transformation and analysis, and closing the gaps between data exploration, analysis, and model building [47]. Although information visualisation and geovisualization have developed some powerful methods to support the visual analysis of spatial and time-dependent data, efficiently supporting the analysis of spatial time series data is a much greater challenge.

Prediction Methods

An ongoing trend in visual analytics is the development of machine learning models whose behavior is understandable and explainable (often referred to as “eXplainable AI”) [55]. Linear regression, as an example of a white–box approach that enables interpretability, was chosen for building a machine learning model. This chapter introduces the field of machine learning and all the methods and techniques used in this thesis.

Machine learning is a type of artificial intelligence (AI) that automates the creation of analytical models. Machine learning focuses on using data and algorithms to mimic the way humans learn, with incremental improvements in accuracy. In machine learning, there are two types of data — labeled data and unlabeled data [56]. Labeled data has both input and output parameters, while unlabeled data has only one or none of the parameters in a machine–readable form [56]. Based on the data, there are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the machine learning algorithm is trained on labeled data and the goal is to correctly label unknown data. In unsupervised learning, the algorithm works on unlabeled data and the goal is to find structures and associate rules learning. In reinforcement learning, the algorithm improves itself and learns from new situations through a trial–and–error method [56].

When the goal of a machine learning task is to predict an outcome using one or more input variables, the task may be referred to as supervised learning. Depending on the nature of the outcome variable, supervised learning can be further grouped into regression and classification problems. In classification problems, the learning algorithm learns a function to map inputs to outputs, where the output value is categorical [57]. In contrast, regression problems involve mapping inputs to outputs, where the output is a continuous real number. Unsupervised learning refers to any machine learning process that attempts to learn a structure without a specific output or feedback. Three common examples of unsupervised learning are clustering, association rules, and self–organising maps [58]. In clustering, objects are divided into different groups so that the common properties of the

data are high in each subset and low between different subsets. K-means is the most commonly used clustering algorithm. It creates a partitioning of a set of objects into k clusters that minimize an objective function, usually a squared error function that implies round shaped clusters. The input parameter k is fixed and must be specified in advance, which limits its actual relevance to streaming and evolving data [58]. In this thesis the focus is on both supervised and unsupervised learning covering regression task and K means clustering for classification task.

5.1 Definition of Linear Regression

In a regression problem [58], the objective is to recover a functional dependency $y_i = f(x_i) + \epsilon_i$ from N -observed training data points $\{(x_i, y_i)\}_{i=1}^N$, where $y_i \in \mathbb{R}$ is the noisy observed output at the input location $x_i \in \mathbb{R}^d$.

In linear regression, the regression function $f()$ is represented by a parameter $w \in \mathbb{R}^H$ in the form

$$f(x_i) := \langle \phi(x_i), w \rangle,$$

for H fixed basis functions $\{\phi_h(x_i)\}_{h=1}^H$. The simplest form of the linear parametric model is when $\phi(x_i) = x_i \in \mathbb{R}^d$, that is, the model is also linear with respect to the input variables,

$$f(x_i) := w_0 + w_1 x_i^1 + \dots + w_d x_i^d.$$

Here the weight w_0 allows for any constant offset in the data.

Fitting

Typically, regression fits a model to the data using an objective function or quality criterion in the following form [58]

$$E(f) = \sum_{i=1}^N \epsilon(y_i, f(x_i)),$$

where smaller $E(f)$ implies better quality. This can be derived as an error/loss function or as a negative log-likelihood or log probability. The squared error function is the most convenient (it leads to a least squares calculation), but there are other possibilities. Typically, methods are divided by three conditions:

1. the representation of the function $f()$,
2. the form of the term $\epsilon(y_i, f(x_i))$
3. the penalty term

Regularized/Penalized Fitting

The problem of overfitting is generally addressed by introducing a regularization or penalty term into the objective function [58]. The regularized objective function now has the form of

$$E_{\text{reg}} = E(f) + \lambda R(f),$$

where $E(f)$ measures the quality of the solution for $f()$ on the observed data points, $R(f)$ penalizes the complexity of $f()$, and λ is called the regularization parameter that controls the relative importance between the two. This regularized formulation is sometimes called coefficient shrinkage, since it shrinks coefficients/weights towards zero. Next, three simple penalty terms $R(w)$ are given.

Ridge Regularization

Ridge regularization is given in the form of [58]

$$R(w) = \sum_{d=1}^D w_d^2.$$

Considering $E(w)$ to be in form of

$$E(w) = (Xw - y)^T (Xw - y),$$

regularized least squares quality function is now

$$(Xw - y)^T (Xw - y) + \lambda w^T w.$$

Since the additional term is a quadratic of w , the regularized objective function is still quadratic in w , so the optimal solution is unique and can be found in closed form. By setting the first derivative of regularized least squares quality function with respect to w to zero, the optimal weight vector has the following form

$$\begin{aligned} \partial_w E_{\text{reg}}(w) &= 2X^T(Xw - y) + 2\lambda w = 0 \\ w^* &= (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

The regularization term has the effect of giving small weights to basis functions that are only marginally useful, since the penalty for small weights is very small. So by reducing the constraint (low λ) on the features, the model will resemble a linear regression model.

Lasso Regularization

Lasso regularization is given in the form of [58]

$$R(w) = \sum_{d=1}^D |w_d|.$$

Unlike ridge regression, lasso regression does not have a closed-form solution. In fact, the nondifferentiability of the regularization term has spawned many approaches. Most methods involve quadratic programming and, more recently, algorithms for coordinate-wise descent on large lasso problems. Lasso regression leads to sparsity in w , i.e., only a subset of w is nonzero, so irrelevant basis functions are ignored.

Elastic Net Regularization

Elastic Net was born out of criticism of lasso, whose variable selection depends too much on the data and can therefore be unstable [58]. The solution is to combine the penalties of Ridge Regression and Lasso to get the best of both worlds. The elastic net estimator minimizes following function

$$\text{EN}(\beta) = \sum_{i=1}^n \left(y_i - x_i^\top \beta \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p |\beta_j|^2,$$

where λ_2 is the mixing parameter between ridge ($\lambda = 0$) and lasso ($\lambda = 1$).

Design of Visual Analytics Process

In Chapter 1, the motivation and the problem statement of this thesis were presented. In Chapter 2 the definition of recession and its causes were given, as well as how to predict it. The analysis of spatial time series was explained in Chapter 3 as a basis for the visual analytics of spatial time series data, which was discussed in Chapter 4. In Chapter 5, the field of machine learning and all the methods and techniques used were presented. In order to support the findings of the systematic literature review and answer the research question, this chapter presents the application Recession Explorer: visual analytics of economic recession and its forecasting, which was designed as the main contribution of the work carried out in this thesis.

Based on the visual analytics process presented in Chapter 4, the bespoke process for implementing the Recession Explorer application was defined. In order to develop a customised process, the original visual analytics process was adapted to the specific problem of spatial time series analysis. The result can be seen in Figure 6.1.

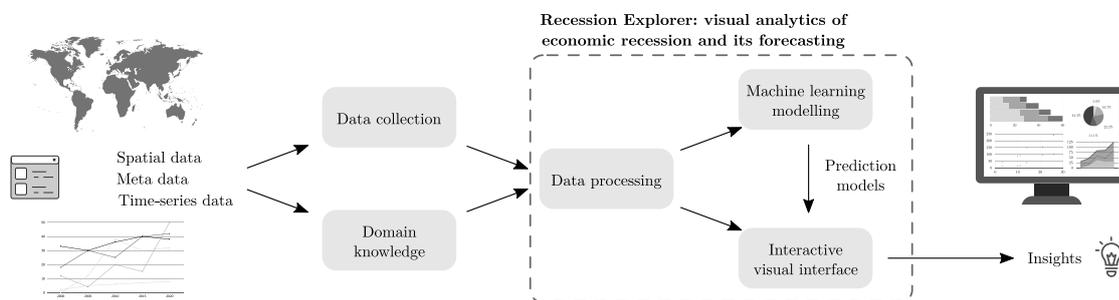


Figure 6.1: Customised Visual Analytics Process for Spatial Time Series

The development of the Recession Explorer application followed the tailored visual analytics process for spatial time series. The first step is data collection, i.e. the spatial time series to be analysed, and domain knowledge based on experience and previous analysis. The second step is the data preprocessing. The third step is modelling through machine learning and visualisation of the data through an interactive visual interface. Insights are gained through the interactive visual interface.

Implementation is divided into two tasks, the machine learning task and the visual analysis task. The former is further divided into a regression task and a classification task. The objective of the machine learning task is to predict the recession factors and future recessions. The dataset contains various recession factors, such as GDP, inflation rate, employment rate, Gini index, trade, investment, tax revenue, etc. which are obtained from Gapminder [3]. As described in Chapter 2, there are exogenous and endogenous recession factors. The recession factors used in building the machine learning model are relevant endogenous factors whose changes are indicative of changes in the phases of the business cycle and thus can indicate recession and non-recession periods.

