Department of Informatics
University of Fribourg

VISUALIZATION OF TEMPORAL ORIGIN-DESTINATION DATA

THESIS

presented to the Faculty of Science
of the University of Fribourg (Switzerland)
in consideration for the award of the academic grade of
Doctor Scientiarum Informaticarum

by

Ilya Boyandin
from Saint-Petersburg, Russia

Thesis № 1786
UniPrint, Fribourg
2013

Accepted by the Faculty of Science of the University of Fribourg (Switzerland) upon the recommendation of:

Prof. Béat Hirsbrunner, DIUF, University of Fribourg, Switzerland, President of the Committee

Dr. Denis Lalanne, DIUF, University of Fribourg, Switzerland, Thesis Director

Prof. Enrico Bertini, Polytechnic Institute of New York University, United States, Thesis Co-supervisor

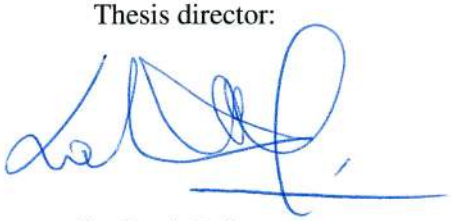
Dr. Gennady Andrienko, Fraunhofer Institute IAIS, Germany, Examiner

Prof. Jo Wood, School of Informatics, City University London, United Kingdom, Examiner

Prof. Rolf Ingold, DIUF, University of Fribourg, Switzerland, Examiner.

Fribourg, March 27, 2013.

Thesis director:



Dr. Denis Lalanne

Dean:

Prof. Fritz Müller

In memory of my father

"Much change in the world is due to geographical movement."

[Waldo R. Tobler]

Abstract

In this thesis we tackle the problem of the visual exploration of temporal origin-destination data (OD-data), which can represent movement of people, energy, material, money, exchange of information or ideas between locations in geographic space. Many OD-datasets represent global processes of profound importance, for instance, flows of refugees and trade between countries, financial aid given to countries and scientific collaborations between universities across the world. However, no universally good solutions have been developed so far for the analysis of OD-data, especially, when it comes to the analysis of temporal changes in them. The premise of this thesis is that the use of interactive exploratory visualization is a key to address the challenges created by the inherent complexity of temporal OD-data. Hence, the real challenge is in designing interactive visualizations which allow the exploration of such complex data and which help to find in these data pieces of useful information.

The work described in this thesis is an attempt to gain a better understanding of the process of interactive visualization of temporal OD-data as a whole and, with this understanding, to facilitate the development of techniques and tools enabling the visual exploration of temporal OD-data. We study what can be learned from these data, what are their possible representations and approaches to their visual analysis, which tasks different representations are better suited for and how they are actually used.

The contributions of this thesis include a survey and a critical analysis of the existing techniques, a taxonomy of the tasks which OD-data visualizations can support, and a systematization of the design-space of temporal OD-data visualizations. Building these taxonomies allows us to formulate design recommendations which can help developers of visual exploration tools to choose representations depending on the tasks which need to be supported. To better understand the insights which can be made while exploring temporal OD-data and to identify the differences in the types of insights which different representations can facilitate we carried out a user study in which we analyzed findings made by the study participants with the use of animated and small multiple flow maps. We developed a novel technique for the visual exploration of temporal OD-datasets which brings together a spatial and a temporal representations and presents an easy to read and easy to navigate depiction of this complex data type. We carried out a design study in which we applied our insights concerning the temporal OD-data visualization to a real-world problem involving real users in the process and performed a retrospective analysis which can serve as guidelines for future design studies.

The work described in this thesis has a substantial impact as the tools we developed have been already used by data analysts in various domains. This work may lead to the emergence of more comprehensive and universal solutions helping analysts to make sense of their data and supporting decision makers in finding and implementing the right policies.

Zusammenfassung

Die vorliegende Dissertation widmet sich der Visualisierung sogenannter *Origin-Destination Data* (Flussdaten mit Ausgangs- und Zielpunkten) unter Berücksichtigung ihrer zeitlichen Komponente. Diese Daten können sowohl Ströme von Menschen, Energie, Materialien oder Geld als auch den Austausch von Informationen und Ideen zwischen verschiedenen Orten darstellen. Viele der in dieser Dissertation visualisierten Datensätze repräsentieren wichtige globale Prozesse wie zum Beispiel Flüchtlings- und Warenströme zwischen verschiedenen Ländern, Finanzhilfe oder die wissenschaftliche Zusammenarbeit zwischen Universitäten. Für die Analyse von Origin-Destination Data gibt es allerdings keine universell anwendbaren Lösungen. Dies gilt vor allem dann, wenn auch die zeitlichen Veränderungen der Datenströme berücksichtigt werden sollen. Bei der Analyse dieser komplexen Daten spielt die interaktive Visualisierung eine wesentliche Rolle. Die Herausforderung bei der Entwicklung interaktiver Visualisierungen besteht darin, die Datenanalyse so einfach wie möglich zu gestalten, damit die Analysten viele nützliche Informationen aus den Daten gewinnen können.

Die vorliegende Dissertation stellt einen Versuch dar, den gesamten Prozess der interaktiven Visualisierung von Origin-Destination Data besser zu verstehen und daraus wertvolle Rückschlüsse für die Entwicklung von Techniken und Tools zur visuellen Analyse solcher Daten zu ziehen. Es wurde untersucht, welche Informationen man aus diesen Daten herauslesen kann, welche Visualisierungsmöglichkeiten und -ansätze es gibt, für welche Aufgaben sich bestimmte Visualisierungsansätze am besten eignen und wie die jeweiligen Visualisierungen tatsächlich eingesetzt werden.

Das Ergebnis der vorliegenden Dissertation ist eine kritische Analyse der bestehenden Methoden, eine Taxonomie der Aufgaben, die mit den beschriebenen Visualisierungen gelöst werden können, und eine Systematisierung verschiedener Ansätze zur Visualisierung von Origin-Destination Data, bei denen auch der zeitliche Aspekt dargestellt werden soll. Diese Taxonomien dienen als Grundlage zur Formulierung von Empfehlungen für Entwickler von visuellen Analyse-Tools, die diesen helfen sollen, eine geeignete Visualisierungsmethode für eine bestimmte Aufgabe zu wählen. Um die Ergebnisse der Datenanalyse besser zu verstehen und die Unterschiede der Ergebnisse verschiedener Visualisierungen herauszuarbeiten, haben wir eine Nutzerstudie durchgeführt. Inhalt dieser Studie war der Vergleich von animierten Flow Maps und Flow Maps mit Small Multiples. Wir haben eine neue Visualisierungstechnik entwickelt, mit der sowohl die räumliche als auch die zeitliche Komponente von Origin-Destination Data einfach dargestellt werden kann. Ausserdem haben wir eine Anwendungsstudie durchgeführt, in der wir unsere Erkenntnisse mithilfe von Testnutzern auf eine reale Fragestellung angewendet haben. Die Nutzer wurden um eine retrospektive Analyse gebeten, die als Richtlinie für zukünftige Studien dienen kann.

Die in der vorliegenden Dissertation beschriebene Arbeit hat bereits viel positives Echo gefunden: Die im Zuge der Dissertation entwickelten Tools wurden bereits von Analysten aus den verschiedensten Bereichen eingesetzt. Die Ergebnisse der vorliegenden Dissertation können zur Entwicklung neuer umfassender Lösungen beitragen, die Analysten helfen sollen, wichtige Erkenntnisse über ihre Daten zu gewinnen und die Entscheidungsträger bei der Entscheidungsfindung und Realisierung von wichtigen Massnahmen zu unterstützen.

Acknowledgements

I wish to thank my thesis supervisors Denis Lalanne and Enrico Bertini for helping me so much on this journey, my fellow students and colleagues at the University of Fribourg who made the journey so enjoyable, the fairy city of Fribourg where I spent four years which were one of the best times in my life, my wife Tanja for her patience and support, and my son Oliver for deciding to be born just one week before the day I had to send my thesis to the jury (that helped me a lot to finalize it!). Many thanks to the jury members for taking their time to carefully review the thesis and making many great suggestions on how to improve it. Finally, thanks to the Swiss National Science Foundation for funding the work on this thesis.

Ilya Boyandin
Zollikerberg, Switzerland, September 2013

Contents

Chapter 1

Introduction

"Our primary goal is to understand and make it easier to produce visualizations".

[Tamara Munzner about visualization research in general]

1.1 Motivation

In the world of today many things which take place in reality are represented as data. It was estimated that in 2012 the world generated 1.8 zettabytes[1] of data and the amount is expected to increase at a rate of 59 percent annually[2]. The easiness of automated data collection, cheap storage facilities, and most importantly, the potential impact of the knowledge which can be extracted from data make data analysis indispensable in many fields of human activity.

A significant number of the processes taking place in our world are *spatial interactions*, that is, movement of people, energy, material, money and exchange of information or ideas between locations in geographic space. Spatial interactions are often captured in the form of origin-destination data (OD-data) which is a way of representing them in the memory of a computer enabling the use of automated methods supporting data analysis. Some of the OD-datasets we consider represent global processes, for instance, flows of refugees and trade between countries, financial aid given to countries and scientific collaborations between universities over the world. In many cases analyzing such data is of profound importance, as the decisions based on their understanding can have huge impact on people's lives.

Making sense of these vast amounts of data and extracting knowledge from them presents a challenge and requires adequate analysis and exploration tools. Since computers became a commodity and the use of interactive graphics revolutionized the field of cartography, lots of software tools for visual exploration of spatio-temporal data have been developed. But the full potential of data representing spatial interactions still remains largely unrealized [Rae, 2009; Andrienko et al., 2010; Marble et al., 1997]. There are currently no universally good solutions for the analysis of this kind of data [Andrienko and Andrienko, 2012]. The existing tools and approaches for visualizing OD-data have shortcomings and their limited capabilities require scrutiny. When it comes to the analysis of temporal OD-data the situation is even more challenging. Techniques which provide comprehensive support for the exploration of spatial interactions and the analysis of changes over time still have to be developed.

The overall goal of this thesis is to facilitate this development by better understanding what can be learned from these data, what are their possible representations and approaches to their visual analysis, which tasks different representations are better suited for and how they are actually used.

1.2 Background

In this section we briefly introduce the domains of information visualization and exploratory data analysis so that we can argue how their offerings can enable the analysis of temporal origin-destination data.

Vision has the highest bandwidth among the human senses and can rapidly transfer large amounts of information into the cognitive centers in our brains [Ware, 2012]. In addition, the visual system provides us with great pattern recognition abilities and with pre-attentive processing which filters and selects what stands out for more complete conscious processing. Visualization is merely a way to employ these impressive capabilities of our visual system and to reinforce our cognition for processing large amounts of data, comprehending them and finding patterns, relationships and bits of useful information.

Visualizing spatial interactions is pertinent to *information visualization* which deals with mapping abstract data into effective visual forms. Card et al. [1999] define information visualization as "the use of computer-supported, interactive visual representations of abstract data to amplify cognition". Abstract

[1] or $1.8 \cdot 10^{21}$ bytes

[2] As estimated in 2011 by IDC (the International Data Corporation).

data is portrayed with the use of abstract visual properties, for instance, country population can be represented by mapping it to the area of the circles positioned in the country centroids on a geographic map or with the fill color of the polygons depicting the countries. Our natural ability to perceive such visual properties makes it possible to rapidly apprehend large amounts of abstract data given that the visual forms which are used to represent them are truly effective.

The main goal of visualization is to provide insight, that is, useful facts or knowledge concerning the data under analysis. Hence, visualizations must take into account the rules of the human visual system to accurately portray the characteristics of the data which are important for the analysis and to avoid misleading their users.

Visualization is generally used with one of the three following purposes [Keim, 2001]:

Exploration Analyzing and finding patterns and relationships in data without having specific hypotheses about them in advance.

Confirmation Examining and confirming hypotheses about the data.

Communication Communicating in the most appropriate way a known fact about the data.

Exploration means that the analyst begins with an imprecise or a very general purpose in mind and manipulates the visualization searching for interesting patterns or any kind of useful information in the data. The confirmatory use implies a much more goal-oriented examination of the data, that is, analysts have very specific questions to address and the manipulations they perform with the data are only meant to obtain answers to these questions. However, the border between these two modes can be easily crossed during the analysis. While exploring analysts can have new hypotheses which they want to examine, whereas to confirm a hypothesis it might be necessary to perform some exploration.

