MASTER

Visualizing geographic information in event logs

Mangnus, J.W.A.

Award date:
2018

Link to publication
Visualizing Geographic Information in Event Logs

Author
J.W.A. (Jos) Mangnus

Supervisors
Dr. D. (Dirk) Fahland
Dr.ir. R.J. (Roeland) Scheepens
Dr.ir. H.M.M. (Huub) van de Wetering

March 16, 2018
Abstract

Process models obtained through process mining describe how cases flow through a process, but only show an abstract view on reality. When handling geographic information, this abstract view obstructs analysis of geographic information. By offering a realistic geographic view, this analysis becomes easier. In this thesis, we examine the ways in which we can assist an analyst in selecting an appropriate visualization technique to facilitate this realistic geographic view. By matching past work in spatio-temporal visualization with analysis problems in process mining data, we propose a framework to allow process analysts to make a sensible choice of visualization technique for their analysis problems. Furthermore, we provide an example of the application of the framework through a business case with real-world data. This application shows the framework’s usefulness, but also highlights one way in which the framework can be adapted.
Acknowledgements

Before starting with the scientific considerations in this thesis, I’d like to take some time to thank the people that helped me reach this goal.

First of all I’d like to thank Roeland Scheepens, Dirk Fahland and Huub van de Wetering for their guidance, wisdom, enthusiasm and patience in their supervision of this project. Without their advice, I would never be able to come to the product below and their continuous valuable input is still greatly appreciated.

I’d also like to thank my colleagues at ProcessGold for providing a warm and welcoming environment to perform this research. Both the noisy and the calm side of the office were great places to work on this research, although some types of work proved to be better suited for the latter. I hope to share many a sporrel with these people in the future.

I would also like to extend a lot of gratitude to the many new friends I encountered during my studies. In particular, the people I met during the many wonderful activities, borrels and lunches at GEWIS. I’d like to thank my fellow 32nd board members, for putting up with me in one of the most intense and educational years of my study, and the members of the BArCommittee, old and new, for sharing many a beer with me and for giving me the opportunity to share beers with many others. I’d also like to thank Luc, Per, Marijn, Rick and Tos for pulling me through the first years of my bachelor and my fellow Data Science in Engineering pioneers, Puck, Sako and Stefan, for the many master courses and projects together.

I would also like to extend a special word of thanks towards EditieNL for inspiring me to live up to the title of ”One of the Best Mathematicians of the TU/e”[8]. Although I’m not 100% sure I’ve fully lived up to the title, it inspired me to become the data scientist I am today.

In closing, I’d like to thank my family, Toine, Jacintha, Eveline and Marjolein, for their continued support during my long study period. I always looked forward to the moments of rest my visits back to Biervliet provided and greatly appreciate the support and, sometimes, reasonable doubt you had for some of the choices I made during my studies.
## Contents

1 Introduction .................................................. 7

2 Previous work .................................................. 10
   2.1 Process Mining ............................................. 10
   2.2 Time in Visualization ..................................... 12
      2.2.1 Time as a second dimension ......................... 12
      2.2.2 Time as a third dimension ......................... 13
      2.2.3 Abstract Representation of Time ................... 15
   2.3 Spatio-temporal Visualizations ............................ 16
      2.3.1 Origin-Destination Mapping ......................... 16
      2.3.2 Origin-Destination Flow Maps ....................... 17
      2.3.3 Maptrix .............................................. 17
      2.3.4 Flowstrates ......................................... 17
      2.3.5 Flow trees [28] ..................................... 17
      2.3.6 Space-Time Paths .................................... 17
      2.3.7 Storygraphs .......................................... 17
      2.3.8 Data Dials ........................................... 17
   2.4 Conclusion ................................................ 17

3 Determining Geographic Targets in Event Data ................. 18
   3.1 Geographic Information in Data: A Structured Literature Review ..................................... 18
      3.1.1 Scoping ............................................... 18
      3.1.2 Planning ............................................. 18
      3.1.3 Searching ........................................... 19
      3.1.4 Screening ........................................... 19
      3.1.5 Eligibility .......................................... 20
   3.2 Use cases for Process Mining with Geographic Data .......................................................... 22
      3.2.1 Do you know cases in which geographic context would have helped the user understand their process better? ........................................ 23
      3.2.2 What kind of data was used in the process? .............................................................. 23
      3.2.3 What kind of information was the user looking to retrieve from their data? ............................. 23
      3.2.4 What kind of process did the data come from? .......................................................... 24
      3.2.5 How was the geographic context represented in the data? ....................................... 24
      3.2.6 How did the user retrieve that information using the current tools? .............................. 24
      3.2.7 Where would you think value could be added using visualization of the geographic context with the current data? ........................................ 25
      3.2.8 Are there any cases you can imagine in which a visualized geographic context would help the user? .............................................................. 25
      3.2.9 If so, what kind of process would the data come from? ............................................. 25
      3.2.10 What would the goal of the user be and what kind of information should the user receive to complete that goal? ........................................ 25
   3.3 Conclusions ................................................. 26
      3.3.1 Identifying Primitives .................................. 27
4 Connecting Visualization Features and Geographic Targets

4.1 Exploring Targets in Literature

4.1.1 Event Space-Correlation Analysis Algorithm Based on Ant Colony Optimization

4.1.2 Intimate partner homicide-suicide: The role of media in depicting life-ending events, along with an analysis of the prevalence and geographic distribution of these events

4.1.3 Online temporal-spatial analysis for detection of critical events in cyber-physical systems

4.1.4 Scalable cluster analysis of spatial events

4.1.5 The geography of taste: Analyzing cell-phone mobility and social events

4.2 Further Targets

4.2.1 Geographic Step ⇒ Process Step

4.2.2 External Stakeholders

4.3 Framework for Geographic Targets in Event Data

4.3.1 Primitives and Visualization Techniques

4.3.2 Comparison of Targets

4.3.3 Suggested Framework

4.4 Conclusion

5 Applying the Framework

5.1 Use Case Description

5.1.1 Data Description

5.1.2 Use Case Questions

5.1.3 Process aspect of the data

5.2 Requirements of the Visualization

5.2.1 Solution Requirements

5.3 Visualization Techniques and their Suitability

5.3.1 Applying the Framework

5.3.2 Usability of Techniques

5.3.3 Common Limitations of Discussed Techniques

5.3.4 Choosing the Visualization

5.4 Extending the Use Case

5.5 Conclusion

6 Implementation of the Prototype

6.1 Implementation Details

6.2 Design Choices

6.3 Creating a new Dashboard

6.4 Conclusion

7 Evaluation of the Prototype

7.1 Analysis of the Business Questions

7.1.1 Q1: Finding a Service Hub

7.1.2 Q2: Spotting Special Events

7.1.3 Q3: Spotting Large Net In- or Outflow

7.1.4 Q4: Identifying Daily Trends

7.2 Conclusion
7.2 User Tests ................................................. 74
  7.2.1 Structure of the User Tests .............................. 74
  7.2.2 User Analyses .......................................... 75
  7.2.3 User Evaluation ......................................... 76
7.3 Conclusions ............................................... 77

8 Conclusion ................................................. 79
  8.1 Future Work ............................................... 79

A Accepted Papers ........................................... 84

B User Test Notes ............................................ 89

C Use Case Questions ......................................... 96
  C.1 1. Which city performs the most service cases from other cities and can we find reasons for the location of this city? 96
  C.2 2. Can we detect special events happening in the usage of the Call A Bike system? 96
  C.3 3. Are there areas which have large net in- or outflow and can we detect trends in these spots? 96
  C.4 4. Can we find patterns in daily usage of bikes? What could be the behavior that causes these patterns? 96
  C.5 5. Are there factors that influence when a bike needs to be serviced? 97
  C.6 6. What does the life cycle of a bike look like? 97
  C.7 7. What influences how long a service reservation takes? 97
  C.8 8. Can we spot patterns in the appearance of reservations which end up in no-shows? 97
1 Introduction

Many businesses in today’s world rely on information systems to keep track of the execution details of their regular processes. Insights from the data in these systems are hard to achieve, however. Much research is being done on process mining, a set of techniques which allow an analyst to extract and test process models with data gained from the information system itself. These process models describe how cases move through the process and allow a view into the structure of the process that is modeled. Visualizations of process models offer an abstract view on reality that only takes into account the flow from one abstract task to the next. In the handling of specific parts of processes, there may be better ways to allow for analysis than regular process models obtained by performing process mining techniques. More specifically, the analysis of processes with a strong geographic aspect is hard to perform using regular process models.

In current analyses, a lot of geographic information is not used at all, while this information could prove invaluable. Suppose someone is analyzing the process of a postal delivery company. The generic process of a package being delivered in steps from a certain origin through distribution centers to the eventual destination can be shown using process models. However, with standard process models, any geographic context of the delivery is lost. It could be that packages are often routed to a distribution center far away from the eventual destination needlessly, but without exact knowledge of the location of the distribution centers, this information is inaccessible. Providing access to the information within the geographic data might allow for easier analysis on a myriad of cases. In the previously mentioned case of a postal delivery process, a visual insight into the geographic context of the process might reveal these kinds of detours.

Geographic data is usually visualized using geographic visualization techniques. The field of geographic visualization focuses on visualizing information from geographic data in the broadest term. This can be done through the use of maps that provide a spatial context to the data, which allows an analyst easier access to important insights from the data. Due to the nature of business process data, we cannot restrict our scope to only geographic visualization techniques. Time is a very important aspect of the analysis of business processes, so we focus on visualization techniques which support spatio-temporal data.

Efforts to combine process mining and geographic visualization are thus far fairly limited. De Leoni et al. [18] use a map as background to replay the occurrence of activities in certain areas, while Zhu et al. [32] use a map and a process model to show an animation of cases moving through the process and through geography. These efforts are, however, fairly limited in their scope and are mostly focused on specific use cases.

There is no known generic approach to visualize geographic data for process mining purposes, while it may prove to be extremely useful in the analysis of business process data. A visualized geographic context offers a view on data that does not inherently rely on a users familiarity with the region related to the data. Visualizing it as part of the process mining workflow could add a lot of value in the analysis of data with a geographic aspect.

In this thesis, we examine the possibility of developing a framework to handle the different kinds of geographic information present in business process data. By looking at the different aspects of geographic information in event logs, assessing the usability of visualization techniques with these aspects and the development and evaluation of this framework, we strive to answer the following research question:
How can we support an analyst in selecting an appropriate spatio-temporal visualization technique for analyzing event data in a geographic context?

We aim to provide an answer to this question in multiple steps. In Chapter 2, we aim to answer the following question: How has spatio-temporal data and event data been visualized in the past? We answer this question by identifying previous work in the field of process mining, visualization of event data, and the visualization of data with both a spatial and a temporal factor in Chapter 2. Through this summary of previous work, we aim to be able to compare the possibilities of the different discussed visualization techniques.

In Chapter 3 we answer the following question: What kind of data can we expect to encounter in the analysis of geographic information in event data? We use a structured literature review and interviews with experts to determine this. From the literature study, we identify 5 different research domains which handle event data with a geographic context, analyze 69 articles based on their geographic and temporal data and to what degree the subject of the articles can be modeled as a process. From the interviews, we identify 8 potential cases in which the experts see that a visualized geographic context adds analytical value.

In Chapter 4, we merge the results of Chapter 2 and Chapter 3 and answer the following question: How can concepts from the visualization of spatio-temporal data and geographic information found in the analysis of events be combined in a useful manner? We do this by taking cases from the interviews and literature and matching them with the visualizations from Chapter 2. From these matchings, we develop a conceptual mapping and a framework for handling geographic information in event data. This framework relies on the grouping of geographic analysis of event data and advising visualization solutions for each group. A user of the framework should identify the group to which their analysis problem corresponds and can then discover a suitable solution.

In Chapter 5, we look for an answer to the following question: Does the application of the framework on a business case result in a useful suggestion? To answer this question, we use a business case with data from public transit bikes from the Deutsche Bahn. The usage and service of public transit bikes provide an interesting use case as a business process with a strong geographic component. Through the application of the business case on the framework and the analysis of this business case in terms of solution requirements, we can see whether the framework delivers a suitable solution. The application of the framework to this business case is discussed in Chapter 5 while an implementation of the result of this application is described in Chapter 6.

In Chapter 7, we answer the following question: Is the prototype developed in Chapter 6 useful for the analysis task at hand? We evaluate the developed prototype through an in-depth analysis of the data with the prototype and through user tests. With this evaluation, we not only test the usability of the prototype, but also the usefulness of the framework as a whole and provide an answer to the question how geographic visualization can help process analysts. The framework is shown to not yet be perfect, but is still able to provide a suitable solution for the analysis task within the business case.

In Chapter 8, we summarize our results, identify the shortcomings of the framework and describe the possibilities for future work. The largest amount of future work is within the application of the framework on new use cases, where each result of the framework can be tested and evaluated with real-life analysis cases.
This project was partially executed in the context of ProcessGold. ProcessGold is a company that provides a combined process mining and business intelligence platform aimed at allowing business to analyze their processes.
2 Previous work

In this chapter, a short introduction into process mining and an overview of previous work in the visualization of data with spatial and temporal aspects are provided. This overview is split up into two parts. The first part focuses on the possibilities of visualizing temporal data to get an idea of how the temporal facet of event data is handled in visualization. The final section focuses on visualization techniques specifically for spatio-temporal data and uses a framework to structurally analyze the provided examples of visualization methods.

2.1 Process Mining

Process Mining is a relatively new field of research which focuses on the gap between data mining and process science. Modern information systems within professional organizations contain a lot of information about the process that drives these organizations. Although this information is available in their systems in the form of logs, it is very hard to extract patterns from these logs manually. Process mining aims to use these logs to provide an organization with a useful insight into this structure. It does so by analyzing so-called events. An event in a process mining context is a record that a particular activity has occurred in a particular case at a specific point in time. An activity is a well-defined step in a process, while a case is an instance of the process. These events are registered in the information systems in event logs and process mining techniques use these event logs to perform various tasks.

An example of an event log can be seen in figure 1. Each record of an event log holds at least a case identifier, an activity name, and a timestamp. Other information that is relevant to the execution of an event can also be included. For instance, the resource that performed the event, the cost of the event, and the location of an event or process can be added to the event log.

There are three types of process mining: discovery, conformance and enhancement [1]. Process discovery techniques take an event log and create a process model based purely on the events within the log. Conformance techniques check to what degree a given process model and a given event log conform to each other and enhancement focuses on extending a known process model with information gained through an event log.

Process mining usually works with process models. These process models provide a visual guide on how a process functions. An example of a process model in the form of a Petri net can be found in figure 2.

<table>
<thead>
<tr>
<th>Case id</th>
<th>Event id</th>
<th>Time stamp</th>
<th>Activity</th>
<th>Resource</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35654473</td>
<td>30-12-2010:11:02</td>
<td>register request</td>
<td>Pete</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>35654474</td>
<td>31-12-2010:10:06</td>
<td>examine thoroughly</td>
<td>Sue</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>35654425</td>
<td>05-01-2011:15:12</td>
<td>check ticket</td>
<td>Mike</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>35654426</td>
<td>06-01-2011:11:18</td>
<td>decide</td>
<td>Sasa</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>35654427</td>
<td>07-01-2011:14:24</td>
<td>reject request</td>
<td>Pete</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>35654483</td>
<td>30-12-2010:19:32</td>
<td>register request</td>
<td>Mike</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>35654485</td>
<td>30-12-2010:12:12</td>
<td>check ticket</td>
<td>Mike</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>35654487</td>
<td>30-12-2010:14:16</td>
<td>examine casually</td>
<td>Pete</td>
<td>400</td>
</tr>
</tbody>
</table>

Figure 1: An example event log from a process for handling compensation requests. Image courtesy of van der Aalst [1].
The position of an activity relative to other activities tells the analyst something about the relative order. In figure 2, an activity to the right of another activity will most of the times takes place after the other activity. In the example in figure 2, ordering of the process is done from left to right, but another common alternative is from top to bottom. In commercial applications, process models tend to be visually enriched using two factors: color and size. An example of this kind of visual enrichment can be found in figure 3, which shows a demo process in the ProcessGold platform.

**Figure 2:** A Petri net modeling the handling of compensation requests. Image courtesy of van der Aalst[1].

**Figure 3:** An image generated through the use of the commercial ProcessGold[23] application. Both color and size are used to indicate the number of cases that flow through parts of the process.
In the example in figure 3, the color of an activity shows the number of cases that have an event associated with that activity and the color and size of a transition are used to indicate the number of cases have passed through that transition. The user can also choose to have color indicate other performance aspects of the process, such as throughput time.

2.2 Time in Visualization

To effectively visualize event data, showing the temporal aspect of events is crucial. In this section, to gain knowledge on how time is handled in visualization, we survey different ideas in how time is represented in contemporary visualization research. Aigner et al. [2] have created a survey on visualization technologies with this very goal in mind. The technologies mentioned in the work of Aigner et al. can be bundled in the sense of how they represent the temporal aspect of their data. In this section, a description and some examples are offered for these groups of technologies. Note that these technologies are not limited to time in event data, but apply to all temporal data.

2.2.1 Time as a second dimension

A very common and well-known manner of handling time data is by plotting a value in one dimension and the changes of this value along a second dimension. Many classic visualization techniques, such as line graphs and bar graphs, use the horizontal axis to display a range of time. An example of this can be found in figure 4.

Small multiples extend this idea by, instead of plotting a single value on the second dimension, showing partitions of the data. These partitions can be any kind of attribute which divides the data and thus can also be used to show change over time. One of the oldest examples of small multiples can be found in figure 5.

Although using a horizontal axis as a range of time is very common, there are other ways of using a position in two dimensions to indicate a value at a time. A technique that extends the idea of using the horizontal axis as a time range, is the Spiral display as shown in figure 6. The spiral display allows for the visualization of patterns over time and easy discovery of reoccurring events in time.

![Figure 4: Time is used as the horizontal axis in this bar chart. Image courtesy of Harris](image1)

![Figure 5: Horse in Motion by Eadweard Muybridge. The movement of a horse can be seen in this example of small multiples.](image2)
Figure 6: Spiral displays show time as a horizontal axis, but also curve the whole chart to allow for the discovery of patterns in time. In this example, temperature is plotted over time to enable the discovery of patterns. The colors in each row correspond to one unique temperature based on Two-Tone pseudo coloring. Image courtesy of Tominski 	extit{et al.} [27]

2.2.2 Time as a third dimension

Using time as a third dimension usually only occurs in situations where the other two dimensions are needed to display other information. One clear example of this is Space-Time cubes. Space-time cubes use two dimensions to provide a spatial context to the visualized content and use the third dimension as the temporal aspect. With Space-Time cubes, the spatial context is most often provided by having one face of the cube show a map of the associated area.

These space-time cubes can be used in all kinds of manners and with different kinds of data. In figure 7 the data is about how an average day might look for the author of the article in Enschede, the Netherlands.