The latter task is the task of visual analysis. This task is dedicated to the development of an interactive visual interface. Due to the characteristics of spatial time series data discussed in Chapter 3, it is important to select appropriate visualization techniques. For this purpose, the systematic review of visualizations for time-oriented data by Aigner et al. [59], was used. According to Aigner et al. [59], the systematic review is structured by three practical questions:

1. What is presented? Space, time, and data.
2. Why is it presented? To facilitate the user's tasks.
3. How is it presented? Through the use of visual representations.

The answer to the first question contains characteristics of spatial time series data discussed in Chapter 3. The answer to the second question expresses the reason for the visualisation. Finally, the answer to the third question answers the question of what visualisations are appropriate for the given data to assist users in their tasks.

User Stories are used to help define user tasks. User stories are a popular way to represent the requirements of a software product. In this case, it is the Recession Explorer application. Requirements formulations should be simple and easy to use in discussions within the development team or with customers. User stories help create a user-centric framework for day-to-day work that encourages collaboration and creativity, and leads to an overall better product. There are several templates for formulating requirements, but in this thesis a simple template from Cohn [60] was used “As a (type of user), I want (a specific goal), [so that (a specific benefit)]” [60]. Based on these user stories, the Recession Explorer application was developed:

- As a casual user, I want to select a particular country on a world map, so that I can study the economic situation of a the selected country
- As a casual user, I want to click on any chart, so that I can see more detailed view of that chart
- As a casual user, I want to examine a year of interest, so that I can see the values of all the recession factors clearly in that year
- As a casual user, I want to see all the changes in economy state, so that I can identify when is the next recession coming
- As a casual user, I want to select multiple countries, so that I can get insight into the trend comparison between the countries
- As a expert user, I want to see the data that other models have produced, so that I can compare the results

The following sections present the technologies used to develop the Recession Explorer application and explain the implementation of two defined tasks.

6.1 Implementation Technologies

Before explaining the details of the implementation process of an application, the following subsections present the technologies used for the implementation.

6.1.1 Python

As one of the most popular programming languages for machine learning and artificial intelligence, Python [61] has been used to accomplish the machine learning task. It is an open-source programming language and is supported by a wide range of resources and high-quality documentation. It also has a large community of developers who offer advice and support during the development process [62]. Python has a variety of libraries that make it easier to handle, access, and transform data. The libraries used for the machine learning task are scikit-learn [63] and pandas [64].

6.1.2 HTML, CSS, JavaScript

HTML, as the core of any web page, provides the basic structure of pages, which is extended and modified by other technologies such as CSS and JavaScript [65]. CSS determines how the HTML elements of a web page should appear on the front end of the page. JavaScript is a programming language that can be used to develop and design interactive web pages. To support the development of the visual analytics process, the use of HTML, CSS, and JavaScript was essential.

6.1.3 D3

D3.js (data-driven documents) is an open-source JavaScript library that allows users to apply pre-built data visualizations to their data. D3.js [66] is used for visualization in this thesis because it provides the full capabilities of modern browsers without a complicated framework to combine powerful visualization components and a data-driven approach.

6.2 Machine Learning Task

6.2.1 Regression Problem

In order to forecast the economic situation of a country in the future, it is first necessary to forecast recession factors. The recession factors that are forecasted are:

1. GDP (USD)
2. GDP per capita (UDS)
3. Inflation rate
4. Employment rate
5. Gini index
6. GDP distribution across economic sectors
7. Employed population by sector and age(% of total employment)
8. Trade (% of GDP)
9. Investment (% of GDP)
10. Tax revenues (% of GDP)

The training set consists of 86 countries. The dataset is a time series with a yearly time step, starting from 1992 to 2016. However, for some countries, data were not available for some years in this period, so the final training set consisted of 1167 instances. It is obvious that the current economic situation of a country depends on the economic situation of the previous years, but still the problem of the optimal time period for the prediction arises. As a compromise between rather long and immensely small time periods, the time period of one year is used for prediction, inspired by the Jung et al. [67], solution. All original variables, both inputs and outputs, were expended by the so-called returns. For an input variable x_n , where n denotes its value in year n , the return value is defined as

$$x'_n = 100(x_n - x_{n-1}). \quad (6.1)$$

In this way, the number of input variables was increased from 20 to 40 input variables. In order to compare the different variables in the model, data preprocessing was performed.

Data preprocessing consisted of data normalization by the arithmetic mean and standard deviation for all countries. For the input variable x its normalised value x^* corresponds to

$$x^* = \frac{x - \bar{x}}{\sigma_x},$$

where \bar{x} denotes the arithmetic mean x and σ_x denotes the standard deviation. Standardizing the input variables leads to a more desirable distribution, as shown in Figure 6.2 and Figure 6.3.

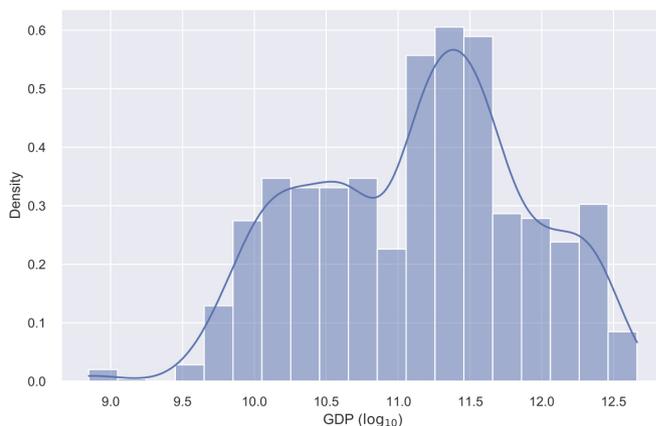


Figure 6.2: GDP Distribution

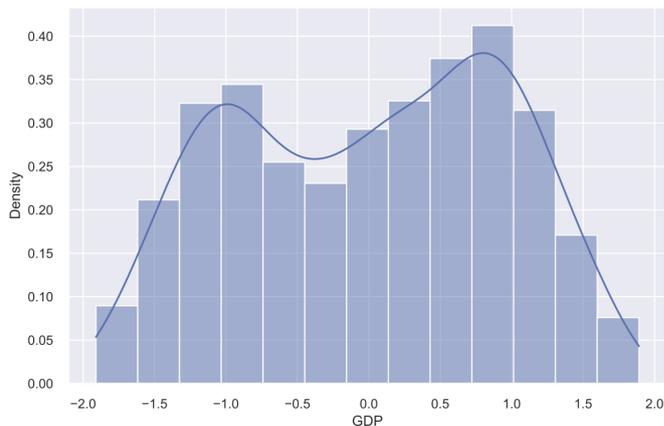


Figure 6.3: Standardised GDP Distribution

Inspired by Jung et al. [67], a linear regression model with elastic net regularization was first used for each of the 20 outputs. The parameters were optimized by 10 fold cross-validation. In accordance with the original solution, the models that tended strongly toward lasso or ridge regularization were replaced with actual lasso or ridge regularization.

The parameters were then optimized again by 10 fold regularization. The final model consists of a combination of individual models for each output variable. The challenge with this approach was the variables that were constant over the years or whose absolute growth was constant over the years. Apart from the fact that a simple linear regression for such variables would tend to the affine function $x_{n+1} = a + x_n$ (neglecting the remaining 39 observed variables), the regularized regression would tend to such a function in particular. In these cases, therefore, the models were forced to make predictions based on the remaining 39 variables by excluding the value of the variable they had predicted in the previous year from the input data and entering only the values of the remaining 39 variables into the input data. Although this approach increased the error to some extent, the models are more interpretable in the sense that they show in what way which economic factor depends on the general economic situation of the country. All variables predicted by the developed model for the entire data set had an error smaller than their standard deviation for the entire data set. Of the more important indicators (tax revenues, for example, do not provide as clear an insight into the economic condition of the country as the inflation rate), the errors for the GDP growth rate and GDP per capita were 2.7%, while their standard deviations were about 3,6%, the error for the employment growth rate was 1.6% while the standard deviation was 2.1%, and the error for the inflation rate was 25.8% while the standard deviation was 30.6%. Given their standard deviation and the results of the elastic net model from the article by Jung et al. [67], the errors in the GDP growth rate and GDP per capita are acceptable, but for example, the error in the inflation rate is very poor and the inflation prediction by the developed model is not reliable.