In theory, the same graphical representations can be used for exploration, confirmation and communication. However, exploration usually implies that a rather broad number of questions can be asked about the data and requires more universal approaches, while the use of visualization for communication focuses on one or several very specific questions and strives to find the most effective visual forms to give answers to these specific questions. Therefore, one can argue that interactive visualizations allowing the examination of different facets of data are usually more appropriate for exploration, whereas static visualizations or visualizations with simple interactive capabilities are more effective for communication purposes.

The approach to data analysis which concentrates on finding patterns and relationships in data by exploring them without having concrete hypotheses in advance is called *exploratory data analysis* (EDA). It was introduced by John Tukey in his influential book [Tukey, 1977] in which he demonstrated how pictures can rapidly give an insight into data without having to build a statistical model or having formulated hypotheses. Unlike pure statistical methods EDA is strongly associated with the use of graphical representations and interactive graphical analysis tools. EDA implies that there is a certain purpose in the form of general questions about the data which motivates the analysis, but most of the concrete questions arise during the analysis process itself [Andrienko and Andrienko, 2006]. For our thesis EDA is an important concept as we believe that the use of interactive visual exploration, which is closely related to EDA, is a key approach addressing the challenges posed by the inherent complexity of temporal OD-data.

1.3 Visualization of temporal OD-data

In this thesis we focus on facilitating the exploratory use of visualization for the analysis of temporal OD-data. Because of the complexity of this kind of data their analysis is a difficult task. In these data

there are multiple components: origin, destination, time, magnitude, and optional thematic attributes describing the context in which the interactions take place. The spatial components (origin and destination) require a geographic representation to allow answering questions related to the spatial arrangement of the interactions. Representing temporal changes involves depicting data for different moments in time which either requires more space or a drastic increase of the density of information portrayed on the screen. The need to represent each of the components introduces constraints. These constraints make it difficult to portray all of the data components at the same time so that the analysis tasks concerning each of the components are equally well supported.

Beside that there is a scalability issue caused by the fact that spatial interactions are connections between geographic locations, and there are usually many more of them than of the locations themselves. A visualization which attempts to represent all these connections at once without sacrificing on the level of detail is likely to suffer from visual clutter or to display an overwhelming amount of information.

In other words, because of the inherent complexity of temporal OD-data, for a dataset of a considerable size it is hardly possible to produce a single static visualization which depicts the data in all its detail in a way which is devoid of clutter, readable and not overwhelming for the user.

The following approaches can be used to deal with this issue:

– reducing the size of the dataset by filtering, aggregating or summarizing the original data,

– focusing only on some of the data components in the visualization and neglecting the others,

– providing the user with the possibility to explore the data interactively.

In our opinion, support for interactive exploration, that is, the third of the above approaches, which is fully in line with Tukey's exploratory data analysis, is really the key here. In fact, the third approach would usually involve applying the other two interactively. This combination can facilitate the analysis of the data in its whole complexity by providing non-overwhelming visualizations representing specific views of the data and allowing the analyst to easily switch from one facet or representation to another or from one aggregation level to another. Compared to a single static representation the use of interactivity makes it much easier to provide support for multiple tasks in a single visualization tool.

This approach is related to OLAP, or on-line analytical processing [Gray et al., 1997], which facilitates the interactive analysis of multidimensional data by answering analytical queries in real time. It describes the query operations *drill-down* and *drill-up*, which change the aggregation level of the view of the data, and *slice* and *dice*, which query for specific subsets of the data or filter specific values. Although OLAP can be used as a backend of a visualization system [Stolte et al., 2003], but using it may result in neglecting some of the important particularities of spatio-temporal data [Wilkinson, 2005]. Besides, it only describes the kinds of queries which can be used to obtain a summary or a subset of the data, whereas the very process of query formulation can be tedious for the user. The lack of easy to use tools to obtain in the right form and visualize data might be the main reason why data analysis is such a difficult task.

All in all the real challenge in enabling the analysis of temporal OD-data, as we see it, is in designing interactive visualization tools which facilitate the exploration of these complex data and help analysts to find in them pieces of useful information. For this we need to gain a better understanding of the process of interactive visualization as a whole.

To describe this process Card et al. [1999] proposed the *information visualization reference model*. The upper part of Fig. 1.1 shows a simplified and adapted version of it. In this model the human is motivated by a certain task to explore the data and uses a visualization to gain insight into the data and extract useful information from them. The human controls the process by choosing or adjusting the visual mapping of the data or by using the interactive features the visualization provides. The process is looped, that is, the human can adjust the visualization if necessary for the task, or even think of a different task during the exploration and change the view, observe it, adjust again and so on.

Figure 1.1: Model of the process of interactive visualization (adapted from the information visualization reference model by Card et al. [1999]) and related research questions.

The basic entities in this model are the human with tasks on the agenda, the data the human needs to analyze, visualizations of these data, and the insights the human gains in the process. The bottom part of Fig. 1.1 shows the questions related to these entities which need to be answered to facilitate the process. Answering these questions in the context of the visualization of temporal OD-data can help significantly to develop effective visual exploration tools for this kind of data, and these are the questions which we intend to investigate in this thesis.

1.4 Research goals

The main objective of this thesis is to facilitate the development of interactive exploration tools for temporal OD-data by achieving a better understanding of the possible analysis tasks, the space of the design alternatives, and the ways in which temporal OD-data visualizations can be used.

More precisely, in the course of the thesis we want to achieve the following goals:

- Study the available techniques for temporal OD-data visualization, systematically describe the design-space;
- Systematize the tasks which OD-data visualizations can support;
- Formulate recommendations for visualization design depending on the tasks which need to be supported;
- Develop interactive visualizations to satisfy the needs of the analysis of real-world temporal OD-datasets;
- Study the types of insights which can be gained with the use of different temporal OD-data visualizations.

1.5 Outline of the thesis

In this section we very briefly describe the contents of the thesis chapters. For the readers' convenience the same short summaries can be found in the beginnings of the chapters.

Chapter 2 – Origin-destination data

In this chapter we introduce temporal origin-destination data more rigorously describing the data model and present several real-world datasets which we will refer to in the course of the thesis.

Chapter 3 – Flow maps
Historically flow maps have been of a huge importance for the visualization of OD-data, therefore we dedicate a separate chapter to them. We talk about their history, ways of representing various characteristics of OD-data in them, the main problems which they have and ways of addressing these problems. We also consider alternative flow mapping approaches.

Chapter 4 – Analysis tasks
The goal of this chapter is to introduce a taxonomy of tasks which can be supported by tools for visual exploration of temporal OD-data. The taxonomy will help us to understand what kinds of questions can possibly be answered by analyzing temporal OD-data. We build this taxonomy by deriving the tasks from the components of the data. This way to identify the tasks is complementary to the user-centered approach which we take in Chapter 8.

Chapter 5 – Design space exploration
Flow maps are by far not the only way of representing OD-data. There are several alternatives which we talk about in this chapter. We use a systematic approach to describe the design space of temporal OD-data visualizations and identify the analysis tasks for which each of the alternatives may be suited best. This allows us to give a number of recommendations for the choice of design alternatives depending on the tasks which must be supported in the first place.

Chapter 6 – Flowstrates
In this chapter we present Flowstrates, our technique for the visualization of temporal OD-data, which brings together a geographic and a time-oriented representation, overcomes some of the deficiencies of other approaches and provides means for identifying and analyzing spatio-temporal patterns in temporal OD-data.

Chapter 7 – User study on animation and small-multiples
Even with the development of novel and abstract visualizations flow maps will still be widely used as it is the most natural representation of OD-data. Of the multiple alternatives for representing temporal changes in flow maps small multiples and animation are the most basic ones. We analyzed the differences in the types of insights which can be gained with the use of these two representations. The chapter describes the qualitative user study we carried out to provide the basis for this analysis.

Chapter 8 – Visualizing AidData
This chapter presents a design study in which we try to shed light on the real challenges and on the process of temporal OD-data visualization taking a user-centered approach. For this we address a real-world problem of the analysis of financial aid allocated to countries. In the chapter we discuss the interviews we conducted with domain experts which let us characterize the problem and identify the important analysis tasks. We present visualizations we developed to address these tasks, consider the user feedback and talk about the lessons learned during this project.

Chapter 9 – Conclusion
A brief summary of the achievements made, concluding thoughts and suggestions for future work.

Chapter 2

Origin-destination data

In this chapter we introduce temporal origin-destination data more rigorously describing the data model and present several real-world datasets which we will refer to in the course of the thesis.

In this chapter we describe origin-destination data more rigorously than before specifying the data models which we will refer to in the rest of the thesis. We also briefly describe several temporal OD-datasets which we will be using as examples in the thesis.

2.1 Introducing temporal OD-data

Origin-destination data is a way of representing spatial interactions, that is, flows of entities between pairs of geographic locations. In OD-data the origins, the destinations and the magnitudes of the flows are specified, whereas the exact routes which the interactions between the locations were taking are not known. This type of data is quite important and very often becomes an object of the analysis.

OD-data is often represented and stored as an OD-matrix. In OD-matrix the rows and the columns correspond to the origins and the destinations respectively (in Fig. 2.1 they are specified with country codes) and the values in the cells of the matrix represent the magnitudes of the flows between the respective origins and destinations. This is essentially the same as the adjacency matrix representation of weighted directed graphs [Cormen et al., 2001].

	ESP	PER	MOZ	DEU	ISR	HUN	ARG
SEN	3.0	3.0	2.0	2.0	2.0	2.0	2.0
CHN	1.0	2.0	2.0	2.0	2.0	2.0	
CHL		7.0	1.0				
ECU	6.0	0.0	8.0	1.0	3.0	0.0	4.0
SRB	8.0	1.0	3.0				
IRQ	8.0	4.0	2.0	5.0	2.0	3.0	6.0
URY	1.0	1.0	1.0		1.0		
...							

Figure 2.1: OD-matrix representation of OD-data.

OD-matrices representing flows between a set of locations are often sparse, that is, many pairs of locations are not connected, hence, most of the flow magnitudes are zeros. In such situations it is more efficient to store only the flows with non-zero magnitudes, for instance, as an adjacency list [Cormen et al., 2001] .

In this thesis we will consider *temporal* OD-data which means that every flow between each origin and destination is associated with a point in time or a time period. In the simplest case such data can be represented as a list of "flow events":

Origin	Dest	Magnitude	Time
SEN	ESP	4.0	2010-01-15 13:43:32
CHN	PER	1.0	2010-01-15 14:05:10
SEN	ESP	4.0	2010-01-16 10:07:31
CHN	PER	1.0	2010-01-16 01:20:32
SEN	ESP	4.0	2010-01-17 03:30:16
CHN	PER	1.0	2010-01-17 22:01:55
IRQ	HUN	1.0	2010-01-17 23:14:07
...			

Figure 2.2: Temporal OD-data as a list of flow events.

Such detailed data about each individual flow is not always available in OD-datasets. Often data is temporally aggregated based on some time period, e.g. a day or a year. Such data can then be represented in the form of a matrix as shown in Fig. 2.3. In this matrix the first two columns specify the origin and the destination and all the others specify the year. The values in the cells represent the total aggregated

magnitude of all flow events that happened in the corresponding year between the respective origins and destinations.

Notice that in Fig. 2.3 some cells are empty, meaning that the data is missing. At the same time, there are also cells which have zero values. Hence, there is a distinction made between the missing data and the flows of zero magnitude. This distinction can be very important for the analysis and visualization should take it into account.

Origin	Dest	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
SEN	ESP	4.0	3.0	3.0	3.0	3.0	2.0	2.0	2.0	2.0	2.0
CHN	PER	1.0	1.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0
CHL	MOZ										
ECU	DEU			6.0	0.0	8.0	1.0	3.0	0.0	4.0	7.0
SRB	ISR										
IRQ	HUN	1.0	8.0	0.0	8.0	4.0	2.0	5.0	2.0	3.0	6.0
URY	ARG					1.0	1.0	1.0	1.0		
...											

Figure 2.3: Temporal OD-data as flows aggregated by year.