A technique similar in idea, are so-called helix icons. Again, two dimensions are used to indicate spatial context. They differ in that the helix icons allow for more variability in

Figure 7: The spatio-temporal path of a person is visualized using a space-time cube. Image courtesy of Kraak [16]

Figure 8: Helix icons are used to show how certain variables change over time. Image courtesy of Aigner 	extit{et al.} [2]
the attributes which need to be displayed at a certain location, at the cost of being more restricted in the number of different locations that can be shown. Another case in which two base dimensions are used to show other information than time is with Worm Plots by Matthews et al. [20].

Worm plots use two dimensions to plot certain attributes measured in experiments. In discrete time intervals, measurements of these attributes are made and so-called diamond-whisker plots are shown on the corresponding time on the time axis. These plots are connected and shown in different colors for different test groups. An example of a Worm Plot with three different test groups can be found in figure 9.

Figure 9: A Worm Plot is used to show 2-dimensional uncertainty regarding certain variables along a time axis. Image courtesy of Matthews et al. [20].

Figure 10: Time curves are used as a technique to show similarity of multivariate data in two dimensions. The darkness of a point corresponds to its place in time. Image courtesy of Aigner et al. [2].
2.2.3 Abstract Representation of Time

Time is not always represented as a dimension or combination of dimensions. Some techniques rely on different aspects to show different points or periods in time.

A similar way of looking at temporal data can be found in Time Curves as proposed by Bach et al. [3]. Time Curves focus purely on the ordinal aspect of time, i.e., they only show in what order two data records occur. This order is indicated with a line through all visualized points, with the darkness of the point indicating its age. The relative position of the records in two-dimensional space indicates the similarity of two points based on multiple other variables. An example of Time Curves can be seen in figure 10.

One example of this is with CiteSpace 2 by Chen et al. [7]. CiteSpace 1 and 2 were created to analyze and visualize co-citation networks and particularly show the emergence of certain trends in the scientific community. They use color to indicate when a citation is made and through the use of color show the age of a citation within the visualization. In figure 11, the blue links represent the oldest connections, while the orange ones are the youngest.

Figure 11: A CiteSpace 2 image regarding the spread of articles containing info from the Sloan Digital Sky Survey. Colder colors indicate earlier publications. Image courtesy of Chen [7]
2.3 Spatio-temporal Visualizations

Now that we have an idea of how time is handled in visualization, it is interesting to see how this is combined with the other aspect of event data that we are interested in; geographic information. More specifically, we are curious as to how time-dependent dynamics in a geographic context can be visualized to see how this can be applied on processes with geographic information.

A lot of research has been performed on how to visualize data with spatio-temporal aspects. In this section, an attempt is made to capture the state-of-the-art in terms of the techniques developed to support this kind of data. To use this research in the context of process mining, one needs to structurally analyze these techniques to see their possible uses. Munzner [21] proposes a structural analysis framework to decompose visualization techniques. This decomposition focuses on answering three questions:

- What is being visualized?
- Why is it visualized?
- How is it visualized?

Apart from answering these questions, the decomposition also addresses the scale of the data that can be handled through a technique. Scale is examined in terms of data elements, visible data elements and number of attributes within the data.

Through the structural handling of these questions, we can impose a structure on the design space of the researched techniques. This structure allows for easier comparison on the different features available with each visualization technique and for an objective view on the advantages and disadvantages of certain techniques.

2.3.1 Origin-Destination Mapping

Origin-Destination mapping [29] focuses on providing a view on large-scale flows that have a specific origin and destination. It does this by projecting a matrix onto a map and creating a smaller version of this matrix in each cell of the larger matrix. Based on the density of the flow arriving at the cell of the large matrix, a cell is given a color. This cell corresponds to the cell in the large matrix to easily distinguish where certain flows arrive from and where they travel to. An example of migration of people across the US is given in figure 12.

![Origin-Destination Map](image)

Figure 12: An Origin-Destination Map is used to show migration in the United States. Image courtesy of Wood et al. [29]
Structured Analysis

What: Data Table containing origin-destination pairs and intensity of flow between the two.
What: Derived Aggregated flow intensity between origin and destination region, based on original pairs.
Why: Tasks Present and discover information about relative intensity of flow between aggregated regions; lookup and browse interesting origin-destination pairs; identify, compare and summarize origin-destination pairs.
How: Encode Shape: Project matrix over origin area containing cells which correspond to the destination areas;
Color: Sequential continuous colormap to indicate intensity of flow.
How: Manipulate Select origin cells to zoom in.
How: Reduce Filter on regions.
Scale Cells: 20x20 maximum

2.3.2 Origin-Destination Flow Maps

Origin-Destination flow maps [10] visually shows a flow between a certain point of origin and a certain point of destination. The color and the width of an arc between origin and destination can be used to indicate the intensity of the flow or other performance metrics of a flow. Jenny et al. [15] performed a study on what principles in the design of these origin-destination flows are esthetically pleasing. By using curved lines, arrows, and clear nodes between which the flows exist, information can be presented more clearly in these kinds of diagrams.

Figure 13: The top 200 flows of migration between states in the United States are shown using an Origin-Destination Flow Map. Image courtesy of Guo et al. [10]
Structured Analysis

What: Data | Table containing origin-destination pairs and intensity of flow between the two.

Why: Tasks | Present and discover information about intensity of flow between two points or areas of interest; lookup and browse interesting origin-destination pairs; identify and compare origin-destination pairs.

How: Encode | Shape: Arrows from origin to destination indicate a directed flow; Size: Width of arrows are used to indicate the intensity of the flow visualized; Color: Sequential continuous colormap also used to indicate flow intensity.

How: Manipulate | Zoom into interesting areas

How: Reduce | Filter on items or areas.

Scale | Items: Depends on scale of image, maximum in the order of 100

2.3.3 Maptrix

Maptrices [31] strive to solve one of the problems with the classical node-link flow map visualization as shown in Section 2.3.2. Classical flow maps suffer when confronted with dense many-to-many flows. It also seeks to provide Origin-Destination matrices with more geographical context. A standard Origin-Destination matrix is colored using the value at a certain cell, usually the intensity of the flow between the corresponding origin and destination pair. Geographic context is provided by linking the referred parts of the OD matrix to the corresponding parts on a map.

![Maptrix Diagram]

**Figure 14:** The flow of migration between states in the United States is shown using a Maptrix and two maps. Image courtesy of Yang et al. [31].
Structured Analysis

What: Data Table containing origin-destination pairs and intensity of flow between the two.

Why: Tasks Present and discover information about intensity of flow between two points or areas of interest; lookup and browse interesting origin-destination pairs; compare and summarize info from origin-destination pairs.

How: Encode Shape: Lines to origin or destination from corresponding row or column on the matrix; Size: The size of points, and the width of arcs from these points, on the map are used to indicate the flow intensity from or to a certain point on a map; Color: Sequential continuous colormap also used to indicate flow intensity.

How: Manipulate Select areas of interest to highlight them.

How: Reduce Filter on items.

Scale Areas: 50x50 matrix maximum

2.3.4 Flowstrates

Flowstrates [4] attempts to visualize the temporal aspects of a flow by using a matrix. The \( y \)-axis represents a combination of origin and destination in between which a flow exists, while the \( x \)-axis represents a time period. The \( y \)-axis is connected to corresponding regions on a map in a similar manner as the Maptrix discussed in Section 2.3.3. The matrix itself is colored as a heat map.

![Flowstrate Diagram](image-url)

**Figure 15:** The flow from East Africa to West Europe is shown using a Flowstrate. Image courtesy of Boyandin et al. [4]
Structured Analysis

What: Data
Table containing origin-destination pairs and intensity of flow between the two over multiple periods of time.

Why: Tasks
Present and discover information about intensity of flow between two points or areas of interest and changes therein over the course of time; lookup and browse interesting origin-destination pairs; locate interesting points in time in regards of flow intensity; compare and summarize info from origin-destination pairs over time.

How: Encode
Shape: Lines to origin or destination from corresponding row on the matrix, a matrix in which a row represents an origin-destination pair and a column represents a period of time;
Color: Sequential continuous colormap used to color cells in the matrix corresponding to flow intensity between an origin-destination pair during a certain period of time, may also be used to indicate flow origin or destination using a categorical coloring.

How: Manipulate
Select areas of interest to highlight them; Select multiple cells in a row of the matrix to see relative change between the periods of time.

How: Reduce
Filter on items.

Scale
Largest limit on readability is the number of crossing lines between areas and the matrix, another limit is on the amount of time periods visualized.

2.3.5 Flow trees

Flow trees work in a similar manner as Origin-Destination Flow Maps albeit more flexible. Where OD Flow Maps revolve around OD-pairs, Flow trees allow for multiple origins or destinations. This allows for a better view on the flow of goods on a certain path from a multitude of origins or destinations. With multiple destinations comes the risk of having too much information drawn on the map and thus a visualization which is no longer legible. Therefore the amount of origins is usually limited.

Figure 16: The top 50 whiskey exports from Scotland in 2009 is visualized using a Flow Tree. Image courtesy of Verbeek et al. [28]
Structured Analysis

What: Data Table with a root, multiple leaves and the intensity of the flows between the root and the leaves; table containing the location of eventual obstacles for the flow.

Why: Tasks Present and discover information about intensity of flow between a designated source and destinations; lookup and browse interesting flows to certain destinations; identify and summarize intensity of flow from a certain origin to multiple destinations.

How: Encode Shape: Use a square as origin, circles as destinations and lines between the two to indicate flow from an origin to a destination; Size: The width of lines and the size of squares and circles denote the intensity of the flow occurring along the routes and points of interest; Color: Can be used categorically to denote the type of flow of a tree.

How: Reduce Filter on destinations.

Scale Limited in number of destinations and distance between destinations.

2.3.6 Space-Time Paths

Space-time paths [16] place a 3-dimensional space-time cube on top of a map to show the spatio-temporal aspects. Allows for a visual representation of both location (in x and y dimension) and temporal aspects (in z dimension). To really get accurate information from this kind of 3D visualization, user interaction is required.

![Space-Time Paths Diagram](image)

**Figure 17:** The path of an average day of a person from Enschede is shown using Space-Time Paths. Image courtesy of Kraak [16]
Structured Analysis

What: Data Table with rows containing an item modeled, a position (in longitude and latitude) and a timestamp.

Why: Tasks Present and discover information about behavior of visualized items; lookup and browse paths of the modeled items, explore patterns in the visualized data; identify and compare paths of visualized items.

How: Encode Shape: Project geographic information as a map on one of the faces of the cube, represent time as a cube on top of the 2d projection of geography; Size: Can be used to represent characteristics of an event or item visualized, such as scale or intensity; Color: Can be used in a similar manner as size;

How: Manipulate Manipulate the viewpoint to get clearer image of the events or paths visualized.

How: Reduce Filter on specific range of both time and space

Scale Limited in number of events or paths.

2.3.7 Storygraphs

Storygraphs [25] allow for information of in nature 3-dimensional order to be displayed in a 2-dimensional visualization, without any direct loss of information. To accomplish this, storygraphs translate a location \((p_x, p_y)\) to a line from \(x\) on the left \(X\) axis to \(y\) on the right \(X\)-axis. A point on this line at \(y\) on the \(Y\)-axis corresponds to an event happening at time \(y\).

Figure 18: The storylines of 7 Afghanistan based combat units are shown here. Each color indicates a certain unit. Each point on a line means that the corresponding was at the location denoted by the line at the point in time indicated on the horizontal axis. Image courtesy of Shrestha et al. [25].
Structured Analysis

What: Data Table with rows containing an item modeled, a position (in longitude and latitude) and a timestamp.

Why: Tasks Present and discover information about behavior of visualized items; lookup and browse paths of the modeled items, explore patterns in the visualized data; identify and compare paths of visualized items.

How: Encode Shape: Cartesian coordinates are projected on parallel vertical axes, time is represented as the horizontal axis; Size: Can be used to represent characteristics of an event or item visualized, such as scale or intensity; Color: Can be used in a similar manner as size.

How: Manipulate Select specific events or paths to retrieve focus on those events.

How: Reduce Filter on items, geographic range or temporal range

Scale Limited in number of events or paths. A large number of different locations also increases cognitive load.

2.3.8 Data Dials

Data dials [30] combine a geographical representation with a symbolic representation of time. Records are grouped by geographical location, time of day and attribute. These groups are placed on their geographic location, are given an angle which corresponds with the time of day on an analog clock and are given a size based on what size the set has.

Figure 19: Using data dials one can see at what times a search is performed. The 5 different lines correspond to 5 different times of day in which people searched for a specific subject at a certain location. Image courtesy of Wood et al. [30].
Structured Analysis

What: Data Table with per record containing a time and date, a location and optionally extra defining attributes.

What: Derived Aggregated counts per location and time of day of queries

Why: Tasks Present and discover information about aggregated items; lookup and browse the items of a specific location, explore outliers in attributes or location; identify and compare trends of the shown items across times of day.

How: Encode Shape: Lines are drawn with their origin in the corresponding location; Size: The length of lines is used to convey amount; Color: Can be used to indicate extra attributes; Orientation: Used to indicate time of day;

How: Manipulate Hover over specific parts to highlight and receive more info.

How: Reduce Filter on items or temporal range.

Scale Limited in number of different locations. Having a large number of different times on which to show statistics also decreases the legibility of the visualization.

2.4 Conclusion

In this chapter, we have taken a look at some of the previous work in the fields of process mining and modeling, visualization of time and visualization specifically for spatio-temporal data. Comparing process visualization and visualizations for spatio-temporal data, one can notice that they use similar encodings for different purposes. All of the spatio-temporal visualizations use position to show some form of geographic context, while process visualizations use it to show an ordering of activities. Many of the spatio-temporal visualizations use color and size in a similar manner as process visualizations. It seems that position is a critical aspect in the understanding of both visualizations. A further look is needed into if one can unify these two groups of visualizations, and if so, how this unification could work for a process analyst.
3 Determining Geographic Targets in Event Data

To determine how one can unify the techniques studied in Chapter 2, we first need to know what kind of data we can expect in the analysis of event data. In this chapter, we take a look at what kind of common analysis problems we can determine in the geographic information one can encounter during this analysis. Our study into these problems is performed in two parts.

The first part of this chapter describes the result of a structured literature review on what kind of geographic information can be found in the analysis of events with geographic context. The literature review does not confine itself to process mining, or even computer science literature, but strives to create a view of the kind of geographic data used in contemporary research.

The second part focuses on the experiences of experts in the fields of visualization and process mining to determine where they see the possibility to improve visualization of geographic information. More specifically, it focuses on finding how geographic context occurs in business process event data and what kind of information cannot be retrieved easily from this data using process mining. These possibilities are researched through interviews with several professionals.

In the conclusion of this chapter, we try to identify geographic analysis targets from the results of the literature review and the results of the interviews. We also discuss to what degree these targets appear in both real-world cases as the hypothetical cases extracted from the literature study.

3.1 Geographic Information in Data: A Structured Literature Review

Structured literature reviews are methods to extract relevant information from existing literature. A structured literature review takes place in 5 steps: Scoping, Planning, Searching, Screening, and Eligibility.

3.1.1 Scoping

In the first step of the structured literature review, the scope of the research is defined. As the goal of this literature review is to find what kind of geographic information is used in the analysis of events in contemporary research, the research question to scope the review is easy to compose:

What kind of geographic information can be found in data used for the analysis of events?

3.1.2 Planning

In the planning stage of the literature review, a query or number of queries is constructed to collect pieces of literature relevant to the research question.

The search terms are constructed using boolean logic to search multiple possibilities at once. The terms are split into groups. Any combination of words containing at least one of the terms in each group is considered in the query and the query performed only focuses on titles of articles. The query used boils down to the following table:
Google Scholar is chosen as search engine as it has a very complete archive and handles boolean queries well. The following query is used as input for Google Scholar:

allintitle: (analysis OR evaluating OR evaluation OR investigation OR analyzing OR survey) (spatial OR geographic OR geography OR geographical OR space) (events OR event)

The focus on terms appearing in titles only is the result of keeping the review manageable. Executing the query on Google Scholar without this constraint results in more than 5 million results. Although reviewing all these results will increase the completeness of the literature review, we can already retrieve a sufficient view on the state-of-the-art using the results from this limited search, due to the broad range of event analyses in the resulting articles.

3.1.3 Searching

This query resulted in 295 total results in Google Scholar on 18-9-2017. In archiving the papers from the results, a paper could be discarded for one of three reasons.

The first reason for discarding a result is if the full paper is not directly available. The second reason is if the paper is the result of wrong indexing by Google Scholar. Some journals contained all names of the articles in Google Scholar as the name of the journal, which resulted in a wrong indexing. The third and final reason for discarding a result is due to it not being available in English.

In total, 59 results were discarded due to availability, 6 results were discarded due to wrong indexing and 9 results were discarded due to not being available in English.

3.1.4 Screening

In the screening phase, more results from the literature research were deemed irrelevant to the research question and in turn the review. This could be done according to one of the following criteria:

- The paper is deemed not applicable to the research question. One example of this could be that the paper is about an analysis of certain solar events in space. Although the title does match the search criteria, the paper itself is not relevant to the question at hand. 116 results were removed due to this criterium.

- The paper does not contain an accurate description of the data used within the research. Among the papers filtered out using this reason, there are results that introduce models for spatial events, but do not provide a concrete use case for those models. 19 results were removed due to this criterium.

- The papers derive their geographic context from other factors. An example of this is a paper that attempts to identify the geographic origin of papaya’s. 6 results were removed due to this criterium.
3.1.5 Eligibility

After filtering in both the search and screening phase and the removal of duplicates afterward, a total number of 69 results remain. The full list of resulting papers can be found in Appendix A. These results can be grouped according to their field of research.

<table>
<thead>
<tr>
<th>Research Domain</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armed Conflict &amp; Crime Analysis</td>
<td>9</td>
</tr>
<tr>
<td>Biology &amp; Psychology</td>
<td>11</td>
</tr>
<tr>
<td>Hotspot &amp; Event Detection</td>
<td>12</td>
</tr>
<tr>
<td>Urban Development &amp; Tourist Events</td>
<td>6</td>
</tr>
<tr>
<td>Weather &amp; Environmental Hazards</td>
<td>32</td>
</tr>
</tbody>
</table>

Note that there is one paper which is in two categories. This paper was about the analysis of the rebuilding of post-tsunami Aceh, Indonesia and therefore fits in both the Urban Development & Tourist Events and Weather & Environmental Hazards categories.

Geographic Data

When looking at the format of the geographic data handled in the papers, there are a number of standards that are used. Some articles handle geographic data in multiple formats. 41 of the papers handle geographic data by using longitudinal and latitudinal coordinates. 10 papers use coordinates as well, but use a grid with degrees of latitude and longitude with certain values at these cross points. Most of the papers that handle these grids, look at the occurrence of extreme meteorological events and look for environmental patterns in the values from measurements or calculated into these grids.

Another common format of geographic data in the eligible papers is through the use of names of locations. Different formats of these names exist and the chosen format largely seems to depend on the necessity of specific locations. Some articles choose to use addresses or postal codes, which allow for an almost exact geographic location. Others choose to only use the name of cities or counties in their analysis. A total of 13 papers use names of location in one shape or form.

Spatial Granularity of Events

An effort is also made to categorize in what form the analyzed events are present in the data. The granularity of these events can be categorized using two dimensions; in time and space. In regard to space, the most common granularity of events is that of a singular point in space. These points in space are identified by using spatial coordinates. A total of 32 of the eligible articles analyze events in this form.