6.2.2 Classification Problem

Since predicting the recession was another part of the machine learning task, the regression model described in the previous section first predicts the economic factors that can indicate whether the recession has occurred or not. Unlike the regression problem, the classification problem considers only the returns defined in equation 6.1. Economic factors whose returns have been observed are:

1. GDP (USD)
2. GDP per capita (UDS)
3. Inflation rate
4. Employment rate
5. GDP distribution across economic sectors
6. Trade (% of GDP)
7. Investment (% of GDP)

Given the characteristics of the dataset, the classification model is constructed so that each year is modeled separately. A recession period can then be described as a series of consecutive years, each of which is classified as a recession. In this way, the recession normalization problem is simplified – as some recessions are shorter and others are longer, so are the non-recession periods. Prior to model building, the data points were transformed by dividing them by their root square mean. That is, for the variable x' , its transformation

$$x'^* = \frac{x'}{RMS(x')}$$

was used. The method used for this classification task was grouping with k-means, an excellent approach for unsupervised learning. Unsupervised learning was generally chosen because of the subtle differences between recession and non-recession periods, which can lead to inconsistencies due to the inconsistent definition of recession, but also because finding relevant data on when one of the observed countries was in a recession proved to be a major challenge. Apart from the fact that this model experimentally gave the best or slightly worse results than others of the observed unsupervised learning models, it is very interpretable. In fact, sometimes even without applying the model, it is possible to infer which category the given data belongs to from the parameters of the models computed during training. The optimal number of groups was determined experimentally. The best results were obtained with three or four groups, which can be considered as different economic phases of a country from recession to economic boom. Figure 6.4 shows how the model divides employment and GDP into four different categories. The economic situation assigned to a group is then chosen by observing which known economic situation belongs to which group.

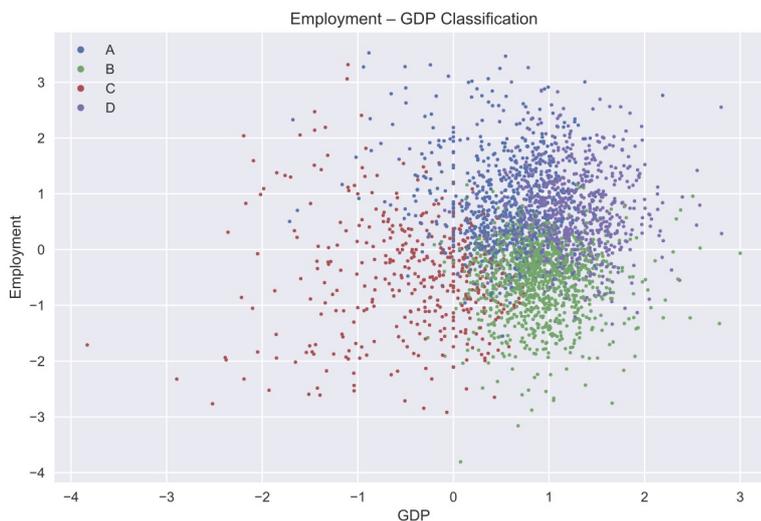


Figure 6.4: Example of Employment-GDP Classification with four categories

Consider, for example, the economic situation in the United States in 2008 when

the Great Recession hit. In the United States, models with three and four groups have successfully distinguished between crisis and non-crisis periods, since, for example, 2001 and 2008 belong to the same group — they are recession periods, and 2004 belongs to a different group — it is not a recession period. However, future recessions are not easy to detect even if the classification model is accurate, because the reliability of the regression model decreases with each predicted year. The economy also depends on external factors. Therefore, it is impossible to predict the crisis caused by the current pandemic COVID-19 because all the data used is older than 2019 before any pandemic could be predicted.

6.3 Visual Analysis Task

The graphical user interface of the Recession Explorer application follows the workflow of the customised visual analytics process. The implementation of the application was based on the user stories. To identify user tasks, the interactive world map served as a starting point. The interactive world map visualisation, as shown in Figure 6.5, was chosen to display the results of the model described in the machine learning task. In general, interactive maps are commonly used to visualise spatial time series data. The graphical visualisation of the data on the map makes it easier for the user to interpret and analyse the results and identify areas of interest in the data for further analysis. The Spherical Mercator projection coordinate system was used to project the world map by calling the `d3.geoMercator()` function and using an external GeoJSON file. The spherical Mercator projection defines a standard projection where the world is projected onto a square and trimmed to approximately $\pm 85^\circ$ latitude. The default centre of `d3.geoMercator` is $[0^\circ, 0^\circ]$ and to make the map centered on the canvas, the `translate` function is applied.



Figure 6.5: Main dashboard—interactive world map

In order to visualize different economic factors, different types of line and bar charts have been used, with the aim of discovering useful information more easily. Generally, line charts help to establish the relationship between two sets of values, where one set of values is always dependent on the other. Line charts are useful because they show data variables and trends very clearly and can help make predictions about the results of data that have not yet been collected. They can also be used to plot multiple dependent variables against one independent variable. When comparing datasets, line charts are only useful if the axes are of the same scale [68]. In the application, the line charts are used to show a trend or pattern in GDP, Gini index, inflation index, and income as shown in the Figure 6.6. In contrast, a bar chart is a type of column chart that is often used for comparisons. The bar chart has two variables, one is the length of the bar and the second is the position of that bar on one of the axes. The bars can be plotted vertically or horizontally in two dimensions. When different bars are plotted together, it is easier to compare them because the variable being compared is identified by the length of the bars. There are several options for comparing data sets with bar charts, such as stacked, grouped, and overlapping bar charts. A grouped bar chart represents a comparison between two categories based on multiple variables (multiple grouped bars). It is important that all of these variables can be measured with the same unit because these variables are plotted on the same axis [69]. In the application, the bar charts are used to show a distribution of data points or to make a comparison of metric values across different indices in import, export, and employment as shown in the Figure 6.7.

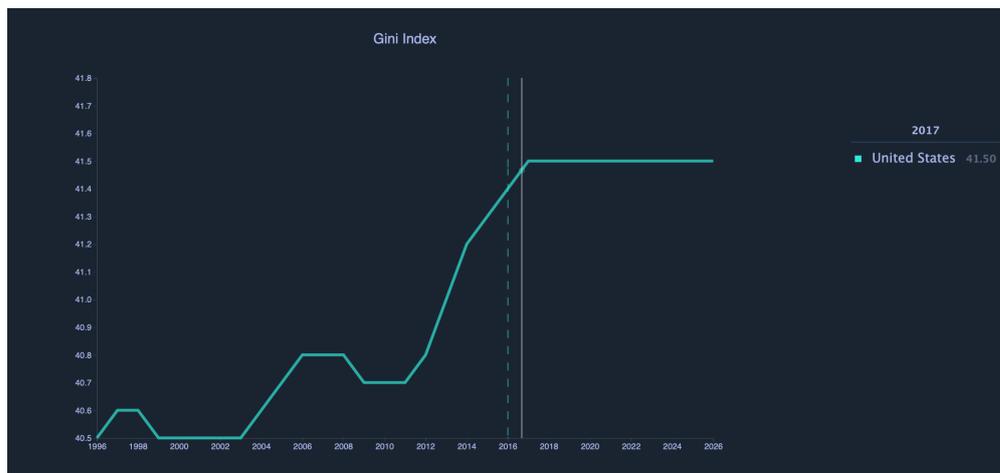


Figure 6.6: Line chart

The economy state is an exception to the other charts as it has 4 different states representing the economic phases of a country from recession to economic boom. A different type of chart was needed that would allow for easier comparison of values. A stacked normalized horizontal bar chart was used as shown in Figure 6.8, created using Vanilla JavaScript.

To render the chart, D3.js can use SVG or Canvas. The SVG rendering is chosen

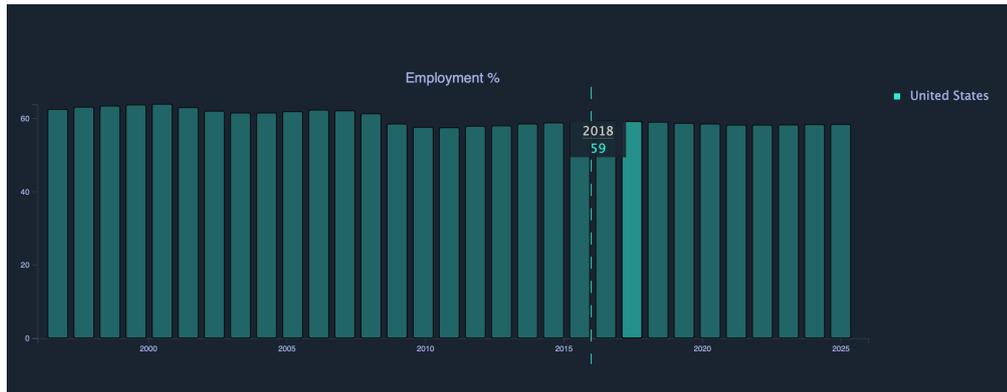


Figure 6.7: Bar chart



Figure 6.8: Economy State Chart

because it is fast enough and makes interaction and styling easy. The line and bar charts created consist of two axes, X and Y. The axis is one of the reference lines of a coordinate system, with each individual point representing a value in the coordinate system and also a position on the canvas on which the chart is displayed. In this case, the x-axis represents time in all line and bar charts. The x-axis was projected with the `d3.scaleTime()` function. The y-axis, which represents the range of values for a given indicator, was created using the `d3.scaleLinear()` function. To render the X and Y axes, new SVG groups were created to contain each axis. Line charts have been created using the `d3.line()` helper function, which allows lines to be drawn efficiently by converting the data into an SVG path. The bar charts were created using the function `d3.scaleBand()`, which divides the domain into n bands, where n is the number of values in the domain

array. Domain is the complete set of values that a chart can use, in our case all values between 1996 and 2026, representing the time frame for the data.