In order to produce a legible geographic visualization usually more information about the origins and destinations is needed, e.g. the full names and the geographic coordinates of the locations. These can be stored separately and referenced by an identifier (in the example below the identifier is the country code):

Code	Name	Lat	Lon
LCA	Saint Lucia	13.903085	-60.9659
MDG	Madagascar	-18.054455	47.108621
UZB	Uzbekistan	41.447353	64.79929
LSO	Lesotho	-29.595733	28.244114
SLB	Solomon Islands	-8.910545	159.537743
MDV	Maldives	3.353159	73.260862
...			

Figure 2.4: Data about the origins and destinations necessary to produce a geographic visualization.

Now we can introduce the data models for event-based and aggregated temporal OD-data which describes in more detail what attributes they can represent.

2.1.1 Event-based model

This model describes non-aggregated data for the individual flows with timestamps of the points in time when they took place. In its simplest form these data consist of a set of geographic locations and a set of flow events:

- Location
  - Name
  - Geographic properties (shape, coordinates)
  - Thematic attributes
- Flow event
  - Timestamp
  - Origin location

  - Destination location
  - Magnitude
  - Thematic attributes

Magnitude is an *attribute* or a *characteristic* of a flow event as it reflects observations or measurements. Time period, Origin and Destination are *references* which specify the circumstances of the event of the observation.

Depending on the type of the flows and the way the data was collected the name "Magnitude" might not always be appropriate, but we will still use it for simplicity. For instance, if a dataset represents migration of people, then the magnitude attribute corresponds to the number of people.

Context of the flow events

Both locations and flow events can have additional "thematic" attributes, for instance, "population of location" or "flow type" which can also be used for the analysis. Together with the time and the locations these thematic attributes define the *context* in space and time in which the flow events were taking place. Later, in Chapter 4 we will discuss analysis tasks concerned with identifying relations of the flow events to context.

2.1.2 Temporally aggregated model

When the information about the individual flow events is not available, or is too detailed for the purposes of the analysis, the data is stored in a temporally aggregated way. The time axis is split into periods uniformly and the total magnitudes of the flows falling into each of the periods are considered. This results in the following data model:

- Location
  - Name
  - Geographic properties
  - Thematic attributes
- Flow
  - Time period
  - Origin location
  - Destination location
  - Magnitude (total over the time period)
  - Thematic attributes

2.1.3 Formal definition

In mathematical terms following the definitions proposed by Andrienko et al. [2011] both of the above models can be defined as a function μ which sets a correspondence between pairs of geographic locations S to the attributes A as a function of time:

$$\mu : S \times S \to (T \to A) \tag{2.1}$$

Here T is a set of time instants (or periods, in case of the aggregated model), and A is a set of thematic attributes (the "Magnitude" is considered one of them). In the event-based model T is continuous, whereas in the aggregated model it is a discrete set.

2.2 Discussion

2.2.1 Comparison to graphs

OD-data essentially defines a weighted directed graph and, as a consequence, many graph-theoretic methods and techniques can be applied to OD-data. As in graphs, there is a set of nodes, a set of "flows" which are the edges of the graph, and the magnitudes which are the edge weights.

One important distinctive characteristic of OD-data is, however, that the locations are geographic and, thus, have fixed positions. One consequence of this is that it is a less common (but still legitimate) practice to apply to OD-data graph layouts which change the node positions [Kaufmann and Wagner, 2001].

Besides, in temporal OD-data the magnitudes change over time, meaning that the weights of the edges in the corresponding graph are also functions of time.

2.2.2 Simplifications in the models

In the above models we made several simplifications (or assumptions about the data). The first one is that we did not explicitly include in the models flow event durations as attributes of flow events. In some situations it might be important for the analysis to know how long it took each of the flows to pass from the origin to the destination. However, we believe that for aggregated datasets this situation is less common, because the durations of the individual flows vary or often are simply not known. In the datasets, which we analyzed, this information was not available. For these reasons, we omitted the duration attribute from the models. Flow durations might still be incorporated through the thematic attributes of the flows, but a more comprehensive model intended to describe all the various properties of the flows in detail should include the duration explicitly as an attribute.

Another simplification is that the time in the models we introduced is only linear. In order to detect events occurring with a certain periodicity it is helpful to use a representation which can highlight various temporal cycles, e.g. seasons of the year, days of the week. This requires a specific cyclic arrangement of the time domain [Aigner et al., 2011b]. If the data models took this into account, it would make it possible to speak explicitly about tasks related to periodical behavior of the flows. However, for the sake of simplicity we decided to stick with linear time in our models and to speak about pattern detection in general (without explicitly distinguishing tasks related to periodicity) when discussing the analysis tasks in Chapter 4.

2.2.3 OD-data and movement

Not all spatial interactions imply some kind of movement. For example, considering scientific collaborations between different universities, each collaboration can have an origin and a destination, but there is no such thing as the *route* of a collaboration. When analyzing movement data it is often necessary to see the exact routes of the movements. For example, analysts looking at bird migration usually need to see the details of the flights: the stopovers, the speed at different segments of the route, and for how long the birds were flying. Usually, OD-data is not a suitable representation for such data. However, in some situations it can be beneficial for the analysis to represent a number of movement tracks as OD-data by extracting from them aggregated flows between certain locations. Andrienko and Andrienko [2011] suggest a method which can transform a large number of trajectories into aggregate flows between areas around significant points. The significant points are automatically extracted from the trajectories based on how often different locations are visited. This method allows producing flow maps representing aggregated movement which are much easier to read and gain insight from than the usual movement trajectory representations. Hence, data representing a large number of movement tracks can be significantly simplified by extracting OD-data from them and visualizing them as flow maps.

2.2.4 Obtaining OD-data by estimation

In transportation planning generating OD-data as a result of an estimation is an important part of the process. The influential book by Voorhees [1956] "A General Theory of Traffic Movement" describes a mathematical model for estimating the numbers of trips between each origin and destination in a given area by applying the gravity model to a trip distribution. In classical transportation forecasting a four-step process is used which takes specific characteristics of the land use of the origins and destinations as the input [Weiner, 1986]. The first three steps of this process produce an estimation of the numbers of trips between pairs of origins and destinations, that is, OD-data in the form of a matrix. The fourth step assigns specific routes to each of the origin-destination trips.

Abel [2010] presents an approach to statistically estimate the magnitudes of international migration flows and their changes over time based on some available reported data for specific countries and years. The result of the approach is a temporal OD-data matrix in which the the cells for which there were no reported data are filled in with estimations obtained with the use of an expectation-maximization (EM) algorithm.

2.3 Example temporal OD-datasets

In this section we briefly describe several temporal OD-datasets which we will refer to in the rest of the thesis to illustrate different visualization approaches.

2.3.1 UNHCR refugee flows

The refugee dataset which has been collected by UNHCR [2010] (the UN Refugee Agency) contains the numbers of people who find asylum in one of the world's countries after leaving their country of origin. The dataset contains data for the last 35 years, and for each of these years there are several thousands of flows between countries.

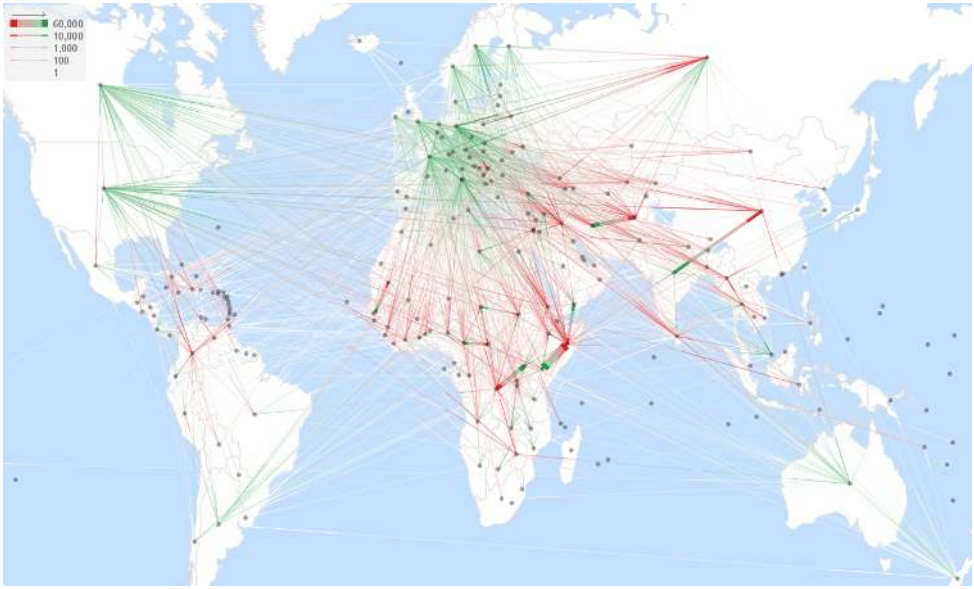


Figure 2.5: Flow map showing flows of refugees between the world's countries in 2008 from the UNHCR refugee dataset. In this flow map color is used to show the directions of the flows (from red to blue).

The records in this dataset are tuples in the following form:

```
[origin country],[destination country],[year],[number of refugees]
```

In Section 6.5.1 we discuss a visualization usage scenario with several findings made in this dataset. We also utilized this dataset for the user study discussed in Chapter 7.

2.3.2 Commuters in Slovenia

This dataset contains the numbers of people who commute to work between the towns and villages in Slovenia. There are about 17 thousand flows for the years from 2000 to 2008. As in most datasets

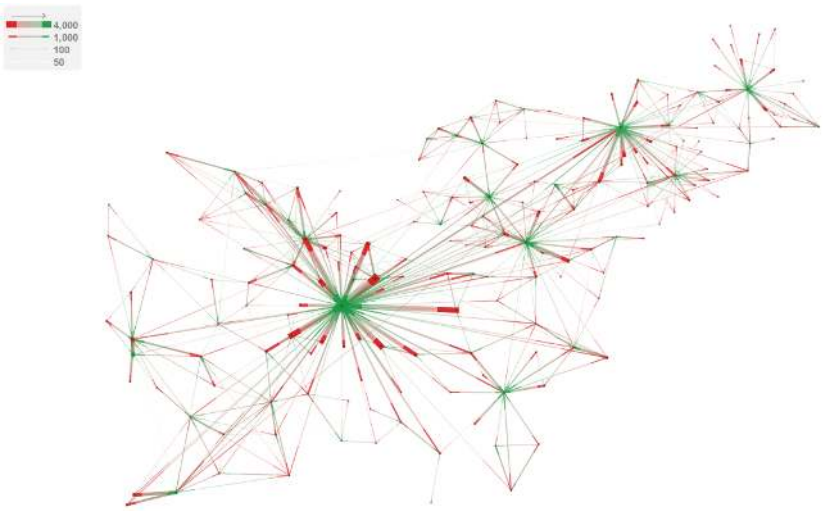


Figure 2.6: Flow map showing commuters in Slovenia in 2008 (the directions are from red to green).

showing commuters we can clearly see how the larger cities attract people from the surrounding areas. It is quite interesting to analyze how this pattern changes over time in this dataset. We discuss this in detail in Section 6.5.2.

2.3.3 AidData

Figure 2.7: AidData: flows of financial aid in 1999 (the directions are from orange to cyan).

The dataset maintained by AidData.org contains more than one million flows of financial aid to countries for the time span between 1949 and 2011. The original dataset includes detailed information

about every individual flow and the flows are classified by their purpose. We discuss questions related to the visualization of these data in detail in Chapter 8.

2.3.4 Moscow metro rides

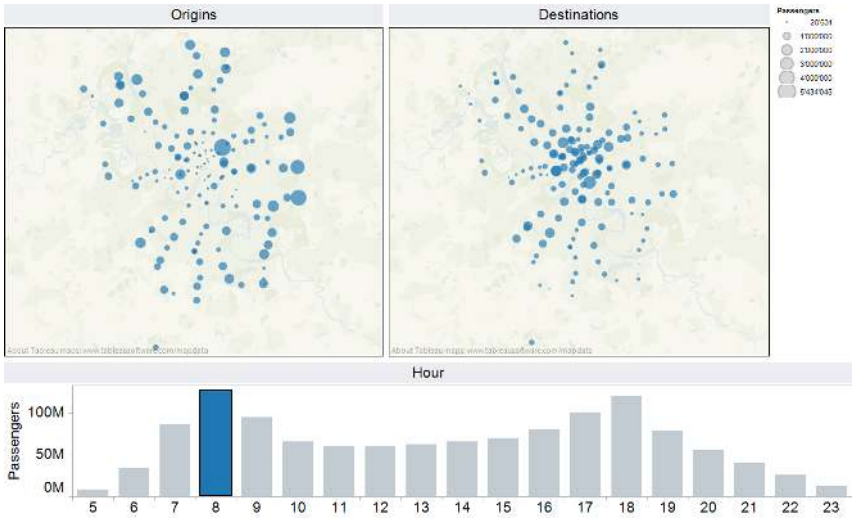


Figure 2.8: Moscow metro rides data. Here 8 a.m. is selected in the "Hour" pane, therefore, in the maps we see only the totals for the stations at 8 a.m. We can see that at 8 a.m. many passengers travel from the outskirts of the city towards the center. (Illustration made with Tableau).