13 articles do not have a specific format to represent the spatial context of events, but have multiple forms and scales of the events analyzed. For instance, the authors could be analyzing events which have a differing scale in terms of space. An example of these different scales can be found in the analysis of armed conflict events in Sudan by Rose [24]. The events analyzed have a differing scale due to the data set used. The data is based on reports of conflicts and these reports do not all contain the same kind of information. Some reports only include the province where the event happened, while other have the location down to the village.

7 articles use similarly sized areas in their analysis, such as roads or counties. 16 papers do not mention the granularity of their event data, or the events themselves do not have an explicit spatial context.
Temporal Granularity of Events

In terms of temporal granularity, there are more variations of granularity in the analyzed events. In 15 articles, events are handled with differing scopes of time. 11 papers relate an event to a single point in time, while 17 papers relate their events to certain similar periods of time. Of these 17, 11 of them use days as the unit of measurement, 3 use hours and another 3 use years. In 27 articles, the analyzed events do not have an explicit notion of temporal granularity or do not mention this granularity.

Articles for further Analysis

An effort was also made to look for articles that have questions on data that are similar to questions which an analyst might have for business processes. In total, a number of 5 papers were encountered with data that fit this notion and these articles will be used to frame the discussion on the applicability of visualization techniques. The 5 papers with this notion are briefly summarized and their similarity to classical business processes is discussed here:

- Liu et al. [19] propose a method to determine causality between extreme environmental events, such as locust plagues and droughts. This method is based on an algorithm which behaves similarly to an ant colony. The data set, as described in the article, contains information on the type of environmental event, the year in which it happened and the location of the event. A chain of causally related extreme events could be seen as a case of a process, while the extreme events themselves can be seen as events within the process. The information of what case each event belongs to is not directly available in the data set.

- Hadley [11] looks towards establishing factors which correlate with partner homicide-suicide cases and more specifically looks towards the contribution of media reporting in these cases. The data set described in the article contains several characteristics of each case, including information about the perpetrator and victim, the date of the incident, the location of the crime and possible risk factors. For process mining, the data set does not directly work due to a lack of specific identifiable events in each case. However, the leadup to the crime may be interesting to look at from a process perspective.

- Fu et al. [9] look to mine critical failure events from cyber-physical systems which have an abundance of automatically generated events. The authors focus their data set on only power failure events. These events also have a location and timestamp associated with them. A power failure event can be seen as an activity in a process, however, cases are hard to determine using this notion. It is possible to equal the log of one system to a case, but this would result in a number of power failures happening after each other with no other activities present in a possible process model. This is the result of a lack of further data in the data set used in the article.

- Peca et al. [22] propose a scalable clustering algorithm for spatio-temporal data and show its possibilities by using a data set of car movement in the center of Milan to detect traffic jams. The appearance of traffic jams and identifiable factors in traffic behavior could be interesting to look at from process mining perspective. The data set used for the analysis in the article uses GPS-signals of cars traveling less than 10 km/h. The records of the data set contain a timestamp and the longitudinal and latitudinal coordinates of a car at that point in time. Within this data, it is hard to identify meaningful cases and activities. A lone car driving less than 10 km/h does not
directly constitute a traffic jam and thus one needs to look at the spatial and temporal neighborhood of a car for the causes of a traffic jam. This is derivable from the data used in the paper, however using only the data present in the data set would likely result in a process model with a number of cars slowing down resulting in a traffic jam. An expanded data set, including driving behavior of cars driving faster than 10 km/h and information about the environment of traffic jams, could result in far more interesting opportunities for process mining applications.

- Calabrese et al. [5] look at the movement of people from and to special events in Boston. These special events range from baseball games to Friday nights at the Boston Museum of Science. The writers mostly focus on looking for correlation between event location and the location of the visitors, but the movement of visitors, just as the movement of cars, can be seen as a process. A record in the data set used contain longitude, latitude, a timestamp and an anonymized identifier of the phone used by visitors. Furthermore, the home location of the phone users is estimated. The location of the phone is based on signaling events which have quite a large spatial uncertainty. On top of this uncertainty, these signaling events do not occur regularly which could result in an irregular sampling within the data set. Putting aside the quality of the measurements used in the paper, one could use two different notions of cases. One could determine a possible visitor as a case or one could use one of the events as a case. Both would result in difficulties when using process mining on the data. When using a person as a case, one can only use the home location as a factor. This would most likely not lead to a usable model. When using one of the special events as a case, one runs into the same problem as before. With only the home location of a visitor known, one cannot construct a generic process model.

3.2 Use cases for Process Mining with Geographic Data

To get an idea of what kind of use cases for geographic data already exist within the field of process mining, several interviews were performed with professionals in the field of process mining at ProcessGold. The interviews consist of two parts, one looks at cases which the professionals already encountered in their work and at details about these cases. The second part is more about cases where geographic visualization could help the user in the analysis of their process. Each part consists of several subquestions to find out more about each of the cases recognized by the experts. The goal of each interview was to get an answer to the following questions:

- Do you know cases in which geographic context would have helped the user understand their process better?
  - What kind of data was used in the process?
  - What kind of information was the user looking to retrieve from their data?
  - What kind of process did the data come from?
  - How was the geographic context represented in the data?
  - How did the user retrieve that information using the current tools?
  - Where would you think value could be added using visualization of the geographic context with the current data?
• Are there any cases you can imagine in which a visualized geographic context would help the user?
  – If so, what kind of process would the data come from?
  – What would the goal of the user be and what kind of information should the user receive to complete that goal?

Depending on the nature of the question, the answers to each question will either be summarized shortly or handled on a per case basis.

3.2.1 Do you know cases in which geographic context would have helped the user understand their process better?

All interviewees recognize these kinds of cases. A number of cases are identified by multiple interviewees. The case of visualizing Dutch cargo-shipping data is identified by all interviewees. Another case shared by multiple respondents is the case of the Dutch Sociale Verzekeringenbank (SVB). The SVB is subdivided into multiple regional offices and there are cases that are assigned to a certain office, but could as well be performed by another office in one of their rounds. Other cases identified include ship traffic in oceans, handling of postal package logistics, the handling of appeals at the UWV and fraud detection in bank transactions.

3.2.2 What kind of data was used in the process?

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch Cargo Shipping</td>
<td>Paths of cargo ships identified by coordinates of certain checkpoints these ships pass, the cargo held by these ships, incidents that happened with certain ships.</td>
</tr>
<tr>
<td>SVB</td>
<td>Addresses which need to be visited and to which regional office these addresses belong</td>
</tr>
<tr>
<td>Ocean Shipping</td>
<td>Data about ships passing across oceans.</td>
</tr>
<tr>
<td>Postal Packages</td>
<td>Routing details of postal packages</td>
</tr>
<tr>
<td>Appeal Handling</td>
<td>Data from case management system from the UWV (Dutch Social Security Agency)</td>
</tr>
<tr>
<td>Fraud Detection</td>
<td>Financial transactions between certain departments/countries.</td>
</tr>
</tbody>
</table>

3.2.3 What kind of information was the user looking to retrieve from their data?

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch Cargo Shipping</td>
<td>Where to optimally place ship checkpoints so as to maximize the number of ships passing through while minimizing the number of checkpoints. Where certain cargo mainly passes through and whether the possibility of a check has an impact on route.</td>
</tr>
<tr>
<td>SVB</td>
<td>How to combine research from different regional offices to minimize travel</td>
</tr>
<tr>
<td>Ocean Shipping</td>
<td>How to detect unusual behavior in ships, e.g. where are ships not allowed to pass through or what are areas commonly evaded.</td>
</tr>
</tbody>
</table>
Postal Packages  How distribution centers of packages work together.
Appeal Handling  Differences in the execution of a set process between offices.
Fraud Detection  The identification of suspicious traces of money.

3.2.4 What kind of process did the data come from?

Dutch Cargo Shipping  The process of cargo ships transporting goods through rivers in the Netherlands and inspections of these ships by the Dutch government.
SVB  The checking of fraud cases in Dutch insurances.
Ocean Shipping  Logistical process of ships sailing across oceans.
Postal Packages  Handling of the delivery of postal packages.
Appeal Handling  The handling of appeals in different offices of the UWV.
Fraud Detection  Financial transactions between international institutions.

3.2.5 How was the geographic context represented in the data?

Dutch Cargo Shipping  Coordinates of the ships.
SVB  Zip-codes.
Ocean Shipping  GPS locations of ships measured at intervals differing from a couple of seconds to a couple of minutes.
Postal Packages  Zip-codes.
Appeal Handling  The location of the handling office of the UWV.
Fraud Detection  IDs of countries and names of departments/institutions in these countries.

3.2.6 How did the user retrieve that information using the current tools?

Dutch Cargo Shipping  Using Excel.
SVB  The user did not retrieve the information beforehand.
Ocean Shipping  Periodically the current location of ships was checked.
Postal Packages  The user did not retrieve this information beforehand.
Appeal Handling  Treat the location as a pure categorical value.
Fraud Detection  Location was treated as a categorical attribute and presented using classical visualization techniques (i.e. bar charts).
3.2.7 Where would you think value could be added using visualization of the geographic context with the current data?

Interviewees agreed that value can be added in cases where location is one of the core aspects of the data. The manner in which location can add value depends on the case. In the Cargo Shipping case, visualizing the flow between these locations is deemed valuable, while in the Appeal Handling case, having the location of an office is deemed valuable. One interviewee noted that using maps can greatly increase the intuitive use of a tool by a user. Maps speak a lot to the experiences of a user and thus allow for easier analysis of spatial data.

3.2.8 Are there any cases you can imagine in which a visualized geographic context would help the user?

The interviewees provided many different cases in which they think visualizing geographic context helps the user. Many of these cases revolve around interaction on a spatial level. It could be interesting to show processes with certain spatial patterns and allow for easier analysis by the user. An example of this could be an assembly process with the different steps of the process happening in different locations. It could be that certain locations or changes of location take extra time and visualizing this could provide more information for an analysis. An extreme case of this kind of process could be a process which flows along physical locations. This could, for instance, be the delivery of packages as already noted earlier, but could for instance also be the way physical packets are transported through a building with regular check-ins at different locations using RFID.

Another example could be a process where the location of the process has influence on how the process is executed. This case can be interesting in multiple aspects of the analysis. An interesting analysis task is to look at the differences in executions between offices which should have the same process. Another interesting analysis task could be to look at the way supply and/or demand is handled on a geographic level for a regional process. It might be interesting to look at the spread of both customers and suppliers for a certain process and visualizing this information using a map might make this information more accessible.

One interviewee also noted that the analysis of processes itself should not always be done on a map, but that maps can help a lot in the presentation of information gained in this analysis.

3.2.9 If so, what kind of process would the data come from?

The interviewees shared the view that the process should have some degree of interaction between geographic locations. This can be on different scales, ranging from the interaction between departments within a company or the flow of goods within a distribution center to the flow of goods between countries.

3.2.10 What would the goal of the user be and what kind of information should the user receive to complete that goal?

The goal of the user, according to the respondents, would be to find patterns in spatial behavior. One example of these patterns could be that a certain set of activities or transitions at a certain point in space are relatively slow, whilst another could be the discovery of unexpected routes existing in the process.
3.3 Conclusions

Both from literature and from the interviews, one can see certain aspects from data appear that are of interest to an analyst. Using the terminology of Munzner [21], we identify 5 targets. One of these targets revolves around processes in which each process step coincides with a change in location. This can be seen in literature, in the article by Calabrese et al. [5] and in the interviews, in for instance the case about the handling of postal packages. Both these cases have the property that each step in the process is about the movement from one location to the next. They also share the fact that the geographic information is the most interesting aspect to the analyst.

Another target which can be spotted, albeit mostly in the interviews, is the case where the activities have a prominent geographic component and where this geographic component has a significant effect on the performance of the process, but where the geographic component of the activity is not the only important part. From the interviews, one case that could fit this target is the case of an assembly process of an airplane. The separate parts are manufactured at different locations and later assembled at another facility. This results in a process in which location is an important part of the process, but where it is not the only factor to affect the results of the process.

With the third target, the most interesting information to the analyst revolves around being able to see the interaction between geographic areas or locations that occurs within the process. An example from literature that touches upon this target, is the article by Liu et al. [19]. The most interesting aspect of the process is the possible cause-effect relationship between pairs of disaster-region tuples. From the interviews, an example of a use case related to this target is the Fraud Detection case. These cases have in common that to the analyst, the presence and form of interaction between regions is the most interesting aspect.

Another target is the case where the events themselves do not have a geographic property, but where the cases do. This can, for instance, be seen in the case of appeal handling at the UWV. The processes should be standard across the different offices at the UWV, but the execution is different in different offices. The geographic information in this target is not necessarily an important aspect of the analysis, but can be used more as a tool to identify the different offices. In literature the same target could be found in the work of Hadley [11], where the area could be an important factor in the leadup to the studied homicides.

The final identifiable target looks more towards the external stakeholders of a process. It can, for instance, be interesting for an analyst to have an idea of how the suppliers and customers of a certain company or office are spread geographically. The geographic information is then not an explicit part of the process behind a company, but by using that information, a company may be able to optimize their process in regards to present external factors. A clear example of this could be the case with the SVB in the interviews.

To summarize this, we identify the following targets:

<table>
<thead>
<tr>
<th>Target</th>
<th>Literature</th>
<th>Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Geographic Step ↔ Process Step</td>
<td>Calabrese et al. [5]</td>
<td>Postal Packages</td>
</tr>
<tr>
<td>2. Geographic Step ⇒ Process Step</td>
<td>-</td>
<td>Assembly Process</td>
</tr>
<tr>
<td>3. Interaction between Areas</td>
<td>Liu et al. [19]</td>
<td>Fraud Detection</td>
</tr>
<tr>
<td>4. Comparison between Areas</td>
<td>Hadley [11], Fu et al. [9]</td>
<td>UWV Appeal Handling</td>
</tr>
<tr>
<td>5. External Stakeholders</td>
<td>-</td>
<td>SVB</td>
</tr>
</tbody>
</table>
3.3.1 Identifying Primitives

Not all targets have cases related to them from literature and comparison of the visualization techniques discussed in Chapter 2 to these targets directly is hard without defining what we would want to see in the cases in each target. What each target has in common is that there are process primitives and geographic primitives which need to be visible to solve the most important questions related to that target. To identify process primitives, let us look at the visually enriched process model in figure 3. From this process model, we identify the following process primitives with corresponding figures which show each primitive:

- **Events (i.e. an occurrence of a process step)** [Figure 20]
- **Dependencies between Events (i.e. co-occurrence of events)** [Figure 21]
- **Attributes of Events (i.e. cost or waiting time)** [Figure 22]
- **End-to-end flow of a case (i.e. complete list of events of a case)** [Figure 22]
To identify geographic primitives related to a process that we could want to have visible, let us look at Figure 14. The following geographic primitives can be identified with corresponding figures which show each primitive:

- **Geographic context (i.e. a point or an area)** [Figure 23]
- **Change of context (i.e. movement between points or areas)** [Figure 24]
- **Relation between contexts (i.e. correlation between points or areas)** [Figure 24]
- **Comparison between contexts (i.e. comparing certain statistics between points or areas)** [Figure 25]

Now with the identification of these primitives and the study on available techniques in Chapter 2, the next step is to find a way to combine the specific components each target needs with the features available in the studied visualization techniques.
4 Connecting Visualization Features and Geographic Targets

In Section 2.3 a survey is performed on the techniques available for showing spatio-temporal data, while in Chapter 3 targets in handling geographic data in process mining are determined through literature research and interviews, and we identify common process and geographic primitives from these targets. The question arising from these chapters is how can one connect the visualization techniques brought forth in Section 2.3 to the targets in Chapter 3 in a useful and meaningful manner. To answer this question, two subproblems need to be handled first.

One of the subproblems is about how to find what combination of primitives needs to be visualized to show a certain target from a process with a geographic context. The other subproblem is about how to select a visualization technique, that use specific visualization features, for a given dataset and target. To solve this subproblem, a closer look is needed into what kind of features can be obtained from different visualizations specifically for the targets identified in Chapter 3.

In this chapter, we first attempt to solve the first subproblem through the use of the cases found in literature and hypothetical questions on these cases. With these cases, we can start to reason about the possible fit between a visualization technique and a set of primitives. The targets which do not have a direct example in literature are explored using the cases from the interviews and hypothetical data related to these cases.

Then, to solve the second subproblem, we look to generalize and summarize the visualization of targets and primitives reasoned about in Section 4.1. With this generalization, we create a framework in which we can relate the techniques of Section 2.3 to the primitives from Chapter 3. This framework offers a structural way to compare the visual primitives needed with the techniques that enable them and should allow users to make their own choices regarding the visualization of their own cases.

4.1 Exploring Targets in Literature

In Chapter 3 5 papers were selected from the eligible papers with data that have similar analysis questions as can be encountered in the analysis of business process. In this chapter, the content of these papers is further examined. We examine the possible application of process mining on the used data sets or hypothetical expanded data sets and the difficulties of applying process mining on the used data sets are discussed. Then we discuss how the problems within these hypothetical use cases could be solved by visualizing specific sets of the primitives described in Chapter 3. We also explore the possible applications of the visualization techniques from Section 2.3 for each paper. In this discussion, we use the same terminology as used by Javed et al. [14] to discuss composite visualization.

4.1.1 Event Space-Correlation Analysis Algorithm Based on Ant Colony Optimization

Liu et al. [19] put forth the hypothesis that certain natural disasters (i.e. earthquakes, floods etc.) are not isolated events, but have some connections. The authors model different environmental disasters as nodes in space-time and use a simulation of an ant colony to see what paths between nodes are the most used and thus the most likely to be correlated.
The data used to test their model and simulation are historical records of droughts locust plagues of the Henan province in China. The events in the records are adapted to have an associated longitude and latitude, but the temporal granularity of these events is not discussed in the article. Suppose a complete data set with georeferenced natural disasters along with semi-accurate timesteps would be available, in what way could process mining techniques be applied to this data set?

One of the first problems encountered when trying to model this supposed process as classical process model is that there is no direct notion of a case. As finding correlation between the disasters in the data set is one of the goals of the authors of the article, one cannot directly determine what events belong to what case. Without the possibility to determine what constitutes a case in this context, it is impossible to use classical process mining techniques.

Now suppose that there exists a method to correctly identify correlated disasters and thus allow for the identification of cases. Identifying activities will become another hurdle. Suppose one would take the type of the natural disaster as the activity (i.e. earthquake, flood etc), the resulting model would lose a large amount of geographic information. It seems very reasonable to assume that, while certain disaster types occur more often in sequence and have some notion of correlation, the geographic location of the disasters also has a large influence on future disasters happening. Disregarding this information could result in a process model that leads to wrong conclusions.

One way to include the previously missing geographic information in the model would be to use the location as an extra factor to identify unique activities next to the disaster type. However, it is extremely unlikely that two disasters occur at the exact same location. Therefore, each case would most likely result in their own string of activities in the model and thus not deliver a good overview. To counter this, a logical grouping of the location would be needed and in achieving this, two factors need to be considered. As already noted, the grouping cannot be too small as this would result in a model that is not able to adequately generalize the cases in the data. The grouping can also not be too large as to allow for the loss of information about the relation between certain regions.