To get more information, a user can click on a chart to get details on demand overview as shown in Figure 6.6 and 6.7. Then by hovering over a year, there is an information about the value of an economic factor. A dashed line is also shown to provide a clear demarcation between real and forecast data. Another useful feature is that when hovering over a particular year in a main dashboard where all charts are displayed, that year is highlighted in each chart, making it easier for the user to see what is happening with each economic factor in that year as shown in Figure 6.10.

All line charts, bar charts, and the economy state chart are combined in the main user dashboard, as shown in Figure 6.9.

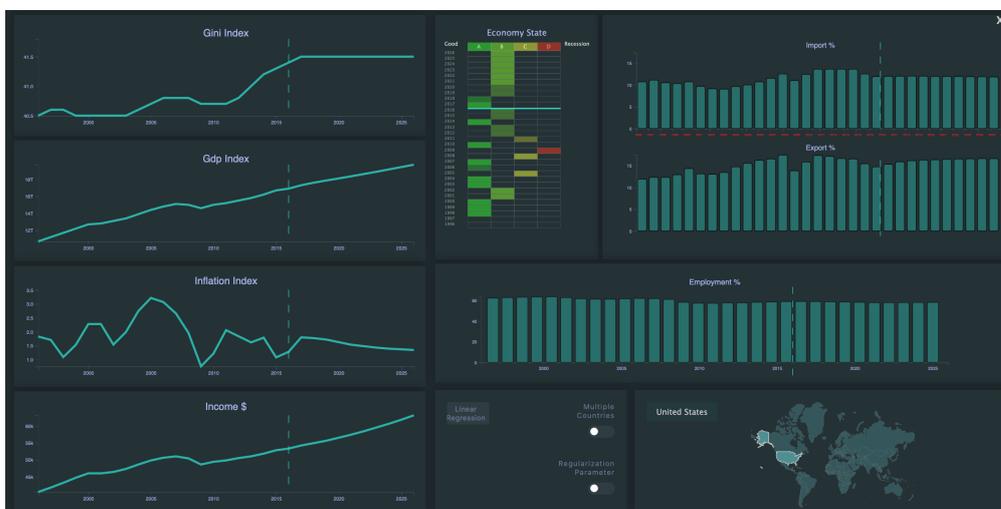


Figure 6.9: Dashboard with all charts displayed

In order to get insights into the trend comparison between the countries, multiple countries selection has been added. Since information about a single country does not provide a comprehensive picture of the actual trend in the world, it is useful to compare a few countries to gain insight into the actual trend. Given the performance of line and bar charts, and to prevent the charts from becoming too cluttered, the user can select up to three countries for comparison in the main dashboard. For this purpose, there is a “Multiple Countries” button that allows the user to select three countries and see all charts on a dashboard as shown in Figure 6.11. All the features already explained are also available in this overview.

In order to be able to compare the results obtained with the model, it is decided to include the data sets of the simple linear regression and all the data sets with regularization parameters between 0 and 1 with 0,1 steps. In this way, the user can see the difference between the selected model, the model of simple linear regression and the models with different regularization parameters. This feature is available for the Gini index chart,

6. DESIGN OF VISUAL ANALYTICS PROCESS

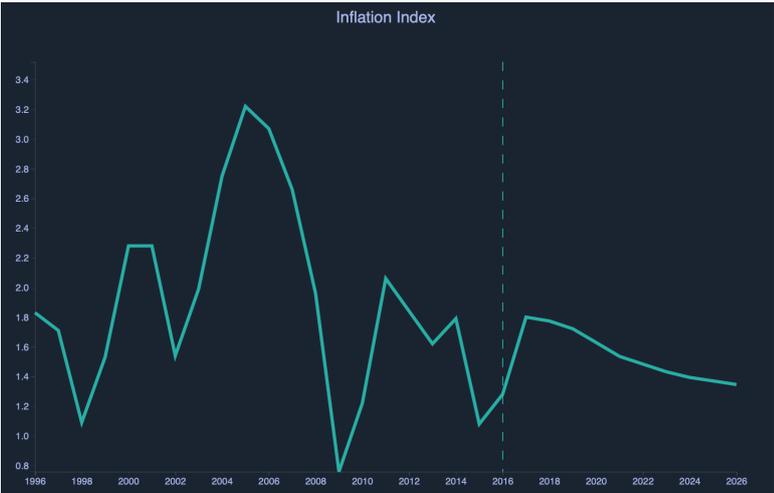


Figure 6.10: The main dashboard with all charts highlighted

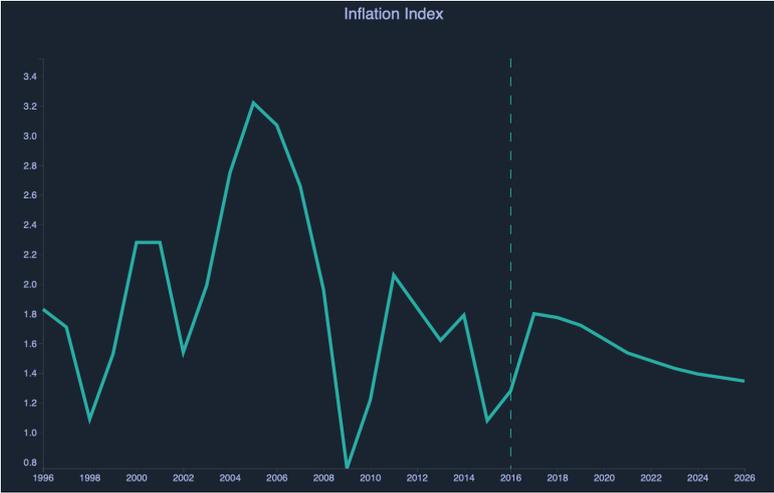


Figure 6.11: Multiple countries dashboard

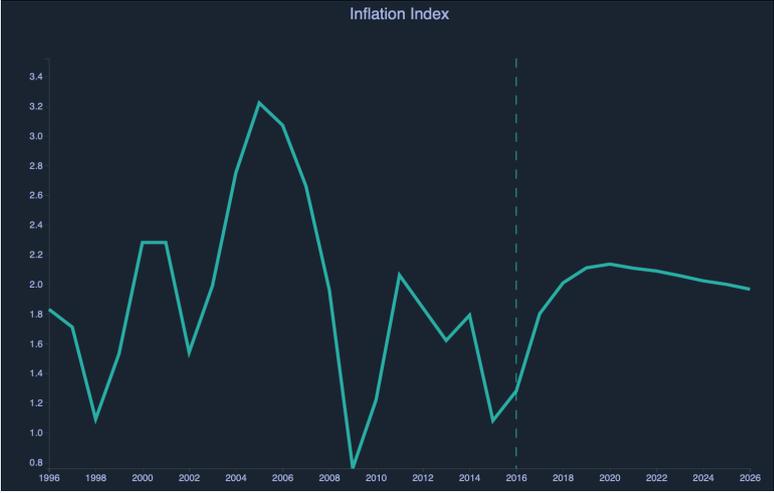
the GDP index chart, the inflation index chart and the income chart, as these recession factors have proven to be quite accurate in predicting the recession. Figure 6.12 shows an example.



(a) Inflation Index chart



(b) Inflation Index chart using Linear Regression Data



(c) Inflation index chart at a regularisation coefficient of 0.9.

Figure 6.12

Evaluation

This chapter presents the evaluation of the thesis based on the requirements formulated in the form of user stories. The application supports the workflow of spatial time series analysis and guides the user through the process with interactive visualizations.

7.1 Use Case Scenarios of The Recession Explorer Application

Based on the user stories, a possible use case scenarios are presented here. The process follows the user stories formulated in the Chapter 6.