The dataset courtesy of the Moscow department of transportation represents about one billion passenger rides in the Moscow metro during one year. It contains about 500K of records in the following form:

```
[origin station],[destination station],[hour],[number of passengers]
```

With these data it is possible to analyze how the spatial patterns of the rides change during the day.

2.4 Conclusion

In this chapter we gave more precise definitions of what we consider temporal OD-data. We discussed specific attributes and references such data consist of, their relation to movement data and talked about some of the situations in which such data is typically used. Finally, we briefly introduced several real-world datasets which will allow us to illustrate the practical aspects of temporal OD-data visualization.

Chapter 3

Flow maps

Historically flow maps have been of a huge importance for the visualization of OD-data, therefore we dedicate a separate chapter to them. We talk about their history, ways of representing various characteristics of OD-data in them, the main problems which they have and ways of addressing these problems. We also consider alternative flow mapping approaches.

3.1 Introduction

Flow map is the most often used representation of spatial interactions (or entities moving between geographical locations). The flows of entities are represented as lines or arrows drawn on a geographic map so that they connect the origins to the destinations and the thickness of the flow lines represents the magnitude of the flows. The main goal of this kind of representation is to support analysts in finding answers to questions related to the flow magnitudes and to the spatial arrangement of the flows. Such questions can be, for example:

- What are the magnitudes of the flows? Where are the largest flows?
- Where are the origins and the destinations of the flows located on the map?
- What are the directions of the flows?
- How far do the flows go?
- What is happening in a specific location or a region?

Answering such questions obviously requires a geographic representation which portrays the flows along with their magnitudes and flow maps are made just for this purpose.

3.2 History of flow mapping

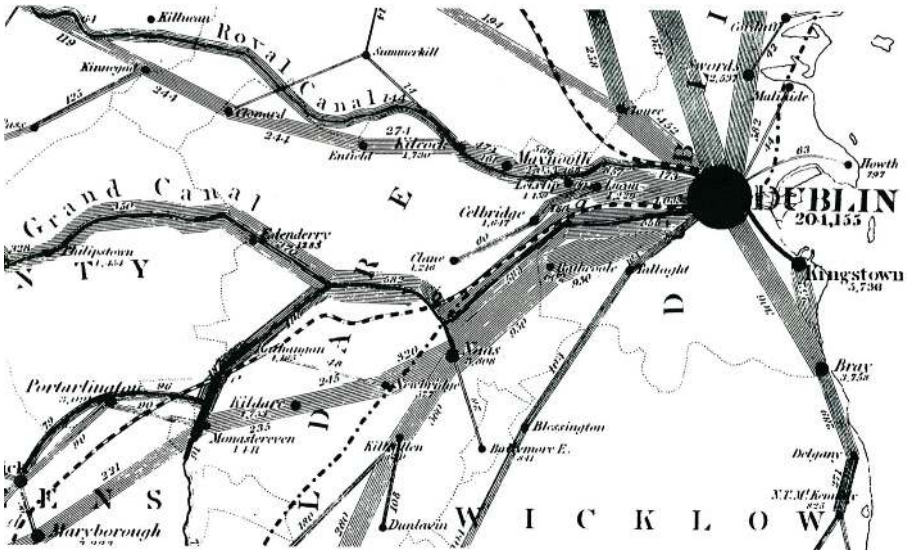


Figure 3.1: A fragment of Harness' passenger conveyance flow map, 1837. Brit. Mus.

Like many of the statistical graphs which are in use today flow maps are not a modern invention. The first known flow maps were made by Henry Harness in 1837 [Robinson, 1955]. Fig. 3.1 shows the flow map which he made for Railway Commissioners to exhibit the relative numbers of travelers conveyed in different directions throughout Ireland. [Robinson, 1982] argues that it was the first use of a proportional line to show linear, quantitative data on a thematic map.

Between 1845 and 1869 Charles Minard, the famous French thematic cartographer, produced numerous flow maps and popularized the technique. He employed them to portray flows of travelers, trade in coal, merchandise on France's transportation facilities, and international trade in cotton and wine [Robinson, 1982]. Many authors praise Minard's chart showing the decline of the size of Napoleon's army in

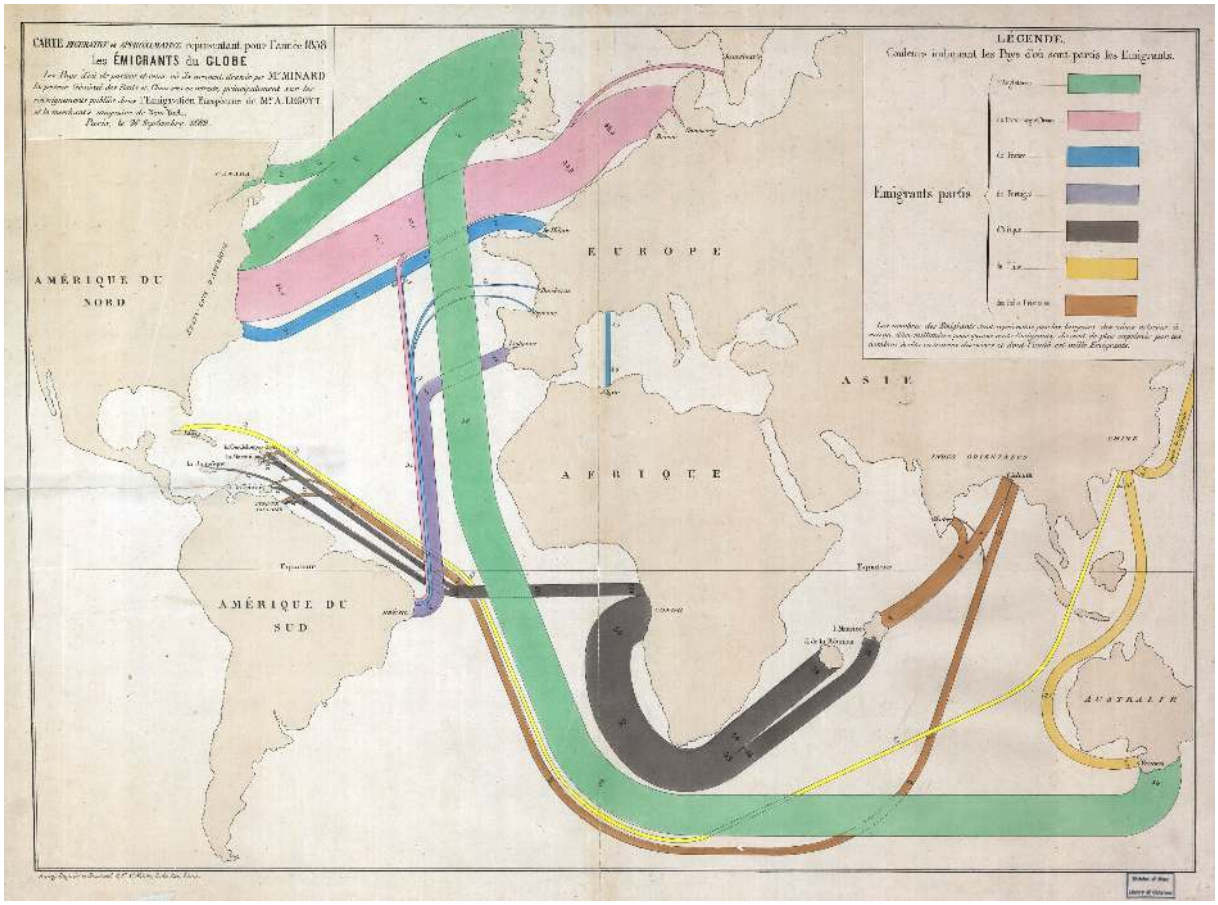


Figure 3.2: Minard's flow map showing the world's migration, 1862.

the Russian campaign [Tufte, 1986]. Fig. 3.2 shows another flow map he made in 1862 which visualizes the migration of people between the world's countries. It was probably the first published migration map showing origin and destination countries of global migrations [Robinson, 1982].

The first computer system capable of displaying an image which can be considered a flow map was developed as early as in the late 1950s. For the Chicago Area Transportation Study [1957] a special cathode ray tube system called "the cartographatron" was constructed which displayed an image representing several millions car trips as lines with the goal of estimating the volumes of the traffic. The system was then used to support the decision-making for planning new interstate highways [Black, 1990].

First computer programs for flow mapping were developed in the 1960s and 1970s. At first, they were used mostly for transportation engineering [Kern and Rushton, 1969; Wittick, 1976; Noguchi and Schneider, 1977; Beddoe, 1978], but later they started to be applied in other domains, for instance, in health care [Francis and Schneider, 1984; Gesler, 1986]. In the 1980s and 1990s many software tools for flow mapping were created which were not limited to a particular domain [Schneider, 1983; Liu, 1995; Thompson and Lavin, 1996].

One system which has been mentioned quite often in research papers was Flow Mapper developed by the cartographer Waldo Tobler. In Fig. 3.4 you can see a flow map of net movement of one dollar notes produced with the first version of Flow Mapper. Tobler [1981] not only created the software for basic flow mapping, but also investigated various approaches for making flow maps more readable by addressing their problems. We will discuss some of these approaches in the subsequent sections.

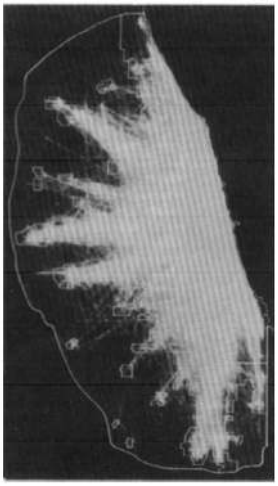


Figure 3.3: Line traces of car trips from Chicago Area Transportation Study [1959].

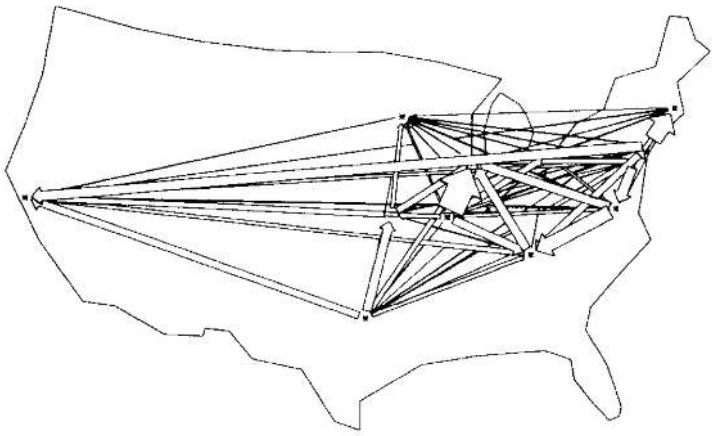


Figure 3.4: Flow map of net movement of one dollar notes produced with the first version of Flow Mapper. ©1981 The Ohio State University. Reprinted, with permission, from Tobler [1981].

3.3 Representation techniques

3.3.1 Representing the directionality of the flows

In most cases the flows are directed and seeing their directions is important for the analysis. Hence, a good flow map representation must make it possible to easily read them.

There are many different ways of showing the flow directions which can be used with flow maps, for example:

- Using arrows pointing into the direction of the flow movement
- Bending the flow line in a specific and recognizable way (e.g. making a curve towards the end)
- Varying the hue, intensity or transparency of the flow line color (e.g. green-to-red, transparent-to-opaque, desaturated-to-fully saturated)
- Varying the thickness of the flow line (e.g. from thin to thick)

– Embedding an animated moving pattern into each of the flow lines which indicates the direction.