Ideally, we would construct logical groups adapted to the geographic scale of the research being performed. In the case study performed in the paper, these groups would likely be the different regions identified of the province of Henan. Correctly identified groups may lead to identifiable trends in the order of certain disaster-region combinations, however using standard process model visualization in this case may not be ideal. Users may not directly be able to relate names of a region to their geographic location without the proper topographic knowledge.

**Required Visible Primitives**

In Chapter 3, this case from literature is linked to the third target where the interaction between areas is the most interesting part of the data to the analyst. In terms of needed primitives, we need to look at what kind of information is most interesting to an analyst. In this case, an analyst would be most interested in seeing what disaster-region tuples are correlated. The authors attempt to show this using a space-time path as can be seen in figure 26.

For the process primitives, we first need to define the process we are looking at. Suppose we take a set of causally related disaster-region tuples as a case and each unique disaster-region combination as an activity. From a process standpoint, the analyst would then be most interested in the events and the dependencies between events.
Figure 26: Liu et al. [19] use a form of space-time cubes to show correlation between historical disasters.

The geographic primitives are closely related to this definition of our process. An analyst needs to have a view on the geographic context of an event and the relation between regions to accurately find out possible connections.

Matching Visualizations

In figures 27 and 28 possible visualizations of a very simple case from within the data can be seen. The simple case here is that droughts in Nanjing and Northern Shandong lead to locust plagues in Northern Henan.

An initial option to visualize the information selected is through a standard process model as in figure 27. The basic activities are identified as being "Drought, Northern Shandong", "Drought, Nanjing Yangzhou" and "Locust Plague, Northern Henan". Although the causality between these three activities seems pretty clear from the model, its geographic context can only be found in the activity names. This kind of view allows for the visualization of all process primitives, but all geographic primitives are lost in the standard process view.

Figure 27: Here the process is shown using regular BPMN.

Figure 28: In this figure, the activities are projected on their geographic location.
Figure 29: The grouped locations are shown through the circles. In the circles, the different disasters are shown with a color indicating the type of disaster. The colored arcs indicate correlation between certain disaster-area tuples.

Figure 28 shows the result of a number of changes in the regular process model to a georeferenced variant. This variant uses the geographic location of the points established in the activities as their actual location in the model. This kind of visualization shows both necessary geographic primitives, as with this view we get to know more about the actual location of the events and we see a causal connection between the regions. Color is used to allow for a distinction in the type of activity that takes place. Red, in this case, is used to indicate droughts, while blue indicates locust plagues. The width of the arc between two activities is used to indicate the dependency between events and thus also shows a relation between different regions. These decisions allow for easier analysis of geographic context. For instance in this case, an analyst can conclude that droughts in regions closer to the coast tend to take place before locust plagues more inland.

For this relatively simple example, the georeferenced process model seems to create a nice result, but this example is not representative of the diversity or the scale of the data that can be found inside the article. Suppose there is a slightly more complex data set, what would the effects be of using this kind of visualization. One result of the georeferencing in the process model is, that, one cannot initially have two different types of disasters on the same place as this would result in these activities being placed on top of each other in the visualization. One way of solving this problem is by abstracting slightly from the regions. An example of what this could look like can be found in figure 29.

Figure 30 shows a figure of an adaptation of the Origin-Destination Map discussed in Section 2.3.1. The problem in showing the previous data using Origin-Destination Maps is that the pairs of locations have multiple categorical variables which need to be shown. To solve this problem, pie charts are used to indicate the ratio of the number of disasters which have a causal relation with disasters in other regions. To indicate total "flow" of events to other regions, there are multiple options. The classical Origin-Destination Map uses opacity to show this flow, but together with the use of multiple colors, it becomes hard to compare the total flow of different regions. Another way of showing this total flow is by using the size of the pie chart.
This approach has some advantages, but also some downsides. By using an adaptation of Origin-Destination Maps, there is less clutter in the resulting image than with the georeferenced BPMN-like model in figure 29. However, in the resulting image, not all information from the georeferenced BPMN-like model is shown. One cannot see what disasters have correlation between them and thus analysis on that correlation is not possible using this image. This means we lose the primitive regarding events, as we can no longer distinguish these events based on type.

Another approach with similar advantages and disadvantages can be found in an adaptation of Origin-Destination Flow Maps. These flow maps have the same problem as the original Origin-Destination Maps in that they are usually only made to convey one type of flow. This could be solved using pie charts to show the distribution of the different types of flow between regions. In the case of Origin-Destination Flow Maps, this can be solved by splitting the arrow conveying flow into different sections corresponding to different activities.

Comparing this method to the adapted Origin-Destination Map, it is easier to use in the case of a relatively small number of different flows. Arrows have a direct clear meaning of origin and destination, while the matrix in an Origin-Destination Map requires extra explanation. If there is a relatively large amount of combinations of regions which need to be displayed, the Origin-Destination Map becomes more suited due to it not having as large glyphs on its resulting image. However, with both techniques we have the disadvantage that although we can see relations between different areas, we are unable to see dependency between the events. This is due to the fact that we cannot see the type of the disaster of either the originating or resulting disaster-region tuple.

Another possibility is to use an adaptation of Flowstrates with the columns in the matrix not indicating a certain period in time, but a certain type of disaster. The results of this adaptation can be found in figure 32. This adaptation has some of the same issues as the previous adaptations to both the Origin-Destination Map and Origin-Destination Flow Map. It is again not possible to see the resulting event.

A way of solving this can be found in figure 33. By using color in a different way than a regular flowstrate, one can show both cause and effect in one image. Instead of using color to denote the origin-region, color is used to indicate the type of disaster. The column still denotes the cause disaster type, but the colors of the pie chart in a cell indicate the resulting disaster type.
4.1.2 Intimate partner homicide-suicide: The role of media in depicting life-ending events, along with an analysis of the prevalence and geographic distribution of these events

In this article, Hadley [11] looks for factors that play in the process that leads up to partner homicide-suicide. As data sources, several websites and newspaper reports were used to collect incidents with enough information for analysis. This information includes characteristics such as perpetrator age, gender, and ethnicity, but also the date and the location of the incident.
The geographic information used in the analysis is limited to just the location of the incident, while the temporal information is limited to the day of the incident. This is also reflected in the visualization used in the thesis. Figure 34 shows the geographical spread of cases across the United States.

Cases themselves are easy to identify as each incident can be used as a separate case. However, activities are a lot harder to identify, especially with the data described in the article. Although all kinds of factors were used in the analysis of these cases, much of these are hard to pinpoint as an activity. While a divorce or the request for a restraining order may be clear as activities, not all cases include these kinds of factors. Furthermore, this kind of analysis also depends on a complete data set, which in this case is highly unlikely.

Suppose there is a complete data set with a complete collection of the events of significance in each case. It may be possible that each case has an almost unique set of events associated with it. The mining of such a data set may then result in a spaghetti-type model, with a branching path for each case. Using the model for analysis is also risky, due to only having data on the cases that result in homicide-suicide.

All in all, this kind of data does not seem suitable for process mining, but a look at how the geographic context is used is still interesting. As mentioned before only the location and time of the incident is recorded, so the activities preceding the incident do not have a location or time associated with them. The only possible analysis based on geographic location would be to see the differences of the process between different regions. In this analysis, the hardest obstacle to overcome would again be to create meaningful groups of cases. These groups would need to contain some difference in the process to analyze, but would also need to be large enough to be able to be representative of the region represented in the group.

Figure 34: The geographical spread of cases researched is shown in this image. Image courtesy of Hadley[11]
Required Visible Primitives
In Chapter 3 this case was linked to the fourth target where comparison of the process between areas is what a user wants to know. Once again, creating a proper grouping on the area is identified as a problem in performing process mining on this kind of data. Suppose we would have this correct grouping, the most interesting information that could be gained through the use of spatial information are the differences in the process, as is noted in this case as well as the target the case corresponds with.

From a process standpoint, this means that we should be able to see all primitives of the process model, while from the geographic primitives we should be able to see the spatial context of the process and we should be able to compare the process between different geographic areas.

Matching Visualizations
As there is no direct interaction between regions available in the data, Origin-Destination visualization does not make sense in this case. The question then remains as to what adapting process models are able to show in terms of the required primitives.

This kind of data leads to two possibilities in the presentation of the process model. These two options can be seen in figure 35 and figure 36. The models as shown in these figures are not representative of any data from the paper and are purely hypothetical.

These samples models for Texas and New Mexico are one of the very few situations where one may project the models on their corresponding regions without having a lot of negative side effects. The BPMN models themselves are short and simple in that they do not have a lot of branching paths. This means they can still be relatively easily be compared to each other despite their distance from each other. The shape and size of both New Mexico and Texas allow for the placement of these models in a relatively easy manner.

However, this example will probably not be representative of the actual processes. The processes are highly unlikely to be this simple and, most likely, the regions encountered will not be as easily usable as the states used in this example. Suppose one would want to compare multiple different states over the whole country at the same time. If one would project the processes onto the states themselves, the result could be similar to the one found in figure 37.

![Figure 35](image1.png) ![Figure 36](image2.png)

**Figure 35:** Regular BPMN models next to each other for New Mexico and Texas

**Figure 36:** The BPMN models are now shown on the states they correspond with.
Figure 37: Expanding the idea of showing the BPMN models on each state seems non-ideal as this results in an overly busy image.

Figure 38: The most general model is shown at the right, while differences between states are projected on their corresponding state.

This does not directly seem as a usable solution as the models themselves become hard to read and finding any difference between the models becomes very hard. As any analysis on the geographic data will most likely be on the differences between the models, it could be beneficial to only focus on these differences. One way of handling these differences can be found in figure 38. By showing the most common variant as a main point, the process remains readable. The differences between the process in a state and the most common one can be seen projected on each state. This method allows for easier comparison of the differences, but relies on the fact that these differences are fairly simple.

One way to handle a relatively small number of complex differences is to cluster these differences and only show these groups on the map. The user could then use interaction to select two variants of the process and compare these side-by-side. An idea on how this could look can be found in figure 39. What this view allows is to see geographic patterns in the similarity of the process, however, it also makes some assumptions on the possibility of these differences to be grouped. If there are a lot of different variants, these colorings may not add any value.

Overall, it seems that projecting a process model onto a geographic visualization has a lot of issues due to them relying on position for understandability. A process model relies on position to guide the user through the flow of the process. An activity that occurs a lot after another activity will be displayed below or to the right of the latter. Realistic geographic
Figure 39: At the right, two selected models are shown. Each color corresponds to a cluster of similar BPMN models.

visualization relies on position as well as geography itself can be seen as nothing other than a set of positions. If one would want to combine the two, it would either impact the readability of the process model negatively of the readability of the geographic visualization. In this case, this is most notable through the lack of appropriate areas to plot the needed process models in realistic geography. Two separate visualization spaces, as shown in figure 39 seem like the most readable solution.

4.1.3 Online temporal-spatial analysis for detection of critical events in cyber-physical systems

Fu et al. [9] propose a method to efficiently identify actual errors in clusters of cyber-physical systems. They use a temporal window to identify error reports close to one another in time and use a clustering algorithm to determine whether spatially they are close as well. The data set used for the case study of the algorithm is a set of 300,000 smart meters over the year 2012. The article focuses on the "power failure" events that the meters send out and link them to calls received by customers about power outages. Each "power failure" event in the data set has an associated longitude, latitude, and timestamp.

Looking at this case from a process mining perspective, the activities themselves seem pretty clear. A power failure event sent by a meter can be seen as an activity, however doing so would create a challenge in finding suitable cases. A single meter cannot be seen as a single case using power failure events as activities. The traces themselves would only consist of a series of power failures with no other activities. As the relative location of the smart meter to other power failures seems to be a leading factor in whether the meters themselves sense a power outage, relative location would seem a likely choice to include into the activities. This relative distance is important as each meter is connected to a substation, which supports multiple meters and can cause a power failure. The problem in using relative distance is that there is a need to cluster the meters spatially, as these clusters are not present explicitly in the data. Usage of these clusters in process mining would then thus only be possible if it is known to what cluster each sensor belongs.

Suppose this is known and there is still the question of how the process of these power failures would look within each cluster. In this case, it may be interesting to look at the full scope of the messages that the sensors produce, instead of only looking at the power failure messages. This allows for a better analysis per cluster on how these outages appear and for easier spotting of failures.
In geographic terms, the location of the sensor initially only may seem interesting for investigating the clusters of sensors. However, there may be a causal relation between the location of a sensor and the power outages it reports. When visualizing this, it may be interesting to have an overview of where these failures keep happening and if there are trends in a certain area. If the number of failures is the most important aspect, a heat map of power failures will already provide enough information. If specific variations of the power outage process are more interesting, process mining techniques may be needed to discover these variations.

**Required Visible Primitives**

There are two ways to look at the process behind the data in the article. The analysis can be purely about which clusters fail more often than others. However, as this does not require process primitives to be visualized, a simple geographic visualization may suffice.

If the analysis, however, concerns the processes behind the failures, process primitives become required for analysis. It may be interesting to look at similar techniques as experimented with in Section 4.1.2 as both cases share the same target.

**Matching Visualizations**

The article itself uses a simple map to show the idea behind how the meters are placed and how multiple power failure events may occur in a spatial context. This map can be seen in figure 40. This image is useful to show a group failure, but is intended to only show whether a meter has failed. It is not possible to see the number of times a certain meter has failed. This could be solved by applying a heat map coloring to the map.

### 4.1.4 Scalable cluster analysis of spatial events

Peca *et al.* [22] propose a method to cluster spatio-temporal event data that is meant to be able to handle the large scale often found in data sets about events in space-time. The data set used as a test case in the article is a data set containing 17,200 GPS-tracks of cars in Milan, with approximately 2,000,000 measurements. The authors attempt to use their method to detect traffic jams by detecting clusters of measurements with a speed less than 10 km/h. A
traffic jam is seen as a collection of low-speed measurements which are close to each other in space and in time. A single traffic jam then has an associated spatial area spanning the location of all associated measurements and a period containing all the timestamps of the associated measurements.

This case seems similar to the previous one in terms of that one measurement with a speed of less than 10 km/h does not mean that there is a traffic jam. The actual interesting events consist of multiple measurements close in space-time. In terms of process mining however, this data set does not seem very interesting. This is due to the fact that although the detection of traffic jams is interesting, there will most likely not be a way to determine the causes of these jams based on the data available.

What might be more interesting is a look at the behavior of vehicles leading up to a traffic jam. From this perspective, a traffic jam can be seen as a case for process mining. The difficulty in this approach comes both from a lack of a direct connection between cars and a case and from the trouble in defining an activity. Although there are of course cars in a traffic jam, these may not be all cars whose behavior leads up to the traffic jam. Only linking the cars present in a traffic jam as applicable to that case may be wrong.

Now suppose there exists a direct link between all cars relevant to a traffic jam and the traffic jam itself. It would be very hard to identify what kind of behavior should constitute an activity. Not all behavior necessary to detect patterns is captured in the relatively simple variable of speed. From the data set in the article, only the GPS location of cars driving less than 10 km/h is used to detect clusters and this is most likely not enough to capture all the driving information needed to look at the lead-up of a traffic jam.

**Required Visible Primitives**

This case does not have an explicit link to the targets set in Chapter 3. This is due to the fact that we can link this specific case to two different targets depending on the data itself. Looking for the causes behind traffic jams, we may want to focus on two different targets of the geographical data. Suppose one would want to purely look at differences in the behavioral process leading up to a traffic jam, it would be quite logical to link it to the target regarding comparison between areas. Having a similar approach to the one in Section 4.1.2 seems logical. However, with this case, the assumption can be made that one of the most important factors in the appearance of a traffic jam at a certain crossing is the state of other crossings in the neighborhood. With this assumption, it could be that linking it to the target regarding interaction between areas fits better and thus that a method similar to the one described in Section 4.1.1 may work.

One could also argue that the state of the other crossings are the only important factor to visualize. If so, this case would fit more within the first target. As we have not yet discussed the needed primitives for that target, we focus on finding these for this kind of analysis.

In this case, events can only be traffic jams taking place at a certain location. This means that from the process primitives, we only need to be able to see dependencies between events. From the geographic primitives, we need to be able to see geographic context. The ability to see relations between different locations in this case is the same as the process component regarding dependencies between events, due to the relatively simple nature of these events.

**Matching Visualizations**

The authors use a three-dimensional visualization to show the spatio-temporal clusters they discover in the data. Two examples of these images can be seen in figure 41. The z-axis is used
to indicate time and the color of a cluster indicates its geographic position (i.e., northwest is blue, southeast is red). Some patterns can be spotted in the data, but it is quite hard to determine a cluster’s exact spatio-temporal position based on purely this image.

As noted before, three-dimensional visualization of spatio-temporal data has some problems. Most notably, one cannot directly identify the location in space-time from an item in one image. One way of solving the problem is by splitting it up and allowing the user to specify what kind of information they want to see. If the user is more interested in the temporal distribution of traffic jams in certain areas, they could select these areas and a two-dimensional plot could be made of how many traffic jams there are at these areas at certain times of the day. If the user is more interested in the spatial distribution of the traffic jams in certain periods of time, they could select these periods and the number of traffic jams could be projected on a two-dimensional map of Milan. This allows for a clearer image on both different distributions, albeit at the cost of not allowing direct insight into both the temporal and spatial distributions.

Looking purely at the case linked to our first target, origin-destination flow visualization techniques seem like a good fit. Assuming that we can associate a traffic jam with the affected roads and intersections, the flow in this case could be interpreted as the cause-effect from one spot to another, with the number of times this happened as the intensity of the flow. Using a flow map could be a very good idea here, as the cause-effect relations will most likely not have a lot of distance between them and lines between these relations will most likely not cross, due to the nature of the data. An example of this can be found in figure 42.
Due to the fact that most traffic jams naturally start at intersections, one can project the lines indicating flow onto the roads connecting these intersections. It is highly unlikely that two traffic jams that are correlated do not share a connection by road between them, so there should almost always be an easily drawable arrow between each related pair of intersections.

This kind of simple visualization is already able to show all required primitives and has the added advantage of being a relatively easy to understand image. Other techniques may also be able to show the same primitives, but may increase the difficulty of use of the image.

4.1.5 The geography of taste: Analyzing cell-phone mobility and social events

Calabrese et al. [5] analyze cellphone traces to identify common factors among visitors of certain events in Boston. The data set used consists of 130 million estimations of locations from close to 1 million devices with associated timestamps. Using this data, the authors identify a similarity in audiences between certain events.

From a process mining perspective, there are different ways of looking at this data set. One could choose to designate each social event as a case, but activities in this case may be hard to determine. A more obvious choice for cases may be a trace from one telephone. Activities may then depend on the kind of analysis one would want to perform. It may be interesting to look at the general flow of people from and to these events, while another interesting aspect could be the characteristics of people visiting the social events.

In the first case, it would be good to define activities as certain points in traffic where a trace may pass through. Identifying these points and using our definition of a case, the process boils down to a fairly simple one where each process step corresponds to a change in location as well.