Select the country on an interactive world map

The first use case scenario deals with the task of selecting a country of interest on an interactive world map. The main graphical user interface dashboard consists of an interactive world map where the user can select and click on a country of interest. The user can also hover over the map to find a country, as hovering over a country displays the name of the country, as shown in Figure 7.1. On the right side are the + and - buttons to zoom in and out.

Examining a year of interest

The second use case scenario deals with the task of examining a year of interest. Clicking on a selected country will open a new dashboard with all the charts as shown in Figure 7.2. Here the user has the ability to examine what is happening with each recession factor, what the predicted state of the economy will be, and also compare the charts with two other countries. When examining the state of a country, the user is shown seven different charts—Gini index chart, GDP index chart, inflation index chart, income chart, import

7. EVALUATION



Figure 7.1: Main dashboard—interactive world map with selected country of interest—Austria

and export chart, employment chart, and economy state chart. Hovering over a year in a dashboard highlights that year in each chart, making it easier for the user to see what is happening with each recession factor in that year.

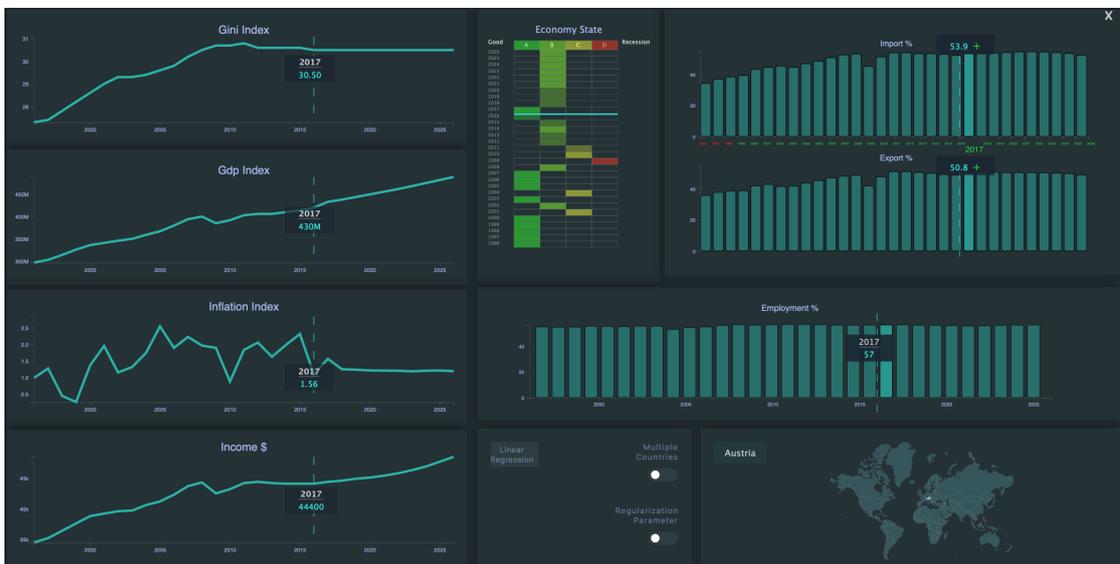


Figure 7.2: New dashboard with all charts displayed

Click on any chart

The third use case scenario deals with the task of clicking on any chart in a dashboard to get a detailed view. When the user clicks on a chart, only that chart is displayed in a detailed, zoomed-in view, as shown in Figure 7.3. When a user moves the mouse pointer over the line in the chart, they get information about the year and the Income value in that case. The dashed line provides a clear demarcation between real and forecast data. Clicking an X in the upper right corner will take the user back to the dashboard with all the charts.

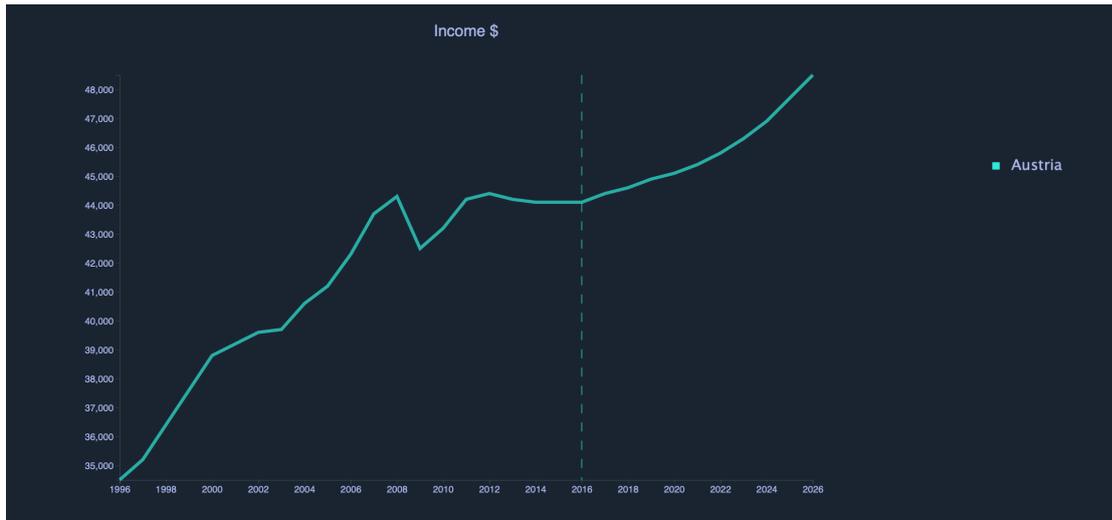


Figure 7.3: Detailed view of Income chart

A country's import and export activity can affect GDP, the exchange rate, the inflation rate, and the interest rate, which in turn can lead to changes in the business cycle. A detailed zoomed-in view of the import and export charts, shown in Figure 7.4, where the user can immediately see whether a country is importing or exporting more. If the difference between import and export is positive and the country is importing more, the year is shown in green colour, otherwise it is shown in red colour.

Select multiple countries

The fourth use case scenario deals with the task of selecting multiple countries. Since the important requirement of this application is to detect patterns and gain insights, the important feature is to compare the visualised data. Given the performance of line and bar charts, and to prevent the charts from becoming too cluttered, the user can select up to three countries for comparison in the main dashboard. The results can be seen in Figure 7.5. Once the first country is selected, the user can select two more countries in the smaller map in the lower right corner. Each country is assigned a different colour, which can be seen in a legend next to the smaller world map. Clicking on a country that

7. EVALUATION



Figure 7.4: Detailed view of Import and Export charts

is already selected on a smaller map will deselect the country and allow the user to select another country. The user can examine how the recession factors have changed over the years and what the predicted trend is. Due to the nature of the economy state chart, it is not included in the comparison. Through the features of economy state chart explained in Visual Analytics Task in Chapter 6, the grouped bar chart option is selected in this application, as shown in Figure 7.6.