Considering node-link graphs in general Holten et al. carried out three user studies in which various edge direction representation techniques and their combinations (see Fig. 3.5) were compared in terms of the speed and correctness of the user performance in tasks for which it was necessary to perceive the edge directions [Holten et al., 2011, 2010; Holten and van Wijk, 2009b]. The studies proved that using tapered edges to show the edge directionality was the best in terms of the speed and the use of animation (which also provided hints for the edge lengths) was best in terms of the correctness.

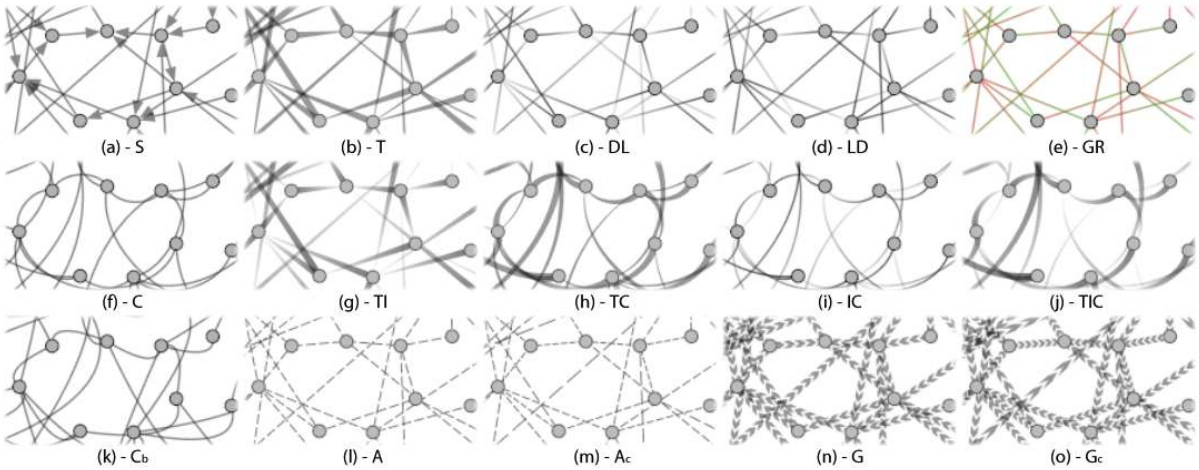


Figure 3.5: Various alternatives for representing directed edges in node-link graphs. A and A_c employed animation to show edge directions. Holten et al. [2011] showed that the design T (tapered) was the best in terms of speed and A_c (animation compressed) was the best in terms of correctness. ©2011 IEEE. Reprinted, with permission, from [Holten et al., 2011].

Flow maps are essentially node-link diagrams representing weighted graphs, therefore, it seems natural to conjecture that the results of the studies by Holten et al. apply to flow maps as well. There is a caveat, however. The way flow directions are represented can interfere with the fact that the flow line thickness is used to portray the magnitudes of the flows. If a comparable study was carried out for flow maps it would have to take this into account.

The extended study [Holten et al., 2011] analyzed the effect of the density of the graph (number of nodes and edges) on the task completion performance. It turned out that the A_c ("animation compressed") representation, which employed animation to show edge directions, was the only one which performed well across different graph densities and outperformed the tapered representation in terms of correctness for the most dense graphs. For the other representations the completion times and error rates worsened with the increase of the graph density. However, even the most dense graphs used in this study consisted of hundreds of nodes and edges. The effect of scale could be different when representing graphs with thousands or millions of edges.

Fig. 3.6 shows a flow map we made to demonstrate the use of tapered edges for representing flow directions. It turns out that beside reducing the clutter there is another advantage of this approach compared to the use of the more conventional arrows. The flow direction can be seen by looking at any part of the flow line. With arrows it is necessary to spot the arrowhead to determine the direction of a flow and this can be a significant difficulty in crowded views in which there are many flows with overlapping arrowheads. However, because of the tapered shape of the arrows, it becomes significantly less obvious how to read and compare the magnitudes of the flow lines. Besides, this representation can make it difficult to perceive the direction and magnitude of very thin and curved lines.

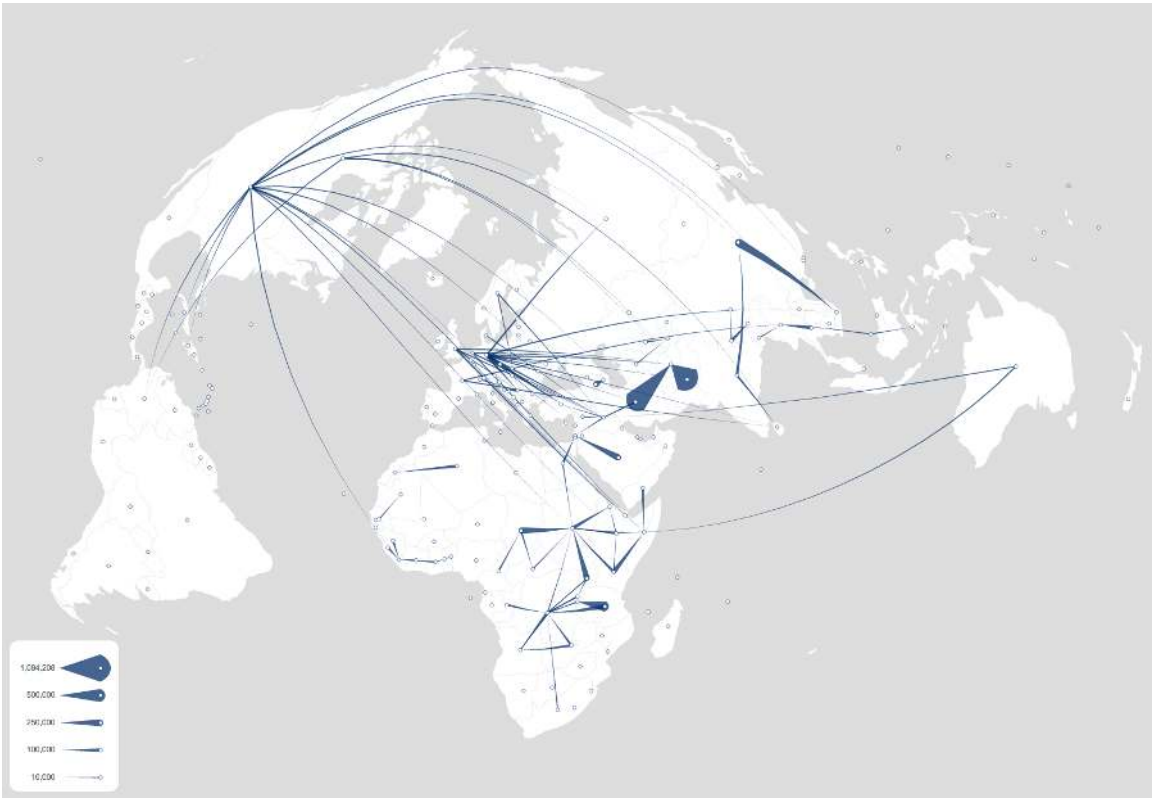


Figure 3.6: Flow map in which tapered edges are used to represent the flows and their directions. The map represents the numbers of refugees from various origins who were residing in other countries in 2005.

3.3.2 Representing the magnitudes of the flows

In flow maps the lines representing the flows usually indicate the magnitudes of the flows and their directions at the same time. Which visual variable is used for one of these two characteristics impacts the way in which the other characteristic can be represented. However, using the thickness of the flow lines is not the only way of representing flow magnitudes. Altogether, there are at least three alternatives:

Flow line thickness

Varying thickness of the flow lines is used in most cases to represent flow magnitudes in flow maps, e.g. [Tobler, 1987]. This depiction of the amount flowing is naturally perceived by humans and usually requires no explanation.

Flow line coloring

A specific color encoding assigns different colors to the flow lines depending on the magnitudes of the flows they represent. An advantage of this approach is that the thicknesses of the flow lines can be the same, thus, they occupy less space and produce less clutter. This approach is used, for instance, in [Holten and van Wijk, 2009a]. However, according to [Cleveland and McGill, 1987] color is inferior to size as a visual variable. The human perception allows us to make much less accurate color comparisons than size comparisons.

Particle animation

The magnitudes are represented by particles moving along the flow lines (which might even not be explicitly shown) from the origins towards the destinations. Either the speed of the particles or their size or their density portrays the flow magnitudes[1]. The study by Holten et al. [2011] implies that an

[1] An example of such animation showing donations to Kiva.org can be found at `http://youtu.be/Yf5QRlkWX8o`

animation in which short line segments moving along the flow lines is one of the two best ways to represent the flow directions. However, no studies so far have proven that such an animation could be used effectively for portraying flow magnitudes and can actually support basic OD-data analysis tasks. In Section 3.6.2 we consider another kind of particle animation in which the particles do not move along the flow lines, but form a vector field derived from OD-data.

3.3.3 Sankey flow maps

Often, varying the thickness of the flow lines to represent the flow magnitudes is used in combination with the Sankey flow drawing technique [Schmidt, 2006] in which the flows going from (or to) a specific location are merged into one big flow at first and are then split into smaller flows going towards different destinations.

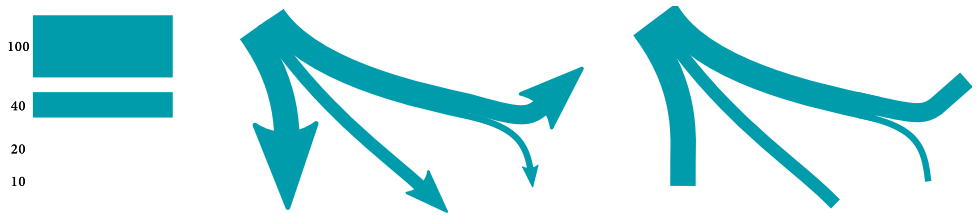


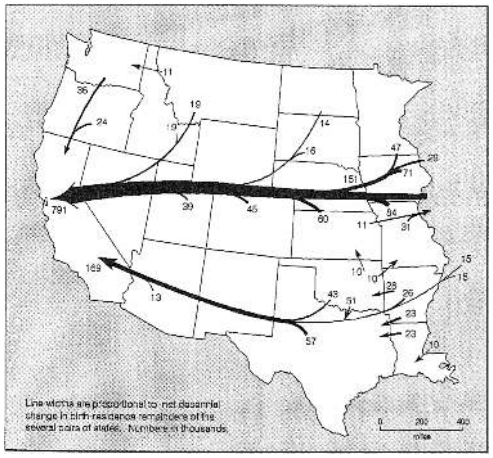
Figure 3.7: Sankey flow drawing technique with and without arrows.

This technique allows reasoning about the contributions of the individual flows to the total magnitude. Even without the arrows the directions of the individual flows are perceived to be the same, because they are represented as parts of one big flow beginning in a specific source (or flowing towards a specific destination) and separating into smaller flows along its path (see Fig. 3.7). Bringing down the number of lines and line intersections in the areas where the flow lines are merged leads to reducing the overall complexity of the visualization and making it more readable.

An application of this technique to flows of one origin can actually be seen in the very first flow map made by Harness (see Fig. 3.1), in which several large flows start from Dublin and are then split into smaller flows going to different destinations. In fact, Harness applied this technique 60 years before Matthew Sankey used it in a diagram which illustrated energy flows of a steam engine [Schmidt, 2006]. Still we use the term “Sankey flow maps” for this type of maps, because the name Sankey is now strongly associated with this particular flow representation technique, in which the flow lines are merged and split along their paths to illustrate the contributions of the individual flows to the total amount flowing. Charles Minard also used this technique in the flow maps he produced before Sankey [Tufte, 1986]. Fig. 3.8a presents a flow map made by Thornthwaite and Slentz [1934], in which they applied the same technique to group flows going towards a specific destination.

Phan et al. [2005] developed an algorithm for the automatic generation of Sankey flow maps which they called Flow Map Layout and implemented in a computer program. Buchin et al. [2011] proposed a different algorithm producing very similar but more visually appealing results (see Fig. 3.8b). The latter algorithm is also capable of drawing the flow lines avoiding the overlapping with the landmasses (very much in the style of the flow maps hand-drawn by Minard).