In the second case, process mining in itself may not be most effective to determine the factors that contribute to a person visiting an event. The authors use a neural network to determine leading characteristics of the people attending certain events. Data mining techniques seem more suited to solve the question of what factors determine what kind of people visit which events. If there is a geographic factor, as the authors conclude, it may be useful to visualize this trend. A combination of bar charts and heat maps should be able to provide all the information regarding the spatial context of the visitors.

Required Visible Primitives

In the first identified case, it is clear that this case belongs in the first target. A similar approach to the one discussed in Section 4.1.4 may offer a good solution to visualize the flow of people through the city.

4.2 Further Targets

To analyze the targets that do not have a direct example in one of the selected pieces of literature, we take a look at the processes that were mentioned in the interviews.

4.2.1 Geographic Step ⇒ Process Step

The example process mentioned in the interviews characterizing this target regards an assembly process that takes place over multiple locations. This kind of process does not have the restriction that process steps can only coincide with geographic steps, e.g. multiple process steps can take place at one location.
Figure 43: A simple example of how juxtaposed visual objects could help in the understanding of a process with multiple geographic locations, but where not all process steps coincide with a geographic step.

Required Visible Primitives
Analysis on process within this target will be reliant on being able to show both complex process primitives and geographic primitives. To maintain an idea on how the process is structured, we need to be able to see events, dependencies between these events and an end-to-end flow of a case. To show how the process flows through multiple areas, we both need the geographic context of a set of events as the change of contexts visualized.

Matching Visualizations
In choosing a visualization for this target, we run into a similar problem encountered in Section 4.1.2. Placing activities in a logical ordering for the process in a space with geographic context implies that these process steps take place in different places and thus increases the difficulty of the visualization. Placing activities on their exact geographic location results in multiple activities overlapping each other and thus becoming unreadable.

Separating process information and geographic information seems to be the best idea in this case as well. Figure 43 shows a simple example of how such a separation may work. In this example, there are three offices of which each, execute a certain part of the process. On the left part of the image, there is a geographic map which shows the interactions between the offices. On mouse-over or selection of one of the offices, the process at that office can be seen where the transitions to or from other offices are clearly marked.

4.2.2 External Stakeholders
The example process mentioned in the process is the process behind researches from the SVB. One of the most important parts of information is the location at which these researches need to be performed, relative to the location of the regional office which performs the research.
Required Visible Primitives
To show this spread, we need relatively few process primitives visible. We only need to show an attribute of the events, with, in this case, the location of the research. From the geographic primitives, we need geographic context of these attributes, relations between these contexts and the possibility to compare different areas on how these researches are spread.

Matching Visualizations
A target that requires relatively little in terms of the visualization of the process, seems well suited for our researched geographic visualization techniques. A way of applying the researched origin-destination methods may be to see the research as an origin and the location of the responsible office as a destination. By taking this definition, we can show all the required primitives. An example of this can be found in figure 44 with an Origin-Destination Flow Map. The major blue dots correspond with certain locations (e.g. offices) and the arrows indicate where the external factors come from. Some aggregation on area is needed to not overload the image with arrows, but some outliers may not be as clear as one would want.

Another way of showing external factors can be seen in figure 45. Here, the colors of the small dots correspond to the location for which they are an external factor. The execution of the idea suffers from a relatively larger range of colors. This makes the image harder to read and analyze. Another idea may be to assign colors to areas instead of points. The colors of certain areas could then correspond to certain offices for which they are an external factor.

Overall, within the techniques, researched an Origin-Destination Flow Map seems to be the best option for the relatively simple data present in this target.

4.3 Framework for Geographic Targets in Event Data
Now that we have covered all targets from Chapter 3 and discussed the possible mappings of these targets to visualization techniques using concrete cases, we return to the question set at the beginning of this chapter: How does one determine which geographic target fits with which visualization technique. To answer this question, we look to generalize the discussion held in Sections 4.1 and 4.2 to achieve a conclusion on which visualization technique fits with
what targets. With this knowledge, we establish a generic framework which we can use to
determine what visualization one should use for process analysis with the identified targets.

4.3.1 Primitives and Visualization Techniques

The table below indicates the process and geographic primitives that the discussed techniques
from Chapter 2.3 are able to visualize, based on the work in Section 4.1 and Section 4.2. The
matrix is mostly based on the discussion in Section 4.1.1 as the limits and possibilities of the
techniques are discussed there.

<table>
<thead>
<tr>
<th></th>
<th>OD-Map</th>
<th>OD-Flow Map</th>
<th>Matrix</th>
<th>Flostrates</th>
<th>Flow Trees</th>
<th>Space-Time Paths</th>
<th>Storygraphs</th>
<th>Data Dials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependencies between Events</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attributes of Events</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>End-to-End Flow of a Case</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic Context</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change of Context</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation between Contexts</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison between Contexts</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

With the information in this matrix, we structure our discussion on the matching between
visualization techniques and geographic targets in event data.

4.3.2 Comparison of Targets

In summary, we have, per target, the following primitives that need to visualized:

<table>
<thead>
<tr>
<th></th>
<th>Events</th>
<th>Event Dependencies</th>
<th>Attributes of Events</th>
<th>End-to-End Flow of a Case</th>
<th>Geographic Context</th>
<th>Change of Context</th>
<th>Context Relation</th>
<th>Context Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic Step</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process Step [4.1.4]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic Step</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⇒ Process Step [4.2.1]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction between Areas [4.1.1]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Comparison between Areas [4.1.2]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>External Stakeholders [4.2.2]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Some targets can be seen as extensions or more complex cases of other geographic targets,
in particular in the targets regarding geographic and process steps. The second target is a
more complex case of the first target. This complexity arises from the process primitives
that need to be visualized in the former, while the latter can rely purely on its geographic
primitives to visualize geographic and process steps.
The target regarding the interaction between areas can also be seen as a more complex case of the target regarding geographic steps only happening with process steps. The added complexity again comes from the need to have access to process components across multiple geographic areas.

Apart from these cases, it is very hard to combine these targets into one visual object. A combination of targets in a single visual object would require an effective combination of primitives, without losing any of the needed insights that these primitives show. In the visualization techniques surveyed, this effective combination of primitives cannot be done.

What could be possible is a combination of a target that requires multiple visual objects and a relatively simple other target. Suppose we have a situation similar to the one in figure 39. The geographic visual object is relatively empty and can be used as a base for another target which only has needed process primitives that can be visualized as a geographic primitive. As the process visual object already shows all process primitives, this is not possible the other way around. One would need to take care though, that the base information of these adjusted visual objects does not become unreadable through the addition of more information.

4.3.3 Suggested Framework

Now that we have identified which visualizations are able to support which process and geographic primitives, we can more accurately look at what geographic targets of event data match with what kind of visualization techniques. We perform this matching per target.

Geographic Step \(\iff\) Process Step

Suggested Solution: Origin-Destination Flow Maps

The target in which a process step always correlates to a geographic step opens up a lot of possibilities due to the relatively small number of process primitives one needs to have visualized. While we still need to visualize dependency between events, this dependency of events is practically equal to the interaction between contexts. Therefore, we only need to consider the geographic primitives.

With these considerations and looking at the matrices above, a visualization technique like Origin-Destination Flow Maps seems ideal for the task at hand. Although many of the Origin-Destination visualization techniques support all necessary primitives, Origin-Destination Maps offer a relatively simple image compared to the other techniques as can be seen in the table with relatively few supported process primitives compared to the other Origin-Destination visualization techniques. Furthermore, Flow Maps can be extended upon if more primitives need to be visualized in more complex cases and do not require a specific temporal or spatial granularity.

Geographic Step \(\implies\) Process Step

Suggested Solution: Composite Visualization

The possibilities for complexity within this target are substantial and should be accounted for. Nearly all process primitives need to be visible together, while still being able to view the geographic context of a set of events and a possible change in geographic context.

Looking at the table above, attempting to show all this information in one visual object is not possible using the techniques surveyed. Composite visualization would be required to show the processes fitting this target.
**Interaction between Areas**  
*Suggested Solution: Flowstrates*  
Interaction between areas is hard to do due to its multivariate nature. Not only do we need to be able to see the kind of events that cause this interaction, but we also want to see the dependencies between these events. Within the geographic primitives, an analyst will be most interested in the geographic context and the relation between different parts of geographic context.

From the visualization techniques researched, only one technique seems to fit with the needed visualizable primitives. Flowstrates are able to handle the multivariate nature of the process primitives within this target, while being able to also show important geographic primitives and are not bound to specific temporal or spatial granularities. Furthermore, Flowstrates are able to handle other process and geographic primitives, which makes it a versatile choice as well.

**Comparison between areas**  
*Suggested Solution: Composite Visualization*  
Within this target, all process primitives need to be visible while still having a notion of geographic context. In this situation, there is a conflict in the use of position within the visualization as noted in Section 4.1.2. Either we prioritize the positioning of process steps and lessen the effect of the geographic context or we prioritize on geographic context and lessen the readability of the process. Both are unwanted and thus a separate visual object to compare areas on their processes seems to be the ideal solution. This is further enforced by the matrix above, as no single technique is able to show all necessary primitives.

**External Stakeholders**  
*Suggested Solution: Origin-Destination Flow Maps*  
The requirements for showing external stakeholders are relatively few. From the process primitives, we only require the attributes of events, while from the geographic primitives we require geographic context, relations between the contexts and the comparison between contexts.

This relatively simple case lends itself to a relatively simple solution. Using an Origin-Destination Flow Map to show the external stakeholders for different parts of the process, should adequately be able to solve the analysis questions for this geographic target. As with the first target, the lack of supported primitives of Origin-Destination Flow Maps is due to the simplicity of the resulting image and thus it seems a good fit for this relatively simple target.

**4.4 Conclusion**  
In Section 4.3 we define a framework with which one can solve problems in relation to event data with one of the geographic targets. This framework is based upon tests with hypothetical cases and the matching of these cases with the discussed visualization techniques. Through the framework, we offer a reasoning on what the current most suitable solution is for our pre-defined geographic targets. Within the framework, we do not have to explicitly handle temporal or spatial granularity, due to the advised techniques’ independence of these granularities. To test whether this framework offers useful answers, a test is needed with a real business case and data set.
5 Applying the Framework

In Chapter 3, a number of geographic targets in process mining data are presented and discussed, while in Chapter 4 we developed a framework to categorize solutions to the problems of these targets. In this chapter, we use the results of Chapter 4 to argument choices made in building a solution for a real-world use case.

The chapter is structured in the following manner. First of all, the use case which we use as a frame for our discussion is thoroughly described. This includes a description of the data, a look at the process behind the data and possible business questions an analyst may have. The next part contains a translation of these business questions into requirements which a possible solution should have. After the requirements are defined, these requirements are matched with the researched visualization techniques. This matching is based on the targets of Chapter 3, the fitting of the use case with the targets available, and the quality of the matching of this target with the techniques as discussed in Chapter 4.

In the conclusion of this chapter, a short summary of the chapter’s content is presented and a choice is made on which technique to base the final concept and build a prototype. Finally, we discuss how the choice described can be translated to a more generic setting.

5.1 Use Case Description

To frame the discussion on the choices that need to be made, a use case is defined on data made available by the Deutsche Bahn. The Deutsche Bahn is responsible for a large part of public transport in Germany. Their facilitations include trains, rental cars and rental bikes all over Germany. In this use case, a closer look is taken at data from their rental bike system. This system is known as "Call A Bike" and allows people in Germany to pick up bikes at certain pick-up points in Germany. The bikes can then be used until they are dropped off at a pick-up point. The system is available in a lot of large cities in Germany including Berlin, Frankfurt am Main, and Cologne.

5.1.1 Data Description

The data, made available by the Deutsche Bahn, was retrieved from their open data portal. To enable analysis the data of two tables is used. The first and largest table is a table containing all reservations, while the second table contains essential information on the position of pick-up points.

Bookings

The bookings table contains a total of 12,140,999 records with each record representing a reservation made on a bike by a user. The following fields of the table are considered the most relevant for analysis purposes:

- **BOOKING_HAL_ID**: Unique identifier for a booking.
- **VEHICLE_HAL_ID**: Unique identifier for the vehicle corresponding to the booking.
- **START_RENTAL_ZONE**: This column together with **START_RENTAL_ZONE_GROUP** contains the name of the pick-up point where the customer started the journey corresponding to this booking.
Service bookings are meant for a range of activities which include, but are not limited to, maintenance, movement, repairs, and cleaning. Other fields of the table contain, among other things, information on the bike used and the company which handled the booking. This information is at the moment not directly relevant within the scope of the research or the data behind the information is deemed too noisy.

**Rental Zones**
The rental zones table contains a total of 1,472 records with each record representing a pick-up point where users can pick up and drop off bikes. The following fields of the table are considered the most relevant for analysis purposes:

- **RENTAL_ZONE_HAL_ID**: Unique identifier for a pick-up point.
- **RENTAL_ZONE_GROUP**: Name of the corresponding pick-up point.
- **RENTAL_ZONE_X_COORDINATE**: Longitudinal coordinate of the corresponding pick-up point.
- **RENTAL_ZONE_Y_COORDINATE**: Latitudinal coordinate of the corresponding pick-up point.
- **CITY**: The name of the city of the corresponding pick-up point.

Other fields of the table contain, among other things, information about which company is responsible for the pick-up point and whether the pick-up point is near certain points of interest, such as airports or metro stations. As with the data from the bookings table, these fields were either not directly relevant within the scope of the research or the data behind the information is deemed too noisy.

**Preprocessing Steps**
To be able to extract useful information from the data, one needs to keep a number of factors in consideration. One very important factor is the quality of the data. As already mentioned, some fields of the data contain a lot of noise that has an effect on the analysis. An example of this can be found in the RENTAL_ZONE_HAL_ID column in the bookings table. With a
total of 17,488 unique values, it does not seem to exactly match with the 1,472 unique values in the RENTAL_ZONE_HAL_ID column in the rental zones table. This is partially due to some of the values being formatted wrong in the source, but also it seems that there are a lot of rental zones that have an ID but do not have a record in the Rental Zones table. Due to this shortcoming and after some testing, it seemed that joining both tables was better on the name of the rental zone as well as with the name of the city.

Another factor that impacts the retrieval of information from the data is the structure of the data. For some questions, the structure of the data does not directly allow for the discovery of an answer. One can find an example of this situation in the question of whether bikes are used to travel between cities. A record in the booking only contains the city of the start of the journey, while some of the destinations are possibly not located in the same city. An easy solution for this problem is to look for a pair of consecutive bookings for a certain bike which have a different city. On the assumption that the data is complete and no other bookings take place between the two, the journey between two cities can be constructed through that pair of bookings.

There are also cases in which the structure of the system behind the data negatively impacts the ease of access to information from the data. An example of this can be found in the naming of pick-up points. By picking generic names such as "Hauptbahnhof" (Central Station) and "Standort" (Site), names are not unique on a global level. The system also includes entries of pick-up points within a city that share names. These pick-up points are close to one another and their names logically reference a location at which they are placed, but this co-occurrence of names introduces ambiguity on where exactly bikes are stationed. In the analysis, these pick-up points are aggregated into one pick-up point with one name and the average coordinates of the corresponding pick-up points.

5.1.2 Use Case Questions

An analyst looking at the data from the Call A Bike system may have a number of questions to which they may want an answer. To accurately find a gap in what kind of questions can be answered using the current tools, one would first have to know what business questions may arise in the analysis of this data. The following questions may be questions which an analyst, seeing the data set, could have. The questions are based both on the interviews in Section 3.2 as on past experiences in process analysis and are also based on exploration of the data in its tabular form. A more thorough explanation of how these questions were conceived can be found in Appendix C.

1. Which city performs the most service cases from other cities and can we find reasons for the location of this city?

2. Can we detect special events happening in the usage of the Call A Bike system?

3. Are there areas which have large net in- or outflow and can we detect trends in these spots?

4. Can we find patterns in daily usage of bikes? What could be the behavior that causes these patterns?

5. Are there factors that influence when a bike needs to be serviced?
6. What does the life cycle of a bike look like?

7. What influences how long a service reservation takes?

8. Can we spot patterns in the appearance of reservations which end up in no-shows?

A closer look is needed for each possible question to see what information is needed to solve each question.

1. Which city performs the most service cases from other cities and can the location of this city be explained?
To answer this question, two pieces of information are needed. The city with the most service cases from other cities needs to be known and the location of this city relative to the cities from which it takes service cases needs to be known. To find the city with the most service cases, one needs to have a list of all service cases which change city and find the one that occurs the most. Once the city is known, the user should be able to see how the city itself is placed among cities which use the Call A Bike system.

2. Can we detect special events happening in the usage of the Call A Bike system?
Events usually happen in a certain period of time at a certain place. To obtain an answer to this question, a way is needed to find peaks at certain periods in certain areas.

3. Are there areas which have large net in- or outflow and can we detect trends in these spots?
Spotting outliers in terms of a large net in- or outflow of bikes relies of course on knowing the net-flow of bikes from and to a spot. To spot trends, location could be an important aspect, so we also need to know the spatial context of a point.

4. Can we find patterns in daily usage of bikes? What could be the behavior that causes these patterns?
Information regarding daily trends in the usage of bikes is needed to answer the first question. Explaining the behavior behind these trends can rely on a lot of factors, including spatial or case-specific factors. To analyze specific behavior, a user would need to have access to information regarding the factors that are most influential for that behavior.

5. Are there factors that influence when a bike needs to be serviced?
To find out if such factors exist, one would need to provide a view of what happens before a bike is reserved for a service booking. These factors could be a combination of all kinds of previous bookings.

6. What does the life cycle of a bike look like?
To know this, information is needed on the full history of a bike. To answer this question, one needs to know where the bike was at what time and for what use for the whole lifetime of the bike.

7. What influences how long a service reservation takes?
To get the answer to this question, a lot of factors of the service reservation should be known
and comparable. An issue here is that there is no indication of the type of service being performed for a certain booking in the data.

8. Can we spot patterns in the appearance of reservations which end up in no-shows?
One could look at this question as a filtered version of the question on patterns in daily usage and thus requires similar pieces of information.

With the information needed for each question known, a comparison needs to be made with what information we can retrieve from process models mined from the data. With the results of this comparison, we can see what information we need to show in the final solution.

5.1.3 Process aspect of the data

The business process behind the data described previously is one that seems very simple. Users make a booking, pick up a bike at a certain point and drop it off at a certain point. Each booking has its corresponding record in the booking table and for use in process mining one can look at each record as a pair of events; a departure and an arrival. Suppose one would use this definition without any further distinction on an activity level and use each trip as a case. The resulting model would be an extremely trivial one as pictured in figure 46. However, looking at the process in this manner, there is almost no information to be gained from the data and it does not provide a correct view on the complexity of the process itself. Expanding upon our notion of a case does not influence the complexity of the process as well. Regardless of whether we take a user or a bike as a case, the resulting model would be the model in figure 46 with an extra transition from Arrival to Departure. Changing our notion of an activity to that of a specific user reserving a bike would be an odd choice as well. The resulting model would be huge as a distinction has to be made for each customer and no information would be easily retrievable from the model.