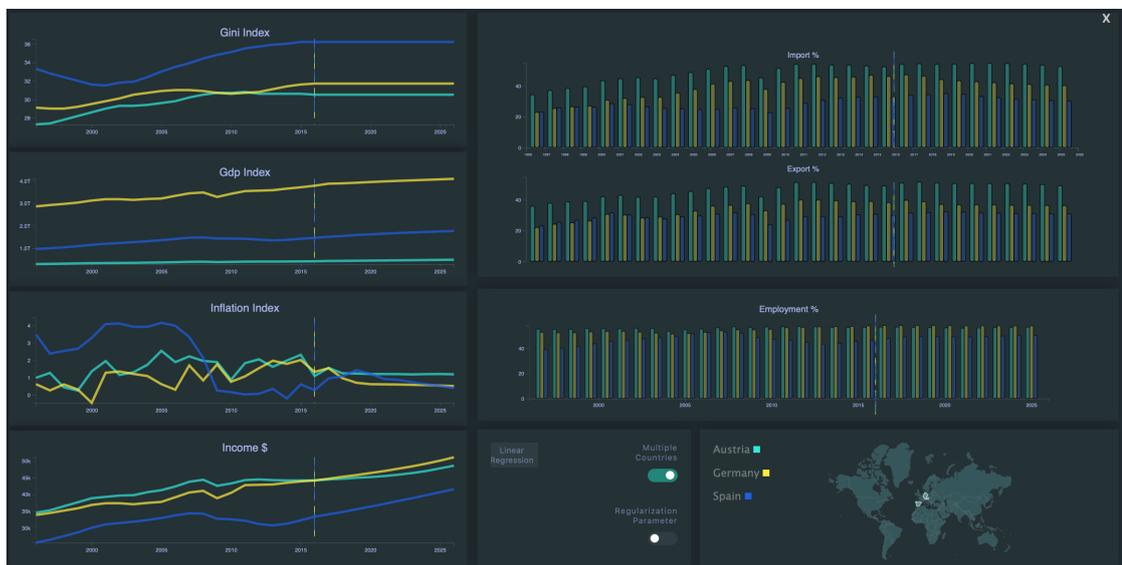


Figure 7.5: Multiple countries dashboard

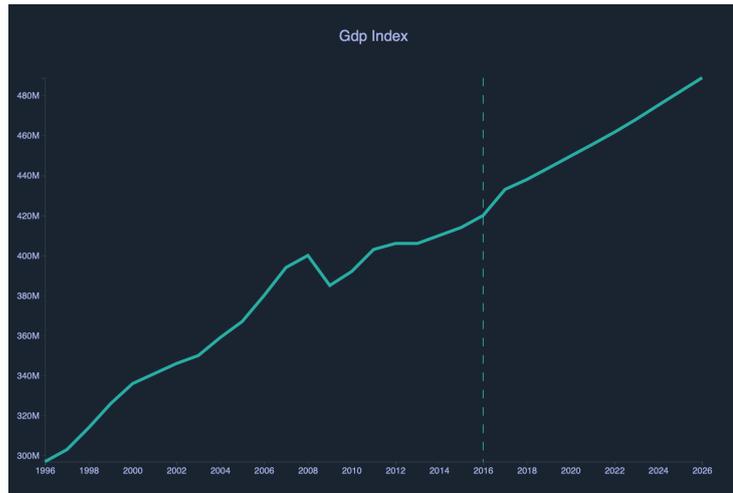


Figure 7.6: Grouped Bar Charts

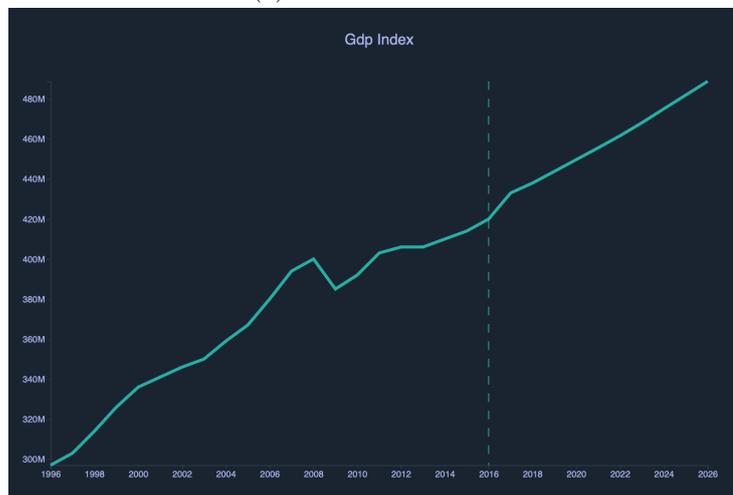
Visualise the data that other models have produced

The final use case scenario deals with the task of visualisation of the data that other models have produced. As explained in Chapter 6, the additional data were added and new charts were created to test whether the model is the best fit. To this end, there is the option of a slider and linear regression button at the bottom centre of the main dashboard. As explained earlier, only the Gini index chart, GDP index chart, inflation index chart, and income chart have this feature. The slider can be moved from 0 to 1 in 0.1 increments. This can be seen in Figure 7.7. However, this feature is not suitable for the casual user and is approaching the realm of the expert, as one must have knowledge of machine learning to understand what results are being displayed. This feature provides direction for future work, which will be discussed in more detail in Chapter 8.

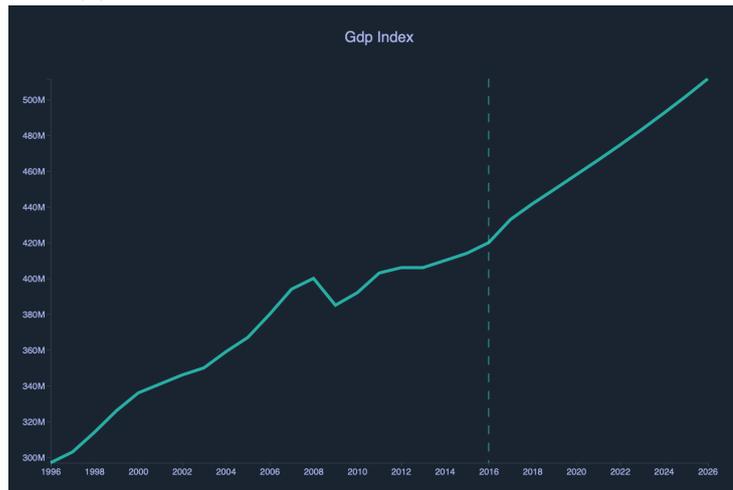
7. EVALUATION



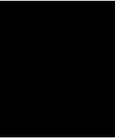
(a) GDP Index chart



(b) GDP Index chart using Linear Regression Data



(c) GDP index chart at a regularisation coefficient of 0.1



Conclusion

With the explosive growth of Big Data, visual analytics of spatial time series data has become increasingly important, and visualisation methods and techniques have also evolved. The exploratory analysis used in spatial time series analysis and statistical computing already requires a tight integration of human reasoning and automated methods. Examining the general visual analytics process and the more specific process for spatial time series data supports this observation. However, visualizing such data remains a major challenge. Given the complexity of the study area, this is very understandable, but significant efforts are needed to improve this situation. Combining spatial and temporal features of the data is often difficult, and usually only one feature is the focus. There are several solutions to visualize these two features, and some of them are presented in this thesis depending on the type of data. Nevertheless, there is no unique visualization method that has been proven to be effective for spatial time series data. Therefore, it is useful to combine different visualization techniques to support the analysis of spatial time series data.

The goal of this thesis was to implement the Recession Explorer application. The data used for the application comes from the field of economics, especially economic recession, as this topic is relevant or interesting to the vast majority of people. Based on literature review and findings, the solution implemented in this thesis proposes a holistic system for representing spatial time series data. The development of Recession Explorer application has been divided into two tasks. The first task is machine learning. The goal of this task was to predict recession factors and based on that predict future recessions. The second task is visual analysis. The goal of this task was to efficiently display the data from the machine learning models. Along with the development of the process design, the requirements for the application were also formulated. These requirements formed the basis for a customised use case scenario. Using use case scenarios, it is shown how the application is used and how the user is supported in solving his tasks. The proposed

application, together with the formulated use case scenarios, represents a possible answer to the research question.

8.1 Discussion and Limitations

The Recession Explorer application is not intended for data mining experts. That is, the user lacks knowledge about machine learning models and the parameters behind the visualization. The current trend in visual analytics is to interact with the user through indirect user interface controls, such as sliders and buttons [70]. However, this assumes that the user has knowledge of the model and parameters. The common method to help the analyst understand the data and build knowledge is through direct manipulation/feedback of visualization objects. When the analyst changes the parameters through direct manipulations, the system recalculates the models and provides an updated visualization. The more complex the models become, the more cognitive effort the user has to put in to adjust them. The approach proposed in this thesis is not transparent and the user cannot adjust the model parameters and manipulate the visualized objects, since all the models are precomputed.

The implementation provided an example of a holistic system that efficiently displays spatial time series data and explores patterns and insights in the data. The use of JavaScript and d3.js for the implementation has proven beneficial, as there is a wealth of examples and documentation. On the other hand, the traditional way of writing d3.js code is different from other common programming languages such as the usual JavaScript. Therefore, it takes a while to get used to it and understand what is happening. As a result, reproducibility and repeatability suffers because if a developer inexperienced in d3.js is to rebuild the implementation, it might be difficult to understand the written code. The question is whether another programming language would not be more suitable. Despite the drawbacks, d3.js was chosen because it is very flexible and fast, supports large datasets and dynamic behavior for interaction and animation, allowing the creation and customization of simple and complex data visualizations [66]. Another advantage is that d3.js is extremely powerful when it comes to handling geometric information and uses GeoJSON to represent geographic features. GeoJSON is a format for encoding a variety of geographic data structures and can represent a collection of features, which was very useful in the implementation.

Chapter 6 already briefly touched on the fact that future recessions are not easy to predict, even if the classification model is accurate. In this case, however, a very special situation has occurred, the pandemic caused by the virus COVID-19. COVID-19 has disrupted daily life worldwide and is the first epidemic event since the Spanish flu pandemic of 1918 to require an urgent global response [71]. Economies worldwide have slowed in all sectors and more than one-third of the world's population has been restricted by various lockdown measures. The consequences are still unclear, as is the time it will take to recover. Moreover, because the data used for the forecast is older than 2019, the model's predictions are not accurate.