One downside of this approach is that it is based on trees, that means that in general it works with flows from or to one selected location. For portraying flows of multiple locations at the same time separate flow maps must be generated for each of the locations and then superimposed on the map [Phan et al., 2005]. This may result in overlapping and line crossings, which the algorithms are not capable of preventing.



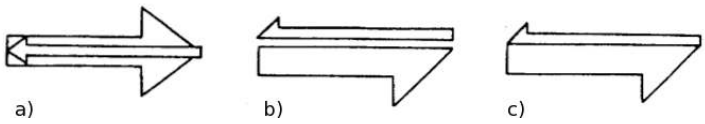
a) b)

Figure 3.8: a) Thornthwaite's flow map of US migrations (after [Thornthwaite and Slentz, 1934]) and b) Flow map showing migrations from California produced with the spiral trees algorithm by Buchin et al. ©2011 IEEE. Reprinted, with permission, from Buchin et al. [2011].

3.3.4 Bi-directional flows

Another related issue is how to represent bi-directional flows, that is, flows between the same two nodes going in the opposite directions. In flow maps each of the two opposite flows has its own magnitude, and it must be possible to perceive both of these magnitudes.

Tobler [1987] discusses two approaches for representing bi-directional flows: putting the smaller arrow on top of the larger one (a.) and using half-barbed arrows (b. and c.):



Tobler states in the paper that neither of these approaches is visually very effective, and notes that the problem can be avoided if only showing net movements between each pair of locations. The latter is, though, unacceptable in many situations when the actual flow magnitudes are of interest for the analysts.

For representing bi-directional flows Bahoken [2011] describes the following alternatives which attempt to minimize the overlapping and clutter:

The first two (from the left) of the above approaches are the ones used most often. We could speculate on the relative effectiveness of these approaches, however, no rigorous studies on this have been performed so far.

As we mentioned above, when arrows are used it is necessary to find the arrow head of a flow to see its direction. This can be difficult in situations when there are many flows going to the same nodes. To mitigate this problem van de Ven [2007] proposed an algorithm which shortens the flow lines to minimize the overlapping:

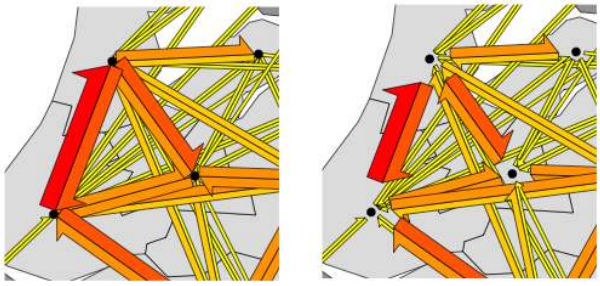


Figure 3.9: Improving the readability of a flow map by shortening the flow lines so that most of the arrow ends can be seen. The left image shows the original flow map without shortening, whereas in the right one the flow lines are shortened (after [van de Ven, 2007]).

The flow lines are shortened where they come close to the nodes, which avoids overlapping, so that as many arrow ends as possible can be seen (see Fig. 3.9). However, this approach does not work very well for nodes which are located very close to each other.

3.3.5 Self-loops

It is often the case that there are flows in OD-datasets which have the same origin and destination. They might, for instance, describe the migration within a specific location. Such flows are usually called *self-loops*. Self-loops in flow maps are usually represented with circles, rings (or donuts), or various kinds of looping arrows. These alternatives are schematically represented in Fig. 3.10.

Figure 3.10: Different ways to represent self-loops in flow maps.

An effective representation must support comparison of self-loop magnitudes to those of the other flows. The ring and the looping arrow representations make this possible, because the thickness of the ring can be compared to the thicknesses of other flow lines. An example of such flow a map is shown in Fig. 3.11.

It must be noted that sometimes similar visual cues (especially, circles) are used in flow maps to represent the total magnitudes of the incoming and outgoing flows for locations. Hence, to avoid confusion it is sensible to clarify in the legend whether self-loops or totals are depicted.

3.4 Problems with flow maps

Despite the popularity of flow maps there are significant problems connected with their use. One problem is the cluttering and occlusion caused by intersecting and overlapping lines (see Fig. 3.12). This problem becomes most apparent when many flow lines are depicted at the same time on the screen producing an “indecipherable hairball” and making it difficult to see the origins and destinations of the flows or to compare their magnitudes.

Another problem of flow maps is the fact that the longer lines take more space on the screen, and thus, might mislead the users drawing their attention away from the shorter flows of larger magnitudes. According to the first law of geography “near things are more related than distant things” [Tobler, 1970]. For spatial interactions that means that the flows between closer regions have larger magnitudes and this is indeed the case most of the time. Hence, shorter flows are often the more important ones and the user’s attention should not be drawn away from them.

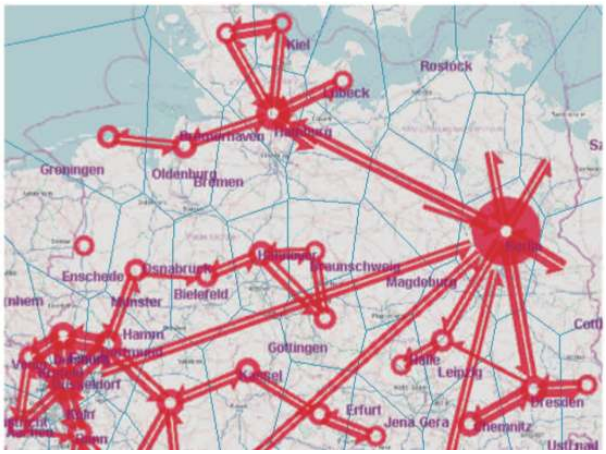


Figure 3.11: A fragment of a flow map in which half-barbed arrows are used to show bi-directional flows and rings are used for representing self-loops. The magnitudes of the self-loops can be easily compared to those of the non-self-loop flows by the thicknesses of the rings and of the flow lines. ©2011 IEEE. Reprinted, with permission, from [Andrienko and Andrienko, 2011].

As it turns out flow map users can easily make incorrect findings, because of the wrong impression this widely used geographical visualization can give in certain situations. In the user study which we described in Chapter 7 we asked the participants to analyze data represented as flow maps. We noticed that the participants made a number of incorrect findings which were caused mainly by these two properties of flow maps:

– Locations with a lot of incoming or outgoing flows appear to be more significant even if the total magnitudes are small. More generally, the totals of the incoming and outgoing flow magnitudes for locations are not easily comparable.

– Longer flows get more screen real estate which makes them appear to be more significant; very short flows might be hard to see even if their magnitudes are large.

The first of these two problems can be easily solved by introducing additional visual elements representing the totals for locations (e.g. circles). The second problem is intrinsic to flow maps and can probably only be solved by using a different representation (see Chapter 5).

3.5 Addressing clutter

Ellis and Dix [2007] catalog various approaches for clutter reduction in visualizations identifying the main benefits which such approaches can have. This taxonomy allows the evaluation of the usefulness of clutter reduction approaches. The list of benefits includes avoiding overlap, allowing a better discrimination of the points and lines, keeping spatial information, providing scalability to the dataset size. All these benefits are relevant to the cluttering problem in flow maps.

A common graph drawing approach to reducing clutter is to rearrange the node positions by applying a specific node layout algorithm [Kaufmann and Wagner, 2001]. Such algorithms usually attempt to minimize the line crossings and node-line overlapping. Sometimes this approach can be a reasonable solution for improving the readability of OD-data visualizations as well [Krempel and Plümper, 1999]. However, applying such an algorithm means that the node positions lose their geographic meaning and the tasks related to the spatial arrangement of the nodes and flows are not supported. In a genuine flow map the positions of the nodes represent their actual geographic locations.

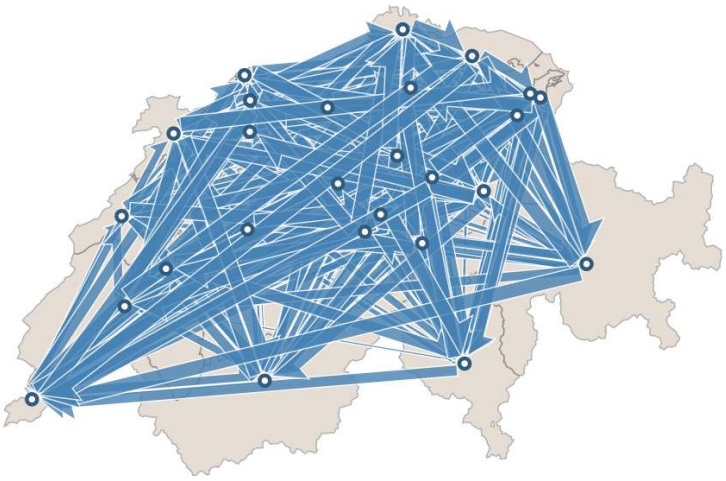


Figure 3.12: A cluttered flow map. Despite sorting the flow lines by magnitude (flows of larger magnitudes are drawn above the flows of smaller magnitudes) it is hard to grasp what is really going on in this visualization apart from just seeing the flows of the largest magnitudes.

To address the cluttering problem specifically in flow maps one of the following approaches can be used:

- Visual techniques improving readability
- Interactive filtering or automatic sampling
- First location totals, then flows on-demand
- Showing only the differences from the expected
- Coarsening the spatial resolution
- Segmenting into flows between adjoining regions
- Sankey flow maps
- Bundling

In the following subsections we discuss these strategies in more detail (except for Sankey flow maps which were discussed in Section 3.3.3).

3.5.1 Visual techniques improving readability

This can include various approaches to drawing the flow lines: e.g. sorting lines by flow magnitude so that flows of larger magnitudes are drawn above the others; showing only parts of the flow lines close to the nodes [Becker et al., 1995; Boyandin et al., 2010]; shortening the flow lines [van de Ven, 2007]; using transparent flow lines; rerouting the flow lines to avoid intersections with land (see Fig. 3.2 or [Buchin et al., 2011]). These approaches can often make the representations more readable, but all of the flow lines are still drawn in them, hence, the cluttering problem is not eliminated completely. Besides, such techniques might affect the accuracy of the users in performing analysis tasks. For instance, an evaluation of a graph representation with partially drawn links [Burch et al., 2012] showed that, albeit the completion times in certain tasks improved, the accuracy suffered in most of them.

3.5.2 Interactive filtering or automatic sampling

The idea is simply to reduce the number of flows which are displayed on the screen at the same time by keeping only "the most important". The flows can be either interactively selected by the user based on a specific property (e.g. magnitude, length, specific origin or destination) or selected automatically using a predefined heuristic.

As an example of such an automatic approach Tobler [1987] suggests removing all flows the magnitude of which is less than that of the average. According to Tobler this typically removes about 75% of the flows while removing less than 25% of the total magnitude. So instead of supporting arbitrary filter thresholds this "optimal cut-off value" can be used. Obviously, this strategy leads to a representation in which a part of the data is left out, and despite attempting to keep the most important data represented it might neglect essential details the importance of which was not foreseen.

3.5.3 First location totals, then flows on-demand

Most of the time the first thing the data analyst wants to see is a comprehensible overview. Once the analyst identifies an interesting pattern in the overview they might be interested in more details about a specific object. This idea is following the visualization mantra formulated by Shneiderman [1996] prescribing the basic workflow which exploratory visualization tools must support: "Overview first, zoom and filter, then details-on-demand". In other words, not all the available details are displayed to the analyst at once, instead they must use interaction to specify which pieces of information they want to explore in detail.