One aspect that is lost in the model in figure 46 is the idea of service bookings, as there is no direct distinction between regular and service bookings. This could easily be solved by introducing separate activities for service bookings, but the previous model also does not have a direct notion of history. It is very likely that the history of a bike has an impact on what kind of reservation needs to take place. A bike needs to be maintained regularly, but heavy use may also speed up the need for maintenance. Therefore, extra insight can be gained by looking at the activities with bikes as a basis instead of a single reservation.

The resulting model would most likely be a simple cycle of a number of regular bookings, with an eventual service booking for service. However, the usage of a bike could be more accurately approached through a look at the trip times of a bike. Applying a good grouping to these times could allow for a good generalization as to not lose all details, but also not have an overwhelming number of possibilities.

![Figure 46](image-url)

**Figure 46:** The most simple look at the process behind the Call A Bike system is pictured above.
However, there is still one large factor that is not taken into account in the model that is essential in the answering of the questions set in Section 5.1.2. Location of departure and arrival is extremely important in finding some of the answers and this is where the complexity of the process is mostly present. Suppose we want a process model where a case corresponds to a bike, and a departure from a certain pick-up point corresponds to an activity, without even making the distinction between service and regular bookings, for a relatively small subset of the data. The model in figure 47 could be the result.

We can see some trends in the model, but overall this kind of model is already quite hard to analyze. This only becomes worse when attempting to analyze traffic in larger cities or close to impossible when trying to analyze multiple cities. Therefore, some kind of aggregation should be needed to allow for a usable view on the data. However, it is extremely hard to define an aggregation which is generic enough to work over different cities, while still maintaining enough detail to craft an accurate analysis.

Looking at the kinds of information we need to solve the questions set in Section 5.1.2, a process model is not able to provide all of the information needed. Looking at the first question, we could build a process model from the service reservations. We can use bikes as cases and the cities between which these bikes are moved as activities. An example of a possible simplified resulting model can be found in figure 48.

From this model, we notice a peak in certain cities. However, we have do not have the geographic context of this peak city available and thus we cannot directly conclude anything about the location of the city. This is a pattern which we can see in many of the questions that are influenced by spatial context.

Another instance of this can be found in trying to answer question 4. Suppose we specifically look at the city of Rüsselsheim to identify trends in traffic and take the model in figure 47 as a base analysis tool. An interesting pick-up point may be the Elisabethenstraße, due to its high usage. However, we’re missing two crucial pieces of information needed to analyze this high usage, temporal and spatial context. Through data exploration in the temporal dimension, we can find out that this pick-up point handles a lot of the traffic one would expect to happen at the Bahnhof pick-up point. There are high peaks in the morning of bikes leaving

\[\text{Figure 47:}\] The model resulting from only looking at consecutive departures in Rüsselsheim represents approximately 0.5% of the total data.
Figure 48: An example of what a model could look like which looks at all service bookings that move bikes between cities, is shown here.

All in all, it seems that a process model for this kind of situation seems non-ideal in attempting to handle the questions set in Section 5.1.2. Another kind of solution is needed to analyze these questions effectively.

5.2 Requirements of the Visualization

In Section 5.1.2, a number of business questions were posed which an analyst may have when looking at the data of the Call A Bike System. In this chapter, we set up a number of requirements for a solution that seeks to allow an analyst to answer the business questions. This is done by looking at what requirements a solution needs to fulfill to answer the questions posed before and comparing these with the visualizations studied in the previous chapters.

5.2.1 Solution Requirements

To find out what requirements the user has for a solution, we need to know what process and geographic primitives a possible solution needs to show. For the process primitives, we take a bike as a case and a pick-up or drop-off action at a certain location as an activity. The requirements will be built up, on a per business question basis.

1. Which city performs the most service cases from other cities and can we find reasons for the location of this city?

As finding which city performs the most service cases is possible using process models, no requirements need to be formulated on this piece of information. To consider reasons for the location of the city, we would need information about the relative position of the resulting city. So in the solution, one should not only be able to see the resulting city, but one should also be able to see the geographic context of the resulting city. Ideally, we should be able to retrieve both pieces of information in one view and compare both the flow between cities
as the location of the city with the most prominent flow. This means we need a view on the dependencies between activities, geographic context, change of context, relation between contexts and comparison between contexts.

2. Can we detect special events happening in the usage of the Call A Bike system?
A solution to solve this question must provide a sense of spatial context as to allow the user to find areas involved in these events. Ideally, a user should again be able to see both the temporal and spatial aspect on different scales to identify specific events happening. This translates to being able to see geographic context and comparison between contexts.

3. Are there areas which have large net in- or outflow and can we detect trends in these spots?
Finding what points have a large net in- or outflow and finding spatial trends in these points should be possible. Ideally, these two pieces of information are combined into one view. In terms of primitives, we need to see attributes of events, geographic context, and comparison between contexts.

4. Can we find patterns in daily usage of bikes? What could be the behavior that causes these patterns?
As noted previously, with process models we can only spot whether there is a high amount of traffic at a certain point. To perform analysis on the temporal and spatial dimension of this data, we need these both visible. This translates to being able to see geographic context and comparison between contexts.

5. Are there factors that influence when a bike needs to be serviced?
To solve this question some form of data mining or statistics are required. This is out of the scope of geographic or process primitives.

6. What does the life cycle of a bike look like?
To solve this question, two approaches can be taken. A process model to describe what happened exactly with the bike at what time, however, we again lose all spatial context with that approach. A geographic view to show where it was at what time, could allow for more accurate analysis on this aspect. Ideally, we would have both these pieces of information combined. This means we need a view on dependencies between activities, end-to-end flow of a case, geographic context, and the change of context over time.

7. What influences how long a service reservation takes?
The most important shortcoming in solving this question seems to be on the side of the data. We miss a lot of information regarding the nature of a service reservation, while this is most likely a very significant factor in how long a service reservation takes. A possible solution cannot account for missing data.

8. Can we spot patterns in the appearance of reservations which end up in no-shows?
The same requirement as with the question on daily patterns appears here. A possible solution needs to have a way to analyze spatial context of pick-up points.
To summarize, we define the following requirements of our solution with the associated question between brackets:

1. The solution allows for a comparison of the location of cities in the data. (Q1)
2. The solution allows for a comparison between the flows between cities. (Q1)
3. The solution allows for a view with both statistics per city and the spatial context of a city. (Q1)
4. The solution allows for a view to compare periods in time. (Q2)
5. The solution allows for different scales of time to be used. (Q2)
6. The solution allows for a view on the spatial context of a point. (Q2)
7. The solution offers different spatial scales for analysis. (Q2)
8. The solution allows for a view on points with a certain statistic and their spatial context. (Q3)
9. The solution offers a view which allows comparison of trends over time and between pairs of points. (Q4/Q8)
10. The solution allows for statistical analysis or the use of data mining techniques. (Q5)
11. The solution offers a view on the history of an object in its location over time for its complete lifetime. (Q6)

5.3 Visualization Techniques and their Suitability

In this chapter, the visualization techniques introduced in Section 2.3 and analyzed in Chapter 4 are combined with the requirements obtained in Section 5.2 to see what kind of technique is best suited for a possible solution. First of all, we look at the result of a direct application of the framework on the requirements as set above. Then, a mapping from a technique to a number of satisfiable requirements is made and an effort is also made to compare that to a process model that is projected onto a map. Finally, the limitations of the techniques are discussed by looking at the requirements that they are not able to fulfill.

5.3.1 Applying the Framework

If we compare the requirements to the targets to which they correspond the most, we get the following matrix:

<table>
<thead>
<tr>
<th>Target \ Requirement</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
<th>9.</th>
<th>10.</th>
<th>11.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Geographic Step ↔ Process Step</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Geographic Step → Process Step</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Interaction between Areas</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Comparison between Areas</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. External Stakeholders</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What can be noted is that some of the requirements can be seen as part of all targets, while others are part of none. The requirements which are linked to all targets share that they are about comparing positions or basic statistics of these positions. This kind of spatial comparison can be seen as a part of each target as analysis within each target links position to a certain part of the process. The requirements which are not linked to any targets do so, because they ask for different scales of time and space. Because the framework does not differ based on spatial or temporal scale, there is no link with the requirements regarding these scales.

From this comparison, we can conclude that most usable requirements correspond to questions in the third target regarding interaction between areas. However, it should be noted that one important facet of the requirements is not represented within the targets. Some of the requirements state a need to compare certain statistics over time, while in the construction of the framework there is not such a notion. A more fine-grained comparison of the available techniques and the solution requirements is warranted to make sure that we make the correct choice of technique for this problem.

5.3.2 Usability of Techniques

Origin-Destination Maps (Chapter 2.3.1)
Origin-Destination Maps work ideally when comparing areas with a common size and shape. Although cities are not all of the same shape and size, Origin-Destination Maps should still be able to handle them and thus satisfy requirement 1, 2 and 3. Using small multiples to display different periods in time, Origin-Destination Maps are able to offer a view to compare these periods and thus satisfy requirement 4. The periods which need to compared can be changed in scale, so requirement 5 can also be fulfilled. Other requirements are not satisfiable through the use of Origin-Destination Maps.

Origin-Destination Flow Maps (Chapter 2.3.2)
Origin-Destination Flow Maps are able to handle show both areas and locations so requirements 1, 2, 6 and 7 can be met using the technique. By using small multiples comparison over different periods of time is possible and as stated before these small multiples are not bound by a certain scale. This allows for requirement 4 and 5 to be met. Other requirements are not satisfiable through the use of Origin-Destination Flow Maps.

Maptrix (Chapter 2.3.3)
Through the ability to handle both points and areas, Maptrices are able to handle requirements 1, 2, 6 and 7. By having statistics available per city and a viewable spatial context of a city, requirement 3 can be met as well. This extends to the points as well, so requirement 8 is also met in the use of Maptrices. Although in theory, small multiples are possible in a Maptrix, using small multiples on a complex Maptrix in practice will display too large an amount of information to be comparable. This means that comparison between periods in time and thus the associated requirements are not directly possible. One Maptrix is not bound to a certain scale of periods in time, so multiple scales can be used as demanded in requirement 5. Other requirements are not satisfiable through the use of Maptrices.

Flowstrates (Chapter 2.3.4)
Sharing characteristics on how they handle areas, points and associated statistics with Map-
trices, Flowstrates are able to meet requirements 1, 2, 3, 6, 7 and 8. Through the explicit temporal dimension present in Flowstrates, they are able to allow for comparison between periods of time. These periods can be configured, so Flowstrates are able to handle requirement 4 and 5. By allowing a view on both the temporal aspect and spatial context, Flowstrates can also satisfy requirement 9. The other requirements are not satisfiable through the use of Flowstrates.

Flow Trees (Chapter 2.3.5)
Flow Trees allow for comparison of cities in the data, so requirement 1 can be met. However, Flow Trees are not ideally suited for comparison of points, so it has trouble with requirement 6 and 7. Flow Trees allow for comparison on a temporal scale in the same way as Origin-Destination Flow Maps through the use of small multiples. Therefore, requirement 4 and 5 can be met. Other requirements are not satisfiable through the use of Flow Trees.

Space-Time Paths (Chapter 2.3.6)
Space-Time Paths have a configurable z-axis and thus allow for different scales of time to be used in the visualization and with that requirement 5. They also allow for a view on the spatial context of a point through the map plotted at the base of the Space-Time Path and this map is configurable in scale. This enables the fulfillment of requirements 6 and 7. In principle, the important points of the data can be shown on the map at the base of the Space-Time Path and be made to show certain statistics. This allows Space-Time paths to fulfill requirement 8. Furthermore, Space-Time Paths allow for a view on the history of an object in space and time and thus are able to manage requirement 11. Other requirements are not satisfiable through the use of Space-Time Paths.

Storygraphs (Chapter 2.3.7)
Storygraphs share many possibilities with Space-Time paths in that they handle time in a similar way and allow for the viewing of paths. This enables requirements 5 and 11. As with Space-Time Paths, Storygraphs do not have a strict spatial scale on how they work and with that are able to fulfill requirement 7. Using color in the lines of a Storygraph could allow for some statistic to be shown and thus requirement 8 to be met. Other requirements are not satisfiable through the use of Storygraphs.

Data Dials (Chapter 2.3.8)
Data Dials are meant to work with points and allow for comparison between these points. This enables requirements 6 and 8. Due to their ability to show trends over time of a point, they are also able to handle requirements 4 and 9. Other requirements are not met through the use of Data Dials.

Process Model with Map-Projected Activities
To have a look at the possibility of projecting process models onto the map, we include this option into the analysis. These process models are similar in look and function to Origin-Destination Flow Maps with the added function that they have the names of activities projected onto their location. These models can handle both areas and points, but they cannot be too close to each other as this may cause too much activity squares to appear and obstruct the geographic context. In the further analysis of requirements, we assume this is not a problem.
Through their similarity to Origin-Destination Flow Maps, they are able to handle the same requirements. By taking activities as cities, they are able to show flow between the cities and due to being projected on a map, there is the possibility to compare the location of cities. The models are not bound by geographical scale, so requirement 6 and 7 are also viable. Small multiples could again be an option to handle requirement 4 and 5.

Summary
The table below shows a summary of the discussion in this section.

<table>
<thead>
<tr>
<th>Satisfiable Requirements</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin-Destination Maps</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Origin-Destination Flow Maps</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maptrix</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flowstrates</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flow Trees</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space-Time Paths</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storygraphs</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Dials</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projected Process Models</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.3.3 Common Limitations of Discussed Techniques

A number of patterns can be discovered in the satisfiability of requirements by the visualization techniques. The most obvious one is that there is no direct solution for the data mining or statistical analysis requirements of requirement 10. This is due to data mining techniques not being in the scope of this research and thus they will not be part of a solution.

Another pattern which can be spotted is the inability of most Origin-Destination visualization techniques to handle the history of an object in requirement 11. Space-Time Paths and Storygraphs are the only visualizations that can handle that kind of information, due to the nature of their design. One of the goals of Space-Time paths and Storygraphs is to show exact position in space-time of a set of objects, while Origin-Destination visualizations focus more on providing an overview on between what pairs of locations this movement takes place. In theory, one could split up the journey of one object in space and time into pieces and use that as a base for an Origin-Destination visualization, however, in most cases, this is not ideal. In some visualizations, such as Origin-Destination Flow Maps and Flow Trees, one could plot these shorter journeys. However, this could cause confusion when cycles happen or paths cross as can be seen in figure 49. Flowstrates do not immediately have a problem with cycles in a path due to the explicit temporal aspect being present in the table. However, Flowstrates still have the same problem while dealing with crossing paths as can be seen in figure 50. While showing a single path using Flowstrates is theoretically possible, the resulting table would contain a lot of unnecessary white space and a user will not be able to easily estimate distances that are traversed throughout the path.

Another requirement that is not met often is requirement 9; not many techniques allow for a comparison over time and between points. Although in theory small multiples should allow for this kind of comparison, when comparing trends over time, the user would have to compare slight variations in color with a lot of space between them. This task requires a lot of attention of the user and this amount of required attention increases the difficulty of the
5.3.4 Choosing the Visualization

The results from Section 5.3 offer a clear view on what the best technique for this use case should be. Flowstrates, although not able to fulfill all requirements, offers the most wanted functionality in this use case. Although the framework does not feature all needed primitives for this particular use case, when looking at the geographic targets this use case corresponds to the most and the tables in section 4.3.1 and section 4.3.2 we can reach a similar conclusion. From the matrix in section 5.3.1 and the framework in section 4.3.3 we know that this case has the most ties to the third target and thus that Flowstrates is the recommended technique.

5.4 Extending the Use Case

As already mentioned in Section 5.1 the business process without geographic information behind the Call A Bike system is not a very complicated one. The question then arises to what degree the conclusion for this use case is usable in more classical and complex business processes that do contain geographic data. To answer this question, three subjects need to be discussed:

- What aspects of the use case are representative of classical business processes?
- What aspects of the use case are not representative of classical business processes?
- How could the chosen visualization technique be used with the aspects that are not representative?
Representative Aspects
Classical business processes consist of activities, cases, and events. The use case at hand can be translated to fit this paradigm. Activities in the use case can be defined as being either picking up or dropping off a bike at a pick-up point. Cases in this context can have multiple definitions. One can look at the business process from a bike’s perspective and take each bike as a case, or one can look at the process from a user’s perspective and see each user as a case. If needed, one could also choose to define each booking as a case. Looking at the process from a bike’s perspective allows for analysis on the life cycles and usage of bikes, while looking at the process from a user’s perspective allows for analysis on specific users and behavior of these users.

The events in the use case are clear as well; an event is simply a user of the Call a Bike system picking up or dropping off a bike at one of the pick-up points at a certain time. An important aspect of analyzing business processes is that the timing of these events can be a critical part of the analysis. This aspect can also be seen in the use case in the form of some of the business questions proposed earlier, for instance, question 6 or 8.

Non-Representative Aspects
When looking at events in the use case and events in classical business processes, we can notice a number of differences. One of these differences is that for each event in the use case, exactly one other event is relevant for most of the analysis. For pick-up events, this is the corresponding drop-off event and vice-versa. Regardless of our definition of cases, events before a pick-up event or after a drop-off event are not relevant to the information to be visualized in almost all factors. This results in almost all handled cases having a trace with a length of 2. In a more classical business process, the events before or after a certain event or set of events could provide a lot of insight into the choices that make up a process.

Another aspect of the events studied in the use case is that they are very limited in terms of variety in comparison to classical business processes. In the use case, events can only be a pick-up or drop-off event. This limited range of events is easily displayable in a visualization due to a point being designated as an origin or a destination. The range of activities of regular business processes can be much larger.

In the visualization which we chose as a solution to the use case, the location of an event is one of the only aspects of the data that is used for analysis. In other use cases or with other data sets, there could be requirements on the visualization that stipulate that other aspects of the data should be shown.

Extensions and Limitations of the Chosen Visualization
The lack of a view on the full flow of cases through the data is hard to solve using Flowstrates or any visualization that relies on spatial information. As discussed in Section 4.1.2, displaying the flow of cases through data would either negatively impact the power of the geographic visualization or the readability of the model. A view using composite visualization to compare the flow of cases between certain spatial clusters as proposed in Section 4.1.2 could be possible, but Flowstrates would not be the ideal technique to show this connection.

Although there is not as much range in activities as one would expect from regular business processes, from the example in Section 4.1.1 we can see that Flowstrates are able to handle a slightly larger complexity in term of activities. By replacing the temporal dimension usually present in Flowstrates, by the activity which causes the change in location, we are able to show different activities within a Flowstrate. This is, however, limited in terms of
activity complexity and results in the loss of temporal information. In alternative use cases, a consideration should be made on what kind of information is deemed most important to the user.

This ability of Flowstrates to handle more activities also translates to an ability to handle a larger diversity in attributes. One can exchange the temporal information supplied by the table in a Flowstrate for any categorical value to analyze the flow between locations on the attribute. To provide an example in the use case, one could decide to compare the number of service reservations and regular reservations between pairs of pick-up points instead of comparing the number of reservations between pairs of pick-up points over different years.