Many machine learning models have been developed to predict the recession, but human inputs are always required. Some of the models are mentioned in Chapter 2. However, there is no single machine learning algorithm for predicting the recession. As with any other machine learning problem, it depends heavily on the characteristics of the data. In this respect, the prediction accuracy could be better by using other algorithms.

8.2 Future Work

The current version of the application is not yet fully developed. As mentioned in the previous section, the implementation of the visual analytics process is not transparent. So the next step would be to make it transparent by allowing users to modify the model parameters through direct manipulation. Of course, the use of the application would then be extended to the domain users. However, the technique of direct manipulation of time series data has not been sufficiently explored.

Another feature would be to extend the interface to help analysts create metrics for data analysis. As described in Chapter 3, one of the biggest challenges in spatial time series analysis is the influence of collocated neighboring spatial time objects on each other. To overcome this challenge, more analysis methods would need to be included in the application. For example, users could be given the option to see which other countries have similar recession factors in a given year, or which countries have the same projected economic state in a given year.

Another direction of future work could be to create a visual analytics platform for different types of data. In this day and age where there is an abundance of data, it is of great importance to create a visual analytics platform that also supports other types of data, such as scientific data, because many researchers in different fields could benefit from such a platform.

List of Figures

3.1	Data Events	11
3.2	Event Map	12
3.3	Space-time Cube	13
3.4	Trajectory Data	13
3.5	Space Time Trajectory Cube	14
3.6	Stack Based Trajectories	14
3.7	Point Reference Data	15
3.8	Raster Data	16
3.9	Raster Data Time Stamps	16
3.10	Raster Map	17
3.11	Isopleth Map	17
3.12	Heat Map	18
4.1	Visual Analytics Areas of Expertise	20
4.2	Visual Analytics Process	21
6.1	Customised Visual Analytics Process	29
6.2	GDP Distribution	33
6.3	Standardised GDP Distribution	33
6.4	Employment-GDP Classification	35
6.5	World Map	36
6.6	Line Chart	37
6.7	Bar Chart	38
6.8	Economy State Chart	38
6.9	Dashboard Containing All Charts	39
6.10	Highlighted Charts	40
6.11	Multiple Countries Dashboard	40
6.12	Inflation Index Chart – Different Datasets	41
7.1	World Map	44
7.2	Dashboard Containing All Charts	44
7.3	Details on Demand	45
7.4	Details on Demand-Import and Export	46
7.5	Multiple Countries Dashboard	46
		53

7.6	Grouped Bar Charts	47
7.7	GDP Index Chart – Different Datasets	48

Bibliography

- [1] Geoffrey Ellis Daniel Keim Jörn Kohlhammer and Florian Mansmann. *Solving Problems With Visual Analytics*. Eurographics Association, 2010.
- [2] Natalia Andrienko and Gennady Andrienko. “A visual analytics framework for spatio-temporal analysis and modelling”. In: *Data Mining and Knowledge Discovery* 27 (2012), pp. 55–83. DOI: 10.1007/s10618-012-0285-7.
- [3] *Gapminder*. URL: <https://www.gapminder.org>.
- [4] *Business Cycle Dating*. National Bureau of Economic Research, 2020. URL: <https://www.nber.org/research/business-cycle-dating>.
- [5] Wesley C. Mitchell Arthur F. Burns. *Measuring Business Cycles*. NBER, 1946.
- [6] Philip A. Klein Geoffrey H. Moore. *Analyzing Modern Business Cycles*. New York: Routledge, 1990.
- [7] Shiskin Julius. *Measuring Current Economic Fluctuations*. NBER, 1973.
- [8] J.P. Keeler. “Empirical Evidence on the Austrian Business Cycle Theory”. In: *The Review of Austrian Economics* 14 (2001), pp. 331–351. DOI: <https://doi.org/10.1023/A:1011937230775>.
- [9] B.P. Simpson. *Real Business Cycle Theory*. Palgrave Macmillan, New York, 2014.
- [10] José Granados. “Does Investment Call the Tune? Empirical Evidence and Endogenous Theories of the Business Cycle”. In: *Research in Political Economy* 28 (2012). DOI: 10.1108/S0161-7230(2013)0000028009.
- [11] Dev Nathan and Govind Kelkar. *Is COVID-19 really an exogenous shock?* United Nations University, 2020. URL: <https://www.wider.unu.edu/publication/covid-19-really-exogenous-shock>.
- [12] Kimberly Amadeo. *Causes of the Business Cycle*. 2020. URL: <https://www.thebalance.com/causes-of-the-business-cycle-3305804>.
- [13] *INVESTMENT BUSINESS CYCLES*. AmosWEB Encyclonomic WEB*pedia, 2020. URL: <http://www.AmosWEB.com>.
- [14] *MACROECONOMIC POLICY: MEANING, TYPES, HOW IT WORKS*. Penpoin, 2021. URL: <https://penpoin.com/macroeconomic-policy/>.

- [15] Sher Verick and Iyanatul Islam. *The Great Recession of 2008-2009: Causes, Consequences and Policy Responses*. 2010. URL: <http://ftp.iza.org/dp4934.pdf>.
- [16] Gary D. Hansen and Edward C. Prescott. “Did Technology Shocks Cause the 1990-1991 Recession?” In: *American Economic Review* 83 (1993), pp. 280–286. DOI: 10.4324/9780203070710.ch26.
- [17] Daniel Kurt. *The Financial Effects of a Natural Disaster*. 2020. URL: <https://www.investopedia.com/financial-edge/0311/the-financial-effects-of-a-natural-disaster.aspx>.
- [18] Mohamad Shaaf. “Predicting Recession Using the Yield Curve: An Artificial Intelligence and Econometric Comparison”. In: *Eastern Economic Journal* 26 (2000), pp. 171–190.
- [19] Jana Hvozdenka. “The Yield Curve as a Predictor of Gross Domestic Product Growth in Nordic Countries”. In: *Procedia Economics and Finance* 26 (2015), pp. 438–445. DOI: [https://doi.org/10.1016/S2212-5671\(15\)00871-0](https://doi.org/10.1016/S2212-5671(15)00871-0).
- [20] Andrew J. Filardo. “How Reliable Are Recession Prediction Models?” In: *Economic Review, Federal Reserve Bank of Kansas City* 84 (1999), pp. 35–55.
- [21] John Galakis Spyridon D.Vrontos and Ioannis D.Vrontos. “Modeling and predicting U.S. recessions using machine learning techniques”. In: *International Journal of Forecasting* 37 (2021), pp. 647–671. DOI: <https://doi.org/10.1016/j.ijforecast.2020.08.005>.
- [22] Yaser S. Abu-Mostafa Alexander James and Xiao Qiao. “Machine Learning for Recession Prediction and Dynamic Asset Allocation”. In: *The Journal of Financial Data Science* 1 (2019), pp. 41–56. DOI: <https://doi.org/10.3905/jfds.2019.1.007>.
- [23] *Spatiotemporal*. URL: <https://www.omnisci.com/technical-glossary/spatial-temporal>.
- [24] Yiannis Kamarianakis and Poulicos Prastacos. “Spatial Time-Series Modeling: A review of the proposed methodologies”. In: *Proceedings 2005 - The 8th AGILE International Conference on Geographic Information Science* (2005).
- [25] A.D. Cliff and J.K. Ord. “Space-Time Modelling with an Application to Regional Forecasting”. In: *Transactions of the Institute of British Geographers* (1975), pp. 119–128. DOI: <https://doi.org/10.2307/621469>.
- [26] Anuj Karpatne Pang-Ning Tan Michael Steinbach and Vipin Kumar. *Introduction to Data Mining, 2nd Edition*. London: Pearson, 2018.
- [27] Anuj Karpatne Gowtham Atluri and Vipin Kumar. “Spatio-Temporal Data Mining: A Survey of Problems and Methods”. In: *ACM Computing Surveys* 51 (2018), pp. 1–41. DOI: <https://doi.org/10.1145/3161602>.
- [28] Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, S. Klooster, and A. Torregrosa. “Data mining for the discovery of ocean climate indices”. In: *In Scientific Data Mining* (2002).