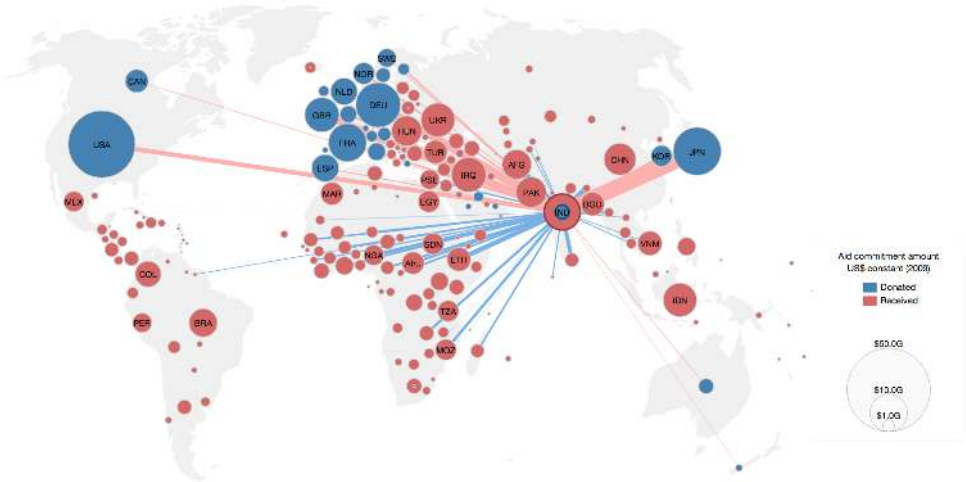


Figure 3.13: A symbol map using circles of varying size to represent the total outgoing and incoming flows' magnitudes of financial aid flows countries (refer to Chapter 8 for more details). One of the countries was highlighted by the user, therefore, the flows of this country are shown.

An OD-data visualization based on this principle might use a representation of the total magnitudes of the outgoing and incoming flows of locations and show the individual flows only when a location is selected by the user (see Fig. 3.13). An obvious disadvantage of this technique is that it only gives an overview of the location totals, and not of the flows. At a time only the flows of one location are shown, this makes it difficult to compare flows of different locations. However, if this approach fits the workflow and the tasks of the analysts it might be a good solution devoid of clutter.

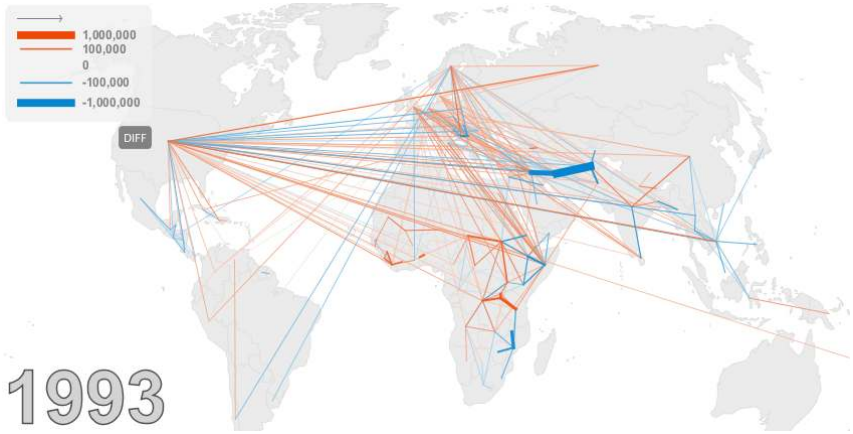


Figure 3.14: Undirected flow map showing only the differences between the flows of refugees in 1992 and 1993. Red means an increase of the flow magnitude, blue – decrease.

3.5.4 Showing only the differences from the expected

Often the most interesting parts of the data are those which deviate from the theoretical expectations. In order to be able to estimate these deviations it is necessary to build a theoretic model describing the expectations [Tobler, 1987]. If such a model exists a flow map can be produced in which only the flows differing from the expectations are portrayed, thus, reducing the number of flows shown at the same time. The same approach can also be used to represent changes of the flow magnitudes between two particular moments in time (see Fig. 7.2). In this case the flow magnitudes of a preceding moment in time (or a time period) can be seen as the ones "defining the expectations".

This approach is of limited use, because it does not show the original flow magnitudes. Hence, it can only be used as a supplementary tool or when the analysts are interested solely in the differences (or in the deviations from the expectations which can also be represented this way).

3.5.5 Coarsening the spatial resolution

To reduce the number of nodes as well as the flows which must be drawn in a flow map, nearby locations can be grouped together. This way only the flows between the groups of locations have to be shown, resulting in a significantly less cluttered flow map.

There are different methods which can be applied for location grouping, for example:

- Using coarser political or administrative geographic regions. For instance, counties can be grouped into states, countries – into continents.
- Locations can be spatially clustered with an automatic clustering method using geographic proximity as the distance measure [Boyandin et al., 2010].
- Flows of locations can be taken into account when grouping them. For example, Guo [2009] proposes a method for *regionalization* of flow maps for discovering geographic regions with "community structures". These are such groups of locations that have much more connections within the groups than among them. Guo proposes a statistical model to measure the modularity of a group of locations and to hierarchically cluster locations into regions based on this modularity measure. The original flows are then aggregated based on the discovered regions and a flow map can be rendered in which only the flows between the regions are shown (see Fig. 3.15).

Coarsening the spatial resolution reduces the complexity of a flow map by presenting the data at a higher abstraction level. However, the way the grouping of locations is performed may have a significant

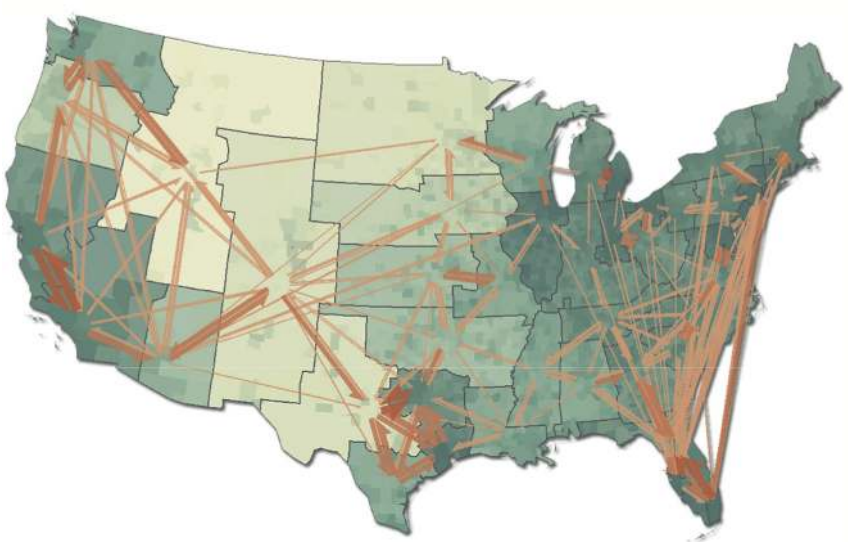


Figure 3.15: Flow map of US migrations in which only the flows between 45 regions discovered by Guo's regionalization approach are shown instead of the 800,000 flows between 3075 counties of the original dataset. ©2009 IEEE. Reprinted, with permission, from [Guo, 2009]).

influence on the movement patterns which can be recognized in the resulting flow map. In Fig. 3.15 the data on the original abstraction level are not represented (specifically, the flows between the adjacent locations within the groups). Providing an interactive support for "looking into" a group could help to alleviate this problem.

3.5.6 Segmenting into flows between adjoining regions

Instead of drawing flow lines between any arbitrary pair of locations the flows can be segmented into sequences of smaller flows between adjoining regions. Segments of different original flows which connect the same adjacent regions are then grouped together and only these resulting flows are drawn in the final flow map. As a consequence a flow map is produced which completely avoids line intersections.

The first flow map of this kind was produced by [Ravenstein, 1885] who examined the county-to-county migration in the United Kingdom and laid down "The Laws of Migration" based on this analysis. In this work Ravenstein included several cartographic illustrations one of which is a flow map showing the flows of migrants segmented into short movements between adjacent counties (see Fig.3.16).

Tobler [1981] discusses an automatic approach for producing such flow maps. To segment the original flows each of them is "rerouted" through a set of predefined locations (e.g. state centroids) with the use of a shortest path algorithm, so that each of the paths consists only of movements between adjacent locations. The resulting flow map has virtually no line intersections, and thus, is much easier to read (see Fig. 3.17).

This method produces a very good general overview devoid of clutter showing the amounts which are flowing between adjacent regions on the map, however, it has disadvantages as well. First, it introduces possibly misleading flow routes, which do not necessarily represent the real routes of the flows. More importantly, it makes it difficult or even impossible to say for any given location what the actual origins of its incoming flows are and what the actual destinations of its outgoing flows are. The resulting flow map shows only flow segments whose magnitudes are accumulated by joining segments of multiple flows which have different origins and destinations, therefore it is impossible to say how much of a location's inbound flow amount actually originates in this location, how much of it is the transit coming from other origins and what part of the transit amount stays in the location.

Figure 3.16: Ravenstein's "Currents of migration" showing UK migrants, 1885 (after [Tobler, 1995]).

In other words, by segmenting flows and then joining segments of multiple flows into flows between adjacent locations the origin-destinaton aspect of the data is disregarded to a large extent. Hence, this approach cannot be recommended in situations when being able to see the actual origins and the destinations of the flows and the correspondences between them is important for the analysis.

3.5.7 Bundling

Bundling (or "edge bundling" when applied to graphs in general) is rerouting flow lines which pass close to each other and go into similar directions, so that they form bundles (see Fig. 3.18). In the last years a vast number of papers has been published proposing various edge bundling algorithms for graph visualization in general [Cui et al., 2008; Holten, 2006; Holten and van Wijk, 2009a; Telea and Ersoy, 2010; Lambert et al., 2010; Ersoy and Telea, 2011; Ersoy et al., 2011; Pupyrev et al., 2011; Hurter et al., 2012].

Bundling is sometimes confused with Sankey flow maps because of the similarity of the two techniques, however, there are important differences between them. Bundling algorithms have the advantage that they can be applied to arbitrary graphs without the limitation of having to select a specific subtree of the graph. However, they cannot be readily applied to flow maps. Most of these bundling algorithms produce visualizations which do not support easy comparison of the numbers of edges constituting bundles nor do they accurately represent the total magnitudes flowing through bundles. Parts of individual edges which constitute bundles are not merged together, they remain separate. Some of them overlap, some of them go next to each other, so it is not really possible to compare the bundle sizes.

Holten and van Wijk [2009a] attempt to mitigate the problem by using color to represent the number of edges passing through each pixel of the visualization (see Fig. 3.18). However, this does not improve the situation much, because each bundle has a certain thickness which is naturally perceived as an indicator of its size and neither the color nor the thickness alone is a reliable indicator, it is the combination of the two which conveys the actual sizes of the bundles. This makes comparison quite difficult and unreliable.

In this sense, segmenting into flows between adjoining regions, which we discussed in Section 3.5.6,

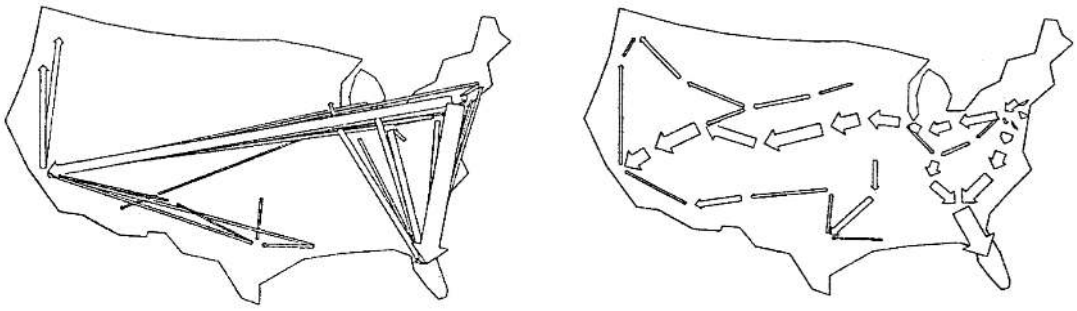


Figure 3.17: Reducing the complexity of a flow map by segmenting flows into series of shorter flows between adjoining regions. The left image shows the original flow map of the net migration between the US states (with all flows of a magnitude below the average deleted). The right image shows the same data represented as state-to-state migration flows. ©1981 The Ohio State University. Reprinted, with permission, from Tobler [1981].