5.5 Conclusion

From the discussion in this chapter, we can conclude that Flowstrates is the technique with which an analyst should try to solve the analysis questions set. This is the result from reasoning from both the framework from Section 4.3.3 and a more fine-grained analysis of requirements. However, we also conclude that the current state of the framework is not ideal yet. It seems from the comparison in Section 5.3.1 that we have not yet captured the full scope of all possible analysis cases with geographic information in event data within this initial definition. For this case however, we have a technique with which we can design a prototype to solve the analysis problems set before.
6 Implementation of the Prototype

In this chapter, we provide a description of the implementation of a prototype of the solution reasoned about in chapter 5. The chapter is structured into three parts, a short summary of the implementation details, a list of design choices with the accompanying reasoning behind these choices and a guide on how to create new dashboards with the prototype.

6.1 Implementation Details

The prototype was developed within the ProcessGold \cite{23} platform, by combining some of the functionality already available in the platform and adding extra features where needed. The ProcessGold platform combines business intelligence with process mining to deliver a complete process analysis solution. The platform is built using a C++ back-end and with a Typescript-based front-end. With the platform, informative dashboards can be built based on multiple, so-called, views. In this context, views are the building blocks of the ProcessGold platform and the different types of views provide different insights into the available data. By using three views, one can already achieve the basic look of a Flowstrate as can be seen in figure 51.

Two types of views are visible in figure 51; a geographic view and a table view. The geographic view uses Leaflet \cite{17} as a base. Leaflet is a Javascript library which allows for interactive use of maps within your browser. The basic geographic view has several required and optional variables. It requires at least a table from which to extract data and the two columns which per record contain respectively the longitude and latitude of a point which needs to be shown on the map. Other variables, which can be added, are, for instance, a filter on the data for the current view and the size and color of the points to be visualized.

Table views are built up through a hierarchical tree structure. Each variable on which the data is aggregated can be seen as a level within this tree. The tables used within Flowstrates

![Image](image_url)

**Figure 51:** By using two geo-views and a specifically constructed table-view in the middle, the basic look of a Flowstrate can already be achieved.
require at least 4 variables to be set correctly. We need a table with the needed data and 3 aggregation levels. These levels correspond to, respectively, origin point, destination point, and temporal aggregation. These levels are made visible in figure 52 through the use of rectangles. Each rectangle indicates a level within the tree. In the example shown in figure 52, the origin city, Berlin, is the first level. The destination cities, Frankfurt, Hamburg, Kiel, Rostock and Hamburg, are the second levels of the tree and the quarters of the year are the third levels. Within each quarter is a set of records belonging specifically to that combination of origin, destination, and quarter and these records are visualized according to a predefined metric. In most cases, we use the number of records as this metric.

The platform itself does not allow for the sharing of information between different views, except through the manipulation of data underlying these views. This data manipulation can, for instance, be done through the use of filters. For the prototype, the geographic views need to have information on the state of the table view to be able to draw lines between the entries in the table and the points on the map. That is why with each re-rendering of the table, we send some information regarding the state of the table to the geographic view as well. This information is comprised of a set of items about either a first or a second level node with, per item, the textual value within a node (e.g. Berlin or Hamburg), the tree level and the height of a node within the rendered table.

With the position of the view and the height of related entries in the table known within the geographic view, we can use Leaflet functionality to draw a line between the point on the map and the corresponding entry in the table. An example of a resulting view can be seen in figure 53.

### 6.2 Design Choices

During the implementation of the prototype, there were some choices made, which do not conform to the implementation of Flowstrates as described by Boyandin et al. In this section, we explain some of the reasoning behind these choices.

- The colors of the heatmap in the table of the Flowstrate are normalized on the maximum flow of visible origin, destination, and temporal aggregation combinations. This allows for a user to have a constant sense on the combination with the highest amount of flow in the current selection.

- When the user hovers over a point on the map, the line connecting the point and the table is highlighted. The decision was made to not highlight the line when hovering the line itself. This decision is due to tests done in developing the prototype. When one
would have multiple lines crossing over a circle, it became hard to show a hint when hovering over the circle. In the end, the ability to hover over a point was deemed more important to the user than the ability to hover over a line.

- The option was considered to sync the zooming and panning of both maps. This option was deemed not desirable due to this option disabling the freedom of the user in choosing what they want to filter on. Furthermore, syncing zooming and panning on the maps is not directly expected by most users.
- When a user hovers over an entry in the table at the left side of the table, all corresponding entries at the right of the table are highlighted as well. This provides a user with an easier overview on the item in which they are interested.

6.3 Creating a new Dashboard

To create a new dashboard on a certain set of data in the prototype, one needs to perform the following steps: These options are all available as a Designer user of the ProcessGold platform. Designers are usually product engineers that have a high understanding of the structure of the tooling and do not necessarily perform the analysis. The dashboards created for the analysis in Chapter 7 were created in a similar manner.

1. Create three views, a table view, and two geographic views. Place the table between the two geographic views.

2. Select the correct variables for the table levels. The information that needs to be shown on the left map should be the first level, while the information that needs to be shown on the right map should be the second level. Depending on the needs of the user, a third level can be added on which the user wants to compare the different flows. In most previous examples, such as in figure 51 and figure 52, this was an aggregation of time.

3. Select the correct variables for the geographic views. The user should select the correct longitude and latitude variable for the points which need to be visualized and set the correct linked attributes for each geographic view. Set the hint of a point to be the value which will also be an entry in the table.

4. Set the appropriate filters for each view based on the information the user wants to display. These filters need to be the same to make sure that there is no disconnect in what can be seen in the different views.

5. The dashboard is now usable, but any additional options can still be set according to the wishes of the user. These options could, for instance, be point size on the geographic view or the size of a row in the table view.

6.4 Conclusion

In this chapter, we have described part of the process of implementing our choice from Chapter 5 to a usable prototype in the ProcessGold platform. The platform was adapted to allow for the creation of Flowstrate dashboards. Dashboards were also created in the prototype to be able to analyze the data and see if the business questions set in Chapter 5.1.2 can be answered.
7 Evaluation of the Prototype

In this chapter, we look at how the developed prototype from Chapter 6 performs in answering the questions set out in Chapter 5. We test how it performs in two parts. We look into answering the business questions ourselves and provide a step-by-step guide on how these answers are reached. Next, a number of user tests are done to see if users are able to come to the same conclusions and in what manner they do so.

7.1 Analysis of the Business Questions

To have an inside look into how a user can use the prototype to answer some of the business questions, we analyze the data ourselves in this section. Specifically, we look for an answer to the first 4 questions, which each highlight a different aspect of the data or the prototype. As a Designer in the ProcessGold platform, we are able to create dashboards, which regular analysts are not able to do. In an ideal implementation, the options available to a Designer should also be available in a simpler format to an analyst.

7.1.1 Q1: Finding a Service Hub

There are two things we need to look at to find a service hub city in the data available between cities. First of all, we need to look at the number of service bookings from a city and we need to determine whether this number is not skewed due to large numbers of regular reservations. To achieve this, we need two different views on the data; one of all service traffic between cities and one with the ratio of service reservations to regular reservations.

![Diagram of service reservations](image)

Figure 53: The total amount of service reservations is made available using a Flowstrate-based approach.
Figure 54: The ratio of service reservations to regular reservations is pictured here.

Figure 53 shows a screenshot of the result of one of these views. It is clear that Frankfurt am Main handles a large amount of service traffic and sends a large number of bikes to other cities during service bookings. From this dashboard itself, however, it is not directly clear if this number is representative. The same view also shows a large peak of service reservations from Berlin to Hamburg in the first quarter of the year and regular service reservations in other quarters. To determine whether these numbers are not just outliers due to large numbers of regular reservations, the view shown in figure 54 can be used.

This view shows us that Frankfurt am Main not only handles a lot of service reservations, but also that the ratio of service reservations to regular reservations is quite high. From this, we can conclude that Frankfurt am Main is a service hub for the Call A Bike program of the Deutsche Bahn. Berlin to Hamburg also remains interesting, looking at the ratio of reservations, but this one route seems to be a kind of outlier. Looking at the images in figure 55 and figure 56, we see the aforementioned large peak in the first quarter, but relatively little traffic for the other destinations and quarters.

Figure 55: The traffic between Berlin and Hamburg seems to be especially heavy in Q1. Berlin has no further notice of heavy service traffic.

Figure 56: Hamburg itself also does not show a large number of service reservations happening.
Through contact with the Deutsche Bahn, the conclusion we reached through the analysis for this question, is confirmed. The outlier of Berlin to Hamburg was also explained through this contact with Hamburg being confirmed as a secondary service hub.

7.1.2 Q2: Spotting Special Events

To see if it is possible to detect special events, we want to use a large event with a known specific location and period. One event that fits this idea is Oktoberfest in Munich. Oktoberfest is a large festival known to attract millions of people to Munich during late September and early October. Oktoberfest takes place on the Theresienwiese as can be seen in figure 57. In the data we have available on Munich, Oktoberfest takes place in week 38, 39 and 40.

From the information gathered here, we would expect to have a peak in usage around the Theresienwiese in these weeks. The closest Call A Bike pick-up point available in the data is the point at the Goetheplatz. An initial look at the usage of the Goetheplatz in the neighborhood of Oktoberfest is shown in the image in figure 58. Although there is a small peak in weeks 39 and 40, this peak is not very visible due to a lot of usage at another route. Further analysis of the data reveals that this point is the pick-up point known as Münchner Freiheit and this pick-up point has a constant trend of high use regardless of the week that is chosen. By looking at the hourly traffic at the Münchner Freiheit, we may be able to determine what kind of traffic this is. Figure 59 shows the traffic by hour of, among others, the Münchner Freiheit. The hourly usage shows a large peak in the evening with the most bikes leaving between 19:00 and 20:00. This traffic is unusual in the sense that the most pick-up points with heavy usage have this in a manner that resembles a work or study cycle. Reasons for this kind of traffic are not immediately clear, but two factors could come into play. It could be that the Münchner Freiheit pick-up point is near a popular nightlife or restaurant area which could explain a peak at the time seen in the data. The overall heavy usage within Munich of the Münchner Freiheit can also be attributed to the nearby English Gardens which are a popular tourist attraction on foot as well as on bike. For continued analysis, it seems that the Münchner Freiheit pick-up points can be treated as an outlier and in the further analysis of this question, bookings with the Münchner Freiheit pick-up point as origin are filtered out of the data.

![Figure 57](image-url)

**Figure 57:** Oktoberfest takes place on the Theresienwiese during weeks 38, 39 and 40 in the data available. The Theresienwiese is marked in blue.
Figure 58: A look at the usage around Oktoberfest at the Goetheplatz does not seem as clear as initially expected.

Figure 59: Hourly traffic at the Münchner Freiheit shows an irregular usage pattern.
Figure 60: The weekly traffic around the Goetheplatz now clearly shows the relative peak of usage around Oktoberfest.

Using this filtering, we get a much clearer view of the situation around Oktoberfest at the Goetheplatz as can be seen in figure 60. Apart from the Goetheplatz, there is also a large peak around these weeks at the Stiglmaierplatz as can be seen in figure 61. This peak seems a bit odd at first, because the Stiglmaierplatz is not as close to the Theresienwiese as the Goetheplatz, but still seems to be affected by Oktoberfest. One explanation of this behavior could be that the Stiglmaierplatz is used as a transit point from the metro or tram to a bike to get to the Theresienwiese as both metro and train stations are nearby the Stiglmaierplatz pick-up point. All in all, one can clearly see two peaks of usage in the neighborhood of the Theresienwiese in the pick-up points of the Goethplatz and the Stiglmaierplatz around Oktoberfest and, therefore, we can conclude that we can spot this event happening in the data.

7.1.3 Q3: Spotting Large Net In- or Outflow

This analysis required for this question is not ideally performed with a classical Flowstrate, so some creativity is needed in how destination is interpreted and how the coloring is applied. As an example city for this question, we take Berlin. Instead of looking at separate pick-up points as destinations, we take the city of Berlin as the destination point in the Flowstrate and we adjust the coloring in the table to reflect the netflow of a point with a red-blue heat map. Red represents a net outflow, while blue represents a new inflow. The result of these adaptations can be found in figure 62.

What can be seen in the image of figure 62 is that there are both points that experience a large net outflow as points that have a large net inflow. What cannot be directly seen
Figure 61: The weekly traffic around the Stiglmaierplatz shows a clearer peak around Oktoberfest as well.

Figure 62: An adapted Flowstrate can be used to gain more information on the net flow of bikes within a pick-up point. This tells us more about which pick-up points over time lose or gain bikes.
Figure 63: From looking at the hourly departures in Kiel, one can notice a students schedule. from the image, but through interaction with the tool, is that a lot of the points that have a large net outflow are on the outskirts of Berlin, while many of the points that have a large net inflow are near the center. Many pick-up points near train or tram stations have a large amount of net inflow, which seems to indicate that people take bikes from near the outskirts to head to stations and continue their journey there.

One question popped up in the analysis for this question. Overall a lot of pick-up points seem to lose a lot of bikes and this is a trend that is shared between a lot of cities. This can be explained by looking more thoroughly at the data. It seems that a relatively large number of reservations have no endpoint in the data. By having to calculate netflow based on the number of bookings that have a certain pick-up point as a destination and the number of bookings that have that point as an origin, these results may not be representative of reality.

7.1.4 Q4: Identifying Daily Trends

To see if we can find daily trends, we chose three cities that differ in scale in terms of the number of bike reservations, Kiel, Rüsselsheim, and Stuttgart. Kiel was chosen as it was a city with a relatively small amount of reservations on which it was relatively easy to develop. Although Kiel only has 1,724 reservations in the data, patterns that reflect its student population were still spotted during development. This combination makes it an ideal initial analysis case. The resulting image of all reservations in Kiel can be seen in figure 63.

In figure 63 we notice two peaks one from Kiel Hauptbahnhof to Campus 4 / Audimax and one from Campus 4 / Audimax to Kiel Hauptbahnhof. These peaks seem to indicate students traveling to university in the morning and returning to the station in the afternoon.
When looking for a larger data set that should mirror some of the behavior, we identified in Kiel, we decided to look at Rüsselsheim. Rüsselsheim is also known as a city with a large student population and, with 65,339 reservations, has a lot more data than Kiel. Looking at the data available of Rüsselsheim, we see a similar pattern emerging to the pattern in Kiel. Figure 64 shows the traffic aggregated by hour in the city of Rüsselsheim. Once again, we can see a large peak in the morning and a slightly more spread out one later in the day. There are some differences with Kiel though. The peak in the morning is not from the point called Bahnhof, but from a pick-up point called Elisabethenstraße Rüsselsheim. An explanation for this difference can be found in the location of the pick-up point. The Elisabethenstraße pick-up point is just across from the Bahnhof pick-up point and thus can be used as an alternative to the pick-up point at the station. Another difference in the usage in Rüsselsheim is that the peak in the afternoon happens a lot earlier than in Kiel. This could be an indication of a different set of class times at the University in Rüsselsheim than in Kiel.

To see if we can identify similar behavior in larger cities in the data set, we decided to use Stuttgart. Stuttgart has 461,370 reservations and showed some unexpected behavior. After isolating the largest peaks in usage in Stuttgart, we get the image in figure 65.

What is unusual about this behavior is that the pick-up points with the highest amount of hourly usage are far from the city center of Stuttgart. The Bad Canstatt Bahnhof and the Mercedes-Benz-Museum pick-up points have the largest peak in hourly usage and are located on the north-eastern edge of Stuttgart. An explanation for this specific trend in Stuttgart could be the Mercedes-Benz factory in the neighborhood of the Mercedes-Benz-Museum pick-up point. The people that go to work there by bike could explain the large amount of traffic at these specific points in time.
7.2 User Tests

To evaluate the usability of the prototype, we have run several user tests with users which all have some experience in handling data analysis and have differing degrees of familiarity with process mining. The structure and results of these user tests are discussed in this section.

7.2.1 Structure of the User Tests

To test whether users are able to come to the same conclusions as we have done in Section 7.1, we provide them with the same questions and the dashboards that were created for the analysis. At the start of the test, we first provide a short introduction on what kind of data they will be analyzing and a short explanation on how the dashboards work using a sample of the data. For each question, the user is directed to the menu or dashboard which they need to analyze the data in regards to that question and are encouraged to voice their thought process in answering the questions. After each question, we compare the solutions of the user and our solutions from section 7.1 and discuss any differences.

After the last question has been answered, a number of evaluation questions are asked to the user to find out how the user experienced his or her time with the prototype. The following questions are asked:

- Did you think the tooling was adequately able to help you answer the questions?
- Was there information missing which you needed to solve certain questions?
What do you think of the amount of information being shown, is it too much, too little or enough?

What do you think could improve the tool in terms of usability?

Do you think this tool fits in the field of process mining and if so, in what cases?

7.2.2 User Analyses

The full notes taken during the user tests can be found in Appendix B.

Q1: Finding a Service Hub

From the user tests, there seems to be a consensus that this is the hardest analysis question. Many had trouble directly understanding what was shown in the dashboard and how to connect this kind of information to the service process behind the data. Regardless of this difficulty in reaching an answer, many users were able to find Frankfurt am Main as the main service hub of the Deutsche Bahn. Many of the users saw Berlin as a potential service hub as well, but that can be explained by the relatively large amount of traffic between Hamburg and Berlin. Several users first thought of cities like Marburg and Darmstadt as likely candidates for another service hub, but reasoned, through the use of the location of these cities and the location of Frankfurt am Main, that service hubs should not be located this close to each other.

Q2: Spotting Special Events

To make sure users would be able to find the same answers as we found in the analysis, we filtered out the reservations which have their origin in the Münchner Freiheit pick-up point. This seems sensible as the user in these tests is not as aware of the spatial context of Oktoberfest and thus would probably not look further than the high amount of usage in the Münchner Freiheit pick-up point.

With Münchner Freiheit removed from the data, many users were able to identify both the Goetheplatz as the Stiglmaierplatz as potential candidates for being close to the actual location of Oktoberfest. Several users first filtered out all data not belonging to the weeks with Oktoberfest and continued their search from there. One user also noticed a high amount of traffic at the Odeonsplatz, but saw that this traffic was more constant than the traffic at the other candidates and decided to go for the Goethe- and Stiglmaierplatz instead.

Q3: Spotting Large Net In- or Outflow

Many users were able to find spots which showed a large net in- or outflow based on the table. After a couple of tests, it was clear that the view itself did not lend itself well to finding spatial patterns. A small adaptation was made to allow users easier comparison between points only based on the map. An image of the result can be found in figure 66.

With this adaptation, users were more quickly able to figure out the pattern that a lot of the points that have a net inflow are located in the city center and that a lot of points that have a net outflow are located on the outskirts of the city. Some users were also able to find the pattern that a lot of the pick-up points in Berlin have a negative net flow.

Q4: Identifying Daily Trends

In the tests, we asked the users to specifically look at the daily trends in Rüsselsheim as this shows a clear pattern, albeit not standard, due to the peak being at the Elisabethenstraße and...
not at a Bahnhof pick-up point. This non-standard pattern could invite the user to deeper analysis on what each station was and what kind of behavior defines the trends in traffic at Rüsselsheim.