- [29] Catie Chang Xiao Liu and Jeff H. Duyn. “Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns”. In: *Front Syst Neurosci.* (2013). DOI: <https://doi.org/10.3389/fnsys.2013.00101>.
- [30] Anthony C. Gatrell, Trevor C. Bailey, Peter J. Diggle, and Barry S. Rowlingson. “Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology”. In: *Transactions of the Institute of British Geographers* 21 (1996), pp. 256–274. DOI: <https://doi.org/10.2307/622936>.
- [31] Mattias Persson. *A Survey of Methods for Visualizing Spatio-temporal Data*. June 2020.
- [32] San Francisco Police. *Crime maps*. URL: <https://www.sanfranciscopolice.org/stay-safe/crime-data-and-maps/crime-maps>.
- [33] P. Gatalsky, N. Andrienko, and Gennady Andrienko. “Interactive analysis of event data using space-time cube”. In: *Proceedings of 8th International Conference on Information Visualization* (2004), pp. 145–152. DOI: 10.1109/IV.2004.1320137.
- [34] Damien R. Farine, Ariana Strandburg-Peshkin, Tanya Berger-Wolf, Brian Ziebart, Ivan Brugere, Jia Li, and Margaret C. Crofoot. “Both nearest neighbours and long-term affiliates predict individual locations during collective movement in wild baboons”. In: *Scientific Reports* 6 (2016). DOI: <https://doi.org/10.1038/srep27704>.
- [35] Merve Keskin, Bahadır Çelik, Ahmet Dogru, and Mete Pakdil. “A Comparison of Space-Time 2D and 3D Geovisualization”. In: *27th International Cartographic Conference* (2014).
- [36] Christian Tominski, H. Schumann, Gennady Andrienko, and Natalia Andrienko. “Stacking-Based Visualization of Trajectory Attribute Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 18 (2012), pp. 2565–2574. DOI: 10.1109/TVCG.2012.265.
- [37] *Global Climate Monitor*. URL: <https://www.globalclimatemonitor.org>.
- [38] National Geography Information Institute (NGII). *The National Atlas of Korea*. 2017. URL: http://nationalatlas.ngii.go.kr/pages/page_1260.php.
- [39] *What is...Heat map*. 2019. URL: <https://www.omniconvert.com/what-is/heat-map/>.
- [40] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [41] Gennady Andrienko, Natalia Andrienko, Peter Bak, Stefan Wrobel, and Daniel A. Keim. *Visual Analytics of Movement*. Springer Science and Business Media, 2013.
- [42] Jie Li, Siming Chen, Kang Zhang, Gennady Andrienko, and Natalia Andrienko. “COPE: Interactive Exploration of Co-occurrence Patterns in Spatial Time Series”. In: *IEEE Transactions on Visualization and Computer Graphics* 25 (2018), pp. 2554–2567. DOI: 10.1109/TVCG.2018.2851227.

- [43] Hartmut Ziegler, Marco Jenny, Tino Gruse, and Daniel A. Keim. “Visual market sector analysis for financial time series data”. In: *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), pp. 83–90. DOI: 10.1109/VAST.2010.5652530.
- [44] Diansheng Guo. “Multivariate spatial clustering and geovisualization”. In: *In book: Geographic Data Mining and Knowledge Discovery, Second Edition* (2009), pp. 325–345. DOI: 10.1201/9781420073980.ch12.
- [45] G. Andrienko, N. Andrienko, Sebastian Bremm, Tatiana von Landesberger, Tobias Schreck, Peter Bak, and Daniel Keim. “Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns”. In: *Computer Graphics Forum* 29 (2010), pp. 913–922. DOI: <https://doi.org/10.1111/j.1467-8659.2009.01664.x>.
- [46] Natalia Andrienko, Gennady Andrienko, Elena Camossi, Christophe Claramunt, Jose Cordero Garcia, Georg Fuchs, Melita Hadzagic, Anne-Laure Joussemme, Cyril Ray, David Scarlatti, and George Vouros. “Visual exploration of movement and event data with interactive time masks”. In: *Visual Informatics* 1 (2017). DOI: 10.1016/j.visinf.2017.01.004.
- [47] Natalia Andrienko and Gennady Andrienko. “Spatio-temporal visual analytics: a vision for 2020s”. In: *JOURNAL OF SPATIAL INFORMATION SCIENCE* 20 (2020), pp. 87–95. DOI: 10.5311/JOSIS.2020.20.661.
- [48] Wei Luo and Alan M. MacEachren. “Geo-social visual analytics”. In: *JOURNAL OF SPATIAL INFORMATION SCIENCE* 8 (2014), pp. 27–66. DOI: 10.5311/JOSIS.2014.8.139.
- [49] JONATHAN NELSON, STERLING QUINN, BRIAN SWEDBERG, WANGHUAN CHU, and ALAN MACEACHREN. “Geovisual analytics approach to exploring public political discourse on twitter”. In: *ISPRS International Journal of Geo-Information* 4 1 (2015), pp. 337–366. DOI: <https://doi.org/10.3390/ijgi4010337>.
- [50] Alexander Savelyev Scott Pezanowski Alan M MacEachren and Anthony C Robinson. “SensePlace3: a geovisual framework to analyze place–time–attribute information in social media, Cartography and Geographic Information Science”. In: *Cartography and Geographic Information Science* 45 (2018), pp. 420–437. DOI: 10.1080/15230406.2017.1370391.
- [51] Alexander Savelyev and Alan M. MacEachren. “Augmenting geovisual analytics of social media data with heterogeneous information network mining—Cognitive plausibility assessment”. In: *PLOS ONE* 12 (2018), pp. 01–27. DOI: <https://doi.org/10.1371/journal.pone.0206906>.
- [52] Ross Maciejewski, Ryan Hafen, Stephen Rudolph, Stephen G. Larew, Michael A. Mitchell, William S. Cleveland, and David S. Ebert. “Forecasting Hotspots—A Predictive Analytics Approach”. In: *IEEE Transactions on Visualization and Computer Graphics* 17 (2011), pp. 440–453. DOI: 10.1109/TVCG.2010.82.

- [53] Natalia Andrienko, Gennady Andrienko, and Salvatore Rinzivillo. “Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics”. In: *Information Systems* 57 (2016), pp. 172–194. DOI: <https://doi.org/10.1016/j.is.2015.08.007>.
- [54] Abish Malik, Ross Maciejewski, Sherry Towers, Sean McCullough, and David S. Ebert. “Proactive Spatiotemporal Resource Allocation and Predictive Visual Analytics for Community Policing and Law Enforcement”. In: *IEEE Transactions on Visualization and Computer Graphics* 20 (2014), pp. 1863–1872. DOI: [10.1109/TVCG.2014.2346926](https://doi.org/10.1109/TVCG.2014.2346926).
- [55] Jaegul Choo and Shixia Liu. “Visual Analytics for Explainable Deep Learning”. In: *IEEE Computer Graphics and Applications* (2018). DOI: [10.1109/MCG.2018.042731661](https://doi.org/10.1109/MCG.2018.042731661).
- [56] Anirudh V K. *What Is Machine Learning: Definition, Types, Applications And Examples*. 2019. URL: <https://www.toolbox.com/tech/artificial-intelligence/tech-101/what-is-machine-learning-definition-types-applications-and-examples/>.
- [57] Kurtis Pykes. *The Difference Between Classification and Regression in Machine Learning*. 2021. URL: <https://towardsdatascience.com/the-difference-between-classification-and-regression-in-machine-learning-4ccdb5b18fd3>.
- [58] Claude Sammut and Geoffrey I. Webb. *Encyclopedia of Machine Learning and Data Mining*. Boston: Springer Science and Business Media, 2017.
- [59] Wolfgang Aigner, Silvia Miksch, H. Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. London: Springer, 2011.
- [60] Mike Cohn. *User Stories Applied: For Agile Software Development*. Addison–Wesley, 2010.
- [61] *Python*. URL: <https://www.python.org>.
- [62] Jakub Protasiewicz. *Python AI: Why Is Python So Good for Machine Learning?* 2018. URL: <https://www.netguru.com/blog/python-machine-learning>.
- [63] *scikit-learn*. URL: <https://scikit-learn.org/stable/>.
- [64] *Pandas*. URL: <https://pandas.pydata.org>.
- [65] *JavaScript*. URL: <https://www.javascript.com>.
- [66] *D3 Data-Driven Documents*. URL: <https://d3js.org>.
- [67] Jin-Kyu Jung, Manasa Patnam, and Anna Ter-Martirosyan. “An Algorithmic Crystal Ball: Forecasts-based on Machine Learning”. In: *IMF Working Paper No. 18/230* (2018).
- [68] *Line Graph*. 2020. URL: <https://www.betterevaluation.org/en/evaluation-options/LineGraph>.

- [69] Shashank Sahay. *Your guide to Data Visualization for Comparison*. 2018. URL: <https://medium.muz.li/guide-to-data-visualization-comparison-part-1-678382ceef00>.
- [70] Rei Takami and Yasufumi Takama. “Visual Analytics Interface for Time Series Data Based on Trajectory Manipulation”. In: (2018), pp. 342–347. DOI: 10.1109/WI.2018.00–70.
- [71] Grace T. Patterson, K. Marie McIntyre, Helen E. Clough, and Jonathan Rushton. “Societal Impacts of Pandemics: Comparing COVID-19 With History to Focus Our Response”. In: (2021). DOI: <https://doi.org/10.3389/fpubh.2021.630449>.