The users were quick to identify the peak from the Elisabethenstraße to one of the Campus pick-up points early in the morning. They were also quick in identifying the peak back from Campus am Bruckweg to the Elisabethenstraße around 13 o’clock. Many wondered as well why the peak in the afternoon was that early. Some users first reasoned it had to with people getting lunch, but then decided against it due to no return journey being visible in the data. Users also spotted another possible trend not immediately noticed during our own analysis between the two Campus pick-up points. From the Campus am Bruckweg to the Campus Opel-Altwerk pick-up point and back slight peaks every two hours can be noticed during class times. This could indicate students moving to and from certain classes.

7.2.3 User Evaluation

In this section, we summarize all answers given by the users to questions regarding the prototype and its chosen visualization technique. These answers both contain comments on the implementation of the prototype and on Flowstrates as a technique.

Did you think the tooling was adequately able to help you answer the questions? Overall, the consensus on this question seems to be that the tooling was adequate to answer the questions, although some users missed some aspects of the data. Many had the most difficulty in answering the first question. Some attributed this to an unfamiliarity with either the data or the used visualization technique.
Was there information missing which you needed to solve certain questions?
Many of the tested users missed a legend on what the amounts were they were analyzing. This was especially so in the analysis of the net traffic in Berlin according to the users. Users were not directly able to link the colors to a specific net-flow and thus the analysis became slightly harder.

What do you think of the amount of information being shown, is it too much, too little or enough?
Users in the initial tests found that some of the lines in the initial prototype between the table and the maps were too dark and overloaded the image. Based on that critique, the decision was made to give the lines a much lower opacity. One thing users noted as well is that the table in the middle of the view was sometimes too large to allow direct comparison of two different origin pick-up points. Many users suggested allowing for selection within the table on multiple origin and destination points as a possible solution for this problem.

What do you think could improve the tool in terms of usability?
Apart from the suggestions offered at previous questions, some users noted that they expected something to happen when a point on the map is clicked. A logical action one user noted would be to scroll to the associated entry in the table and highlight that entry.

Several users also noted to missing a way of filtering out all entries in the table based on a certain lower bound. These users felt that it allow the user to find the relevant information more easily.

Multiple users also pointed out that due to some of the lines crossing, the exact link between the two items became hard to recognize. One user suggested using edge bundling to result in more readable lines from the table to the map where needed.

Do you think this tool fits in the field of process mining and if so, in what cases?
There was no clear consensus on this question. Many users saw the possibility of the prototype adding value to a more classical process mining case, albeit limited to specific processes. The processes identified by the users in which the prototype could have added value include a process where the main process is executed in multiple offices or the process behind mail delivery. The users indicate that the interaction between different locations within the process can be analyzed easier with this kind of tooling. Some users also identified the potential of the prototype in visualizing a social network that is spread between multiple geographic locations.

Some users noted that the value of the prototype would be more clear in a classical business intelligence setting than in a process mining setting. The lack of a possibility to see the history of an object was also noted as a weakness in more classical business processes.

7.3 Conclusions
From both our own analysis as well as the user tests, it seems clear that the prototype adequately allows for an analyst to find answers to the questions posed. What should be noted is that almost all analysis was performed without domain knowledge. Domain knowledge, regarding the spatial organization of some of the analyzed cities, would increase the strength of the analysis. However, even without domain knowledge, some interesting results can be identified using the developed prototype. The prototype is far from perfect however, as can be noted in some of the results of the user tests. The critique of users during these tests can
be split up into two parts; critique on Flowstrates and critique on the implementation of the prototype.

The critique on Flowstrates themselves was mostly directed towards the number of lines that can appear in the larger data sets. This number of lines increases the cognitive load of the visualization and users note that it increases the difficulty of analysis. As one user noted, edge bundling could be a solution for this specific problem.

In the critique of the implementation, a common factor in much of the feedback is the lack of user interaction in the prototype. Users are able to handle the relatively busy overview of the Flowstrate and find the important points on which they want to focus. However, the prototype does not yet support a lot of the wanted interaction which is required to allow for this focus. This increases the difficulty of the analysis. An example of this lack of interaction impeding analysis is in the inability to filter multiple points in the table. Another point of critique was that some users missed a direct connection between a value in a table and the color which is displayed. This could be solved by the use of a legend to show direct color to value mappings.

Future improvements could also include an automatic aggregation on city or pick-up point level based on the zoom level of the maps or the choice of different aggregations through the use of a simple menu. The aggregation based on city or pick-up point and the choice for temporal aggregation are now chosen by a designer of the dashboard and not by the user. Using semantic zoom as described by Spence [26] and allowing for a choice in temporal aggregation, there is no longer a need for different dashboards as a user can adjust their temporal and spatial scale to zoom in on the correct areas. These improvements on the implementation are an interesting opportunity for future work.

Although the prototype is far from perfect, as noted by the users, it seems to still be able to convey sufficient information to the user about a large and complex data set. Most users also see the added value of this kind of tool in commercial software, be it in process mining software or in business intelligence software.
8 Conclusion

To conclude our research, we now return to the research question set initially:

*How can we support an analyst in selecting an appropriate spatio-temporal visualization technique for analyzing event data in a geographic context?*

We know from Chapters 2 and 3 that geographic visualization can help in the analysis of geographic data in business process data. In Chapter 4, we establish a framework to enable an analyst to make a choice regarding which visualization technique can aid in the analysis of data with geographic aspects. In Chapter 5, we provide a business case with which we tested some of the capabilities and limits of the framework. The solution from that chapter is further developed in Chapter 6 and evaluated in Chapter 7.

The main contribution of this thesis is the development of a framework through which an analyst can make a choice regarding spatio-temporal visualization techniques when they encounter event data with geographic information. The application of the framework can be done in the same manner as described in Chapter 5. An analyst should first identify the corresponding domain situation of their analysis problem, be it with requirements as in Section 5.2 or through the creation of research questions on the data. Ideally, the more fine-grained approach taken in Section 5.3.2 is not necessary. This matching of requirements and techniques directly circumvents part of the framework and requires knowledge of the available visualization techniques. The framework in its current form however did not capture the whole scope of the analysis questions and thus the more fine-grained approach was needed. Where the framework did result in a usable advice as seen in Chapter 7, the reasoning towards this advice did not cover all facets of the analysis problem. More specifically, it was noted that the framework does not hold a direct notion of temporal comparison in its structure.

Adding a temporal comparison information element could allow for a broader scope of the problems which the framework can handle, but new considerations should be made to make sure that the advised techniques are still relevant. More specifically, a look is warranted into whether the advice of origin-destination flow maps for its respective targets is still relevant. As noticed in Section 5.3.3 origin-destination flow maps are not ideal for comparing different periods of time. Further analysis is required to decide whether there is a better technique for use cases which correspond to these targets and have the need for temporal comparison.

8.1 Future Work

For the framework, we note that a temporal primitive was missing to truly capture the diversity of techniques researched. It could be that there are more primitives missing to allow for a complete comparison between the visualization techniques surveyed. These primitives will most likely be discovered through repeated use of the framework and are left for future work.

Furthermore, the framework is limited by the visualization techniques surveyed. It may be the case that the ideal visualization technique exists, but was not researched as a part of Chapter 2. An attempt was made to collect all relevant visualization techniques, but it could be possible that through the focus on spatio-temporal visualization, we did not capture the full scope. The framework allows for the addition of new techniques through the identification of the primitives which visualization techniques provide. This type of expansion
of the framework is also left for future work.

All in all, the proof of the functionality of the framework comes through available use cases. Although we have seen hypothetical cases in which the suggested solutions allow for analysis, more use cases are needed to truly test the limits of the developed framework. Ideally, we would test each geographic target within business process data with a separate use case, but this is left for future work.

For the visualization technique used for our use case in Chapters 5, 6 and 7 it is noted that there are opportunities for improvement. An interesting option could be to see the effects of edge bundling on Flowstrates with large data sets to increase readability. Another interesting option for improvement would be to add semantic zooming to an implementation of Flowstrates, iff the analysis warrants it.
References


A Accepted Papers

The list of all accepted papers from Section 3.1 is provided below. The papers which are researched more in-depth as a hypothetical use case are shown in bold. Note that the numbers below do not refer to the references in the main body of the thesis.

References


B User Test Notes

Notes taken during the user tests can be found below. These notes were used to create the summary in Section 7.2.

Test 1

• Finding Service Hub
  – User looks for large spread of reservations
  – Concludes Frankfurt am Main & Berlin
– Uses both From-To, To-From and Ratio to confirm theory about Frankfurt

• Berlin

• Finding Oktoberfest
  – User zooms in on Stiglmaierplatz on the ”From” map
  – Goetheplatz next
  – Checks theory by looking at to map of Goetheplatz

• Daily Trends Rüsselsheim
  – Campus to Elisabethenstraße (student complex?)
  – User thinks student traffic

• Net Traffic Berlin
  – Haus von Kulturen seems continuous negative flow
  – TODO: Fix Hint of Berlin Dashboard
  – Neue Hochstraße negative flow is decreasing over time, needs extra look over time
  – Swinemünder Str. increasing negative flow, might warrant extra look
  – Inflow seems much smaller than outflow, problem?

• Eval Q1:
  – Yes, but it can be better

• Eval Q2:
  – Certain filters, hover info should be there

• Eval Q3:
  – Good, but filtering is needed on the image

• Eval Q4:
  – Selections on the table

• Eval Q5:
  – On purely process data no, but with meta data yes. Combination of Geo and table is very powerful

Test 2

• Finding Oktoberfest
  – Selection in relevant weeks
  – Misses a legend
  – Concludes Goetheplatz/Stiglmaierplatz

• Daily Trends Rüsselsheim
  – Campus to Elisabethenstraße and back
  – User suspects early student traffic between 7:00-13:00

• Net Traffic Berlin
  – Color is slightly different in table than map
- Blue is harder to read than red
- Haus der Kulturen and Axel-Springerstraße good candidates for exchange

- Finding Service Hub
  - Uses To-From view
  - User suspects Hamburg-Darmstadt, or Berlin Frankfurt
  - User confirms through use of ratio view

- Eval Q1:
  - Yes, but some questions are harder than other, specifically last one

- Eval Q2:
  - More information about data set might be useful

- Eval Q3:
  - Maybe a bit too much, table is too large to allow for comparison

- Eval Q4:
  - Legend is needed, too many lines, might be able to solve some of the questions without geo.

- Eval Q5:
  - Mostly to show specific social networks.

**Test 3**

- Finding Service Hub
  - Suspects Frankfurt by location and heavy traffic

- Finding Oktoberfest
  - Filtering on weeks does not provide perfect clarity
  - Concludes Goetheplatz/Stiglmaierplatz
  - Selection on location would be very nice

- Daily Trends Rüsselsheim
  - Campus has heavy traffic
  - Back to Elisabethenstraße pretty early, probably located near station

- Net Traffic Berlin
  - Lightblue is hard to read
  - S-Bhf for pick
  - User suggests two routes, one north and one south for replenishment

- Eval Q1:
  - Yes, but first question was hard

- Eval Q2:
  - No, but the information was a bit vague, a better explanation of the data at the start might help.
• Eval Q3:
  – Missed an ability to filter effectively
• Eval Q4:
  – Without selection the lines on the geo view are very busy, filtering on a minimum amount of traffic might be useful
• Eval Q5:
  – No, the questions fits more with BI than with process mining. Maybe for specific goals this could be used in process mining.

Test 4

• Finding Service Hub
  – Looks for a central point geographically
  – Suspects Frankfurt am Main and Berlin
• Finding Oktoberfest
  – Filters on relevant weeks
  – Concludes Goetheplats
• Daily Trends Rüsselsheim
  – Focuses on morning/afternoon
  – Elishbethenstraße to Campus might be studenttraffic.
• Net Traffic Berlin
  – Haus der Kulturen and Bhf’s good candidates for exchange
  – In looking for patterns, user wanted a selection of the extremes to compare geographically
• Eval Q1:
  – It took some getting used to, especially using combinations of data in table and in geo
• Eval Q2:
  – A legend or scale
• Eval Q3:
  – A lot of information at the beginning, lots of data in the middle which is very powerful
• Eval Q4:
  – -
• Eval Q5:
  – Especially with interactions between locations is very visible, process within a company is harder to see using this technique. You can’t really see where it starts and where it begins.

Test 5
• Finding Oktoberfest
  – User concludes Goetheplatz and Stiglmaierplatz.
  – Both have high peaks in 40.
• Daily Trends Rüsselsheim
  – Elisabethenstraße to Campus between 7:00 and 8:00
  – Campus to Elisabethenstraße between 13:00 and 14:00
  – User suspects lunch break, but does not see traffic back to campus.
  – Concludes Elisabethenstraße is near station
• Net Traffic Berlin
  – Haus der Kulturen and Antwerpenerstraße good candidates to bring bikes
  – Hbf’s good candidates to take bikes.
• Finding Service Hub
  – Uses To-From view
  – User concludes Frankfurt am Main, Marburg and Berlin
• Eval Q1:
  – Yes, but the last question was considerably harder.
• Eval Q2:
  – Legend at the Net Traffic view could help a lot.
• Eval Q3:
  – For analysis purposes, it is good. User considered it illogical to use place in analysis except for the service question.
• Eval Q4:
  – Explanation/context might be needed to get to know more about the data. Minimum bound on amount of traffic might help.
• Eval Q5:
  – Difficult to say, this goes more towards BI, it is more about amounts and there is no direct process to link it to.

Test 6

• Finding Oktoberfest
  – User concludes Goetheplatz, Odeonplatz, maybe Stiglmaierplatz.
  – User mainly looked at change in relevant weeks
• Daily Trends Rüsselsheim
  – Elisabethenstraße to Campus and back
  – Elisabethenstraße probably near main Station
  – Campus near study place
• Net Traffic Berlin
- Haus der Kulturen prime candidate to bring bikes to
- Blue entries good to take bikes from
- User notes that the colors on table are clearer than colors on map.

- Finding Service Hub
  - User suspects Frankfurt am Main, Berlin (?) and Hamburg.
  - User notes that the table is pretty weird.

- Eval Q1:
  - Yes

- Eval Q2:
  - Sorting/Grouping on location would be nice.

- Eval Q3:
  - Not too much, but for columns it is a bit unclear.

- Eval Q4:
  - User noted that there was the expectation that clicking on a point on map would result in an action
  - User would let the click scroll to relevant entry in table and highlight it.

- Eval Q5:
  - Mostly for specific processes, with physical parts that move with the process, for instance postal.

**Test 7**

- Finding Oktoberfest
  - User concludes Stiglmaierplatz, Goetheplatz
  - User notes margins of the header containing weeks could be slightly larger

- Daily Trends Rüsselsheim
  - Elisabethenstraße to Campus around 7:00-8:00, probably a station
  - Campus to Elisabethenstraße around 13:00
  - User asks questions regarding very short work-day.

- Net Traffic Berlin
  - HBf’s and other Bahnhofs to Haus der Kulturen
  - User would take a lot of the bikes from the city center to outskirts

- Finding Service Hub
  - User suspects Frankfurt am Main and Berlin

- Eval Q1:
  - Yes, except for last question.
  - The Net Traffic dashboard was harder than the others.

- Eval Q2:
Not really, sorting on total amount of traffic would help

- Eval Q3:
  - Enough, sometimes the amount of lines made the image very busy and the lines were not traceable using a mouse.

- Eval Q4:
  - Merge overlapping nodes on map, adjust width of squares according to choice of temporal distribution.

- Eval Q5:
  - Works with a from and to distribution and an amount between each, but the user has trouble seeing it work with a complete process.

**Test 8**

- Finding Oktoberfest
  - User concludes Stiglmaierplatz, Goetheplatz

- Daily Trends Rüsselsheim
  - User identifies traffic from Students and teachers
  - Traffic in midday will most likely be from people looking for lunch.
  - User suspects people are heading home near Elisabethenstraße.
  - Bahnhof also moves a lot of people to the campus
  - User also notices traffic between campuses.

- Net Traffic Berlin
  - User would move bikes to Haus der Kulturen, Swinemunderstraße and Antwerpen-erstraße
  - User would take bikes from HBF’s, Rudi-Dutschke and Potsdamerplatz
  - The user notes that there are a lot of red points compared to the amount of blue points.

- Finding Service Hub
  - User suspects Frankfurt am Main
  - User notices heavy traffic from Berlin to Hamburg, but does not identify one of the two as a hub.

- Eval Q1:
  - Yes, but it takes getting used to. User notes that he used to table more than map(s).

- Eval Q2:
  - No

- Eval Q3:
  - The initial state of a dashboard is good, user would change map filtering, to sorting on the visible.

- Eval Q4:
C Use Case Questions

A short description of the process leading up to the creation of the use case questions from chapter 5 can be found here.

C.1 1. Which city performs the most service cases from other cities and can we find reasons for the location of this city?

This question is based on communication with the Deutsche Bahn on what the exact nature is of service reservations. The answer to this by the Deutsche Bahn was: Service bookings can have a variety of backgrounds. For example: cleaning, repairs, location changes, etc. After examining the data and noticing bikes travelling between cities, the idea was conceived to see if this was part of the process or not.

C.2 2. Can we detect special events happening in the usage of the Call A Bike system?

This question is based on experience of one of the supervisors of this thesis project, Dr. Dirk Fahland. He noted that an event such as Oktoberfest should be visible in this kind of data and provided details on where and when we should look for specific Oktoberfest activity. After manual data exploration, it was obvious that Oktoberfest was visible at the Goetheplatz and thus an interesting case to examine.

C.3 3. Are there areas which have large net in- or outflow and can we detect trends in these spots?

This question is based on data exploration and, more specifically, the building of several mockups during the project based on the actual data. Berlin was chosen for this case as it does show a relative, albeit skewed, balance of how many bikes are taken and arrive at each pick-up point.

C.4 4. Can we find patterns in daily usage of bikes? What could be the behavior that causes these patterns?

This question is based on data exploration as well as some expectations regarding the behavior that could be visible from the data. Kiel was chosen as a target city, due to its relative small data set which allows for easy testing and development, while still having a clearly visible pattern of student behavior.
C.5 5. Are there factors that influence when a bike needs to be serviced?

Just as with question 1, this question was based on communication with the Deutsche Bahn regarding service reservations. From that communication, we expected to see more in the data regarding what happens before a service reservation.

C.6 6. What does the life cycle of a bike look like?

As with question 1 and 5, the communication with the Deutsche Bahn lead to an expectation regarding the visibility of a clear standard life cycle of a bike.

C.7 7. What influences how long a service reservation takes?

This question was first a different question based on a large difference between booking length and trip length of a reservation in the data. After asking about this difference, the Deutsche Bahn answered the following: *Booking length is the total time of the booking. Trip length is the journey time. It is normal that both are usually very different. I would recommend only to consider the booking length, because at that time the bike was not available for other customers.* Based on that answer, the choice was made to focus more on the booking length of service reservations as other information may not be as interesting.

C.8 8. Can we spot patterns in the appearance of reservations which end up in no-shows?

After exploring the data and spotting the no-show tag, the idea was conceived to see if there are patterns in what kind of reservations end up in a no-show.