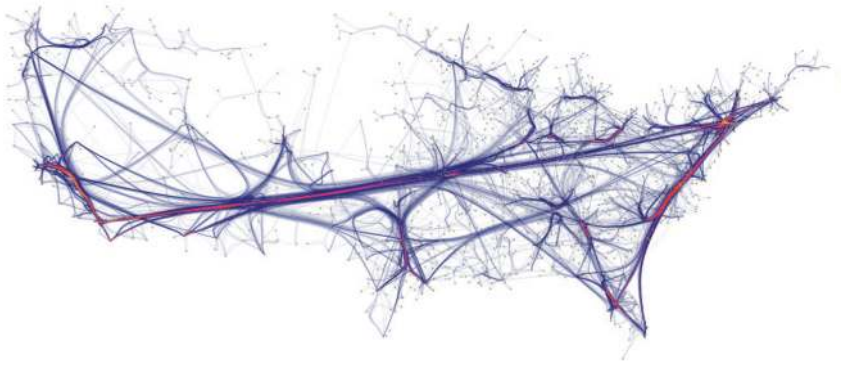
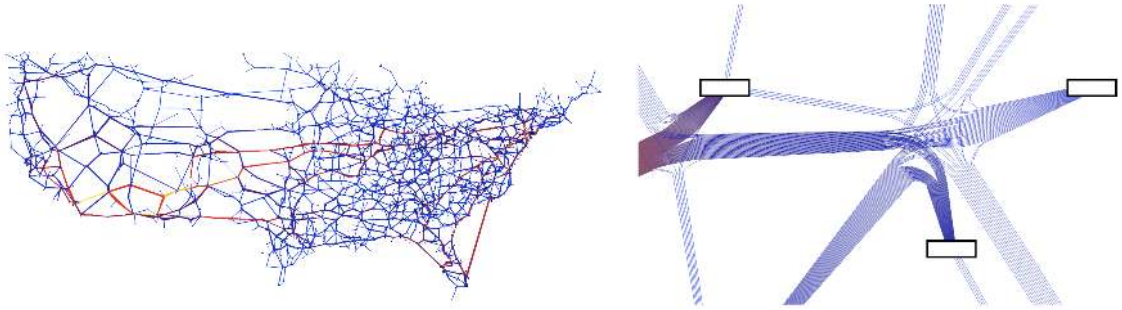


Figure 3.18: Force-directed edge bundling of the graph of US migrations. ©2009 The Author(s) Journal compilation ©2009 The Eurographics Association and Blackwell Publishing Ltd. Reprinted, with permission, from Holten and van Wijk [2009a].

has an advantage over bundling. After segmenting the thicknesses of the flow segments actually do accurately represent the aggregated amounts which are flowing between adjacent locations.

There is, however, at least one bundling algorithm which can be potentially useful for producing flow maps accurately representing the total magnitudes of the flows in bundles. This algorithm was very recently proposed by Pupyrev et al. [2012]. It bundles the edges of a graph in a way so that the individual edges within each bundle do not overlap and can actually be seen as separate edges when zooming in. Hence, the thicknesses of the resulting bundles correspond to the numbers of edges in them. An extension of this algorithm might be able to bundle edges of varying thicknesses so that the total thicknesses of the bundles would more accurately represent the total magnitudes of the flows in the bundles.

However, even in this form bundling would have disadvantages similar to those which segmenting into flows between adjoining regions has (see Section 3.5.6). First, rerouting the flow lines may mislead the user into thinking that the actual flows take the visualized routes. Second, it makes it more difficult to find correspondences between the actual origins and destinations of the flows, because as soon as a flow gets to be a part of a bundle it becomes very difficult to distinguish it from the other flows in the bundle.



a) b)

Figure 3.19: US migrations bundled with an approach proposed by Pupyrev et al. [2012]. All the individual flows in the bundles remain separate and are drawn as parallel as possible, hence, the thicknesses of the bundles represent the numbers of edges in them. Figure b) shows a fragment of the whole graph shown in a). ©2012 Springer-Verlag GmbH Berlin Heidelberg. Reprinted, with kind permission of Springer Science+Business Media.

3.6 Related mapping techniques

In this section we discuss map distortions and several other techniques which represent OD-data in maps, but differ from what we considered as flow maps in the previous sections.

3.6.1 Map distortions

Distorting map projections are sometimes used to make geographic maps more readable or to highlight the most important elements in them. Thus, in the flow map showing world migrations by Minard (Fig. 3.2) the shapes of the continents and countries were notably deformed so that the flow lines could avoid crossing the landmasses. This made the underlying map easier to read and helped to give the impression that the migrants are "flowing out" from the landmasses making the origins and the destinations of the flows less ambiguous.

Stefaner [2010] created an interactive geographic symbol map representing outgoing and incoming migrations to and from New York (see Fig. 3.20). This is a symbol map (the symbols are the circles representing geographic locations and their sizes represent the numbers of migrants). Like in a Dorling cartogram [Dorling, 1996] the country centroids are slightly repositioned to avoid the circles overlapping. Since the most origins and destinations of the moving New Yorkers are within the city area, New York is enlarged and put in the center of the view. The rest of the world is mapped with a distorting projection using a damped distance function, hence, very effectively utilizing the screen space.

Wood and Dykes [2008] proposed an algorithm for producing *spatially ordered treemaps* (see Fig. 3.21) which can be applied to a set of geographic territories to generate a space-filling layout consisting of squarified rectangles representing these territories. The algorithm attempts to produce a layout so that the positions of the rectangles are as close to the actual geographic positions of the territories they represent as possible. By nesting a minified copy of the same layout in each of the individual rectangles it is possible to create an effective representation of OD-data, as we will discuss in Section 5.1.

Brunet [1986] introduced the notion of *chorems* which are schematic representations of geographic areas, their properties and relationships. With chorems the area shapes are simplified and stereotypical graphical objects are used to represent their properties and relationships. Hence, all the details which are not important for conveying the main message are eliminated from the view. Following this approach De Chiara et al. [2011] discuss the use of chorems in the context of an integrated visual analytics system and present several flow maps among other examples.

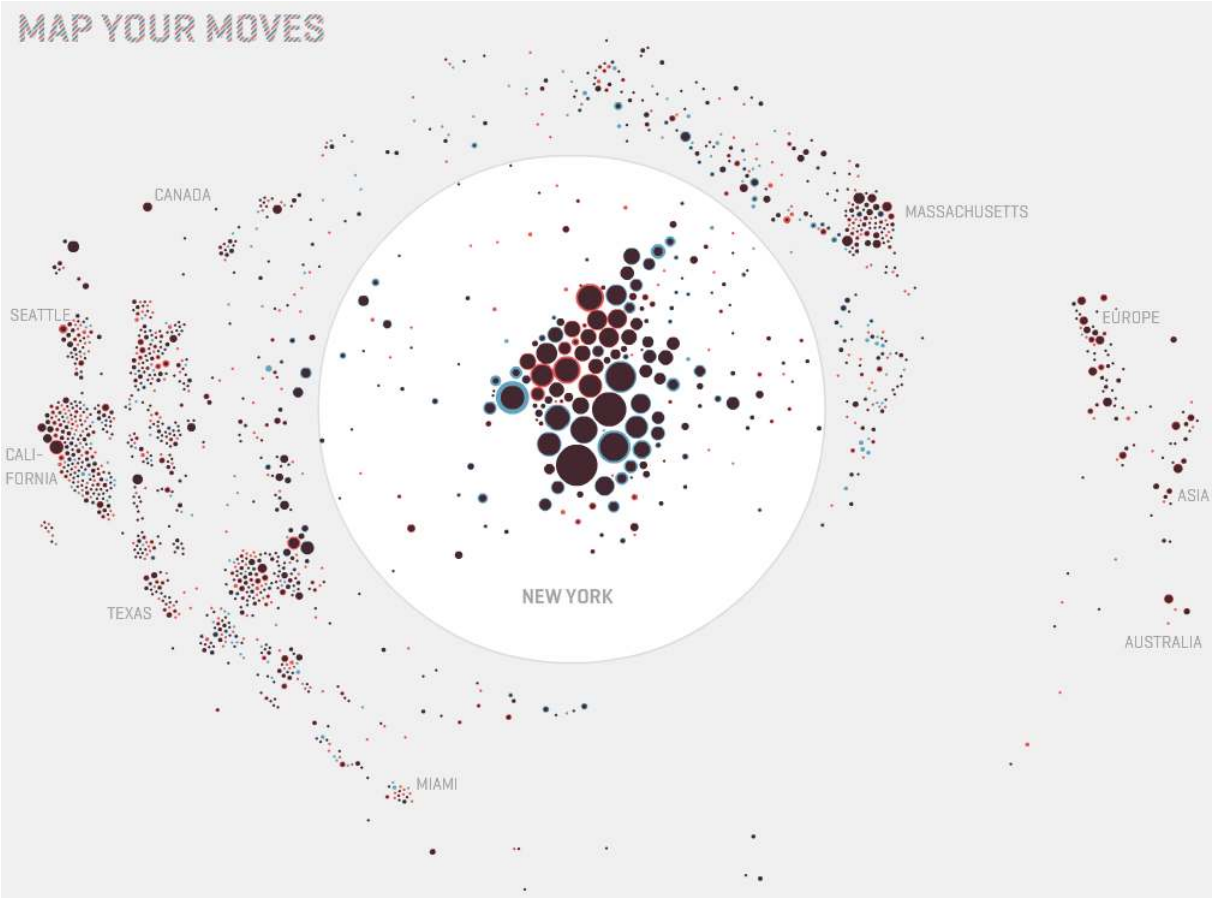


Figure 3.20: An online interactive geographic symbol map showing outgoing and incoming migrations to and from New York [Stefaner, 2010]. The world map is distorted magnifying New York and putting it in the center of the view.

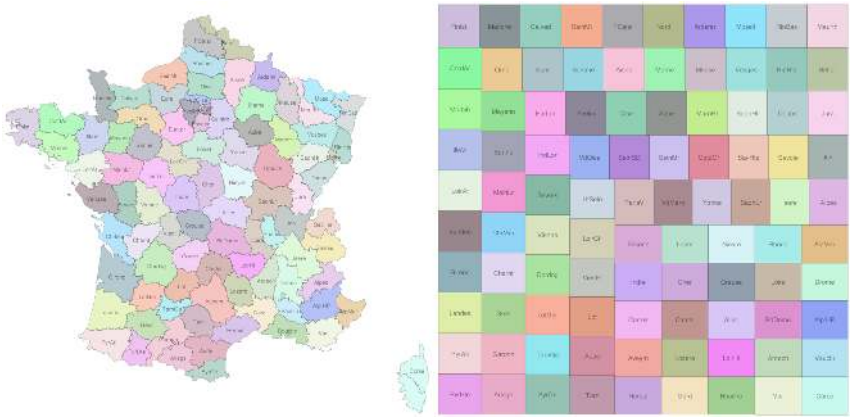
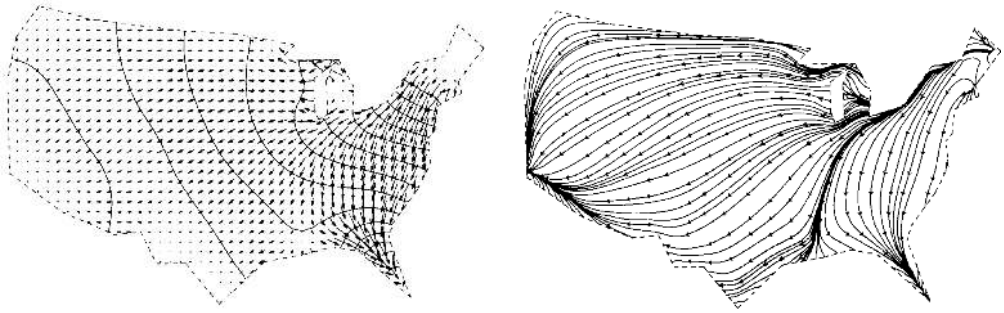


Figure 3.21: Using the spatially ordered treemap layout to produce a space-filling squarified map of France. ©2008 IEEE. Reprinted, with permission, from [Wood and Dykes, 2008].

3.6.2 Vector field maps

Ravenstein's map shown in Fig.3.16 was probably inspired by the meteorological approach to mapping winds and currents with the use of small arrows or pen strokes which stems from the 17th-century work of Edmond Halley [Tufte, 1986]. Ravenstein's map shows mostly local moves between adjacent counties

and is based on the idea that migration patterns are defined by many small movements, which form large-scale patterns when observed in combination.



a) b)

Figure 3.22: a) Estimated field and b) estimated trajectories (computed from the field a.) of the 1970-1976 net population flow in the US ©1981 The Ohio State University. Reprinted, with permission, from Tobler [1981].

Waldo Tobler was inspired by this idea and in [Tobler, 1981] proposed a method for producing spatially continuous map depictions of migration flows representing migration as a vector field on a map. Given an OD-dataset of migration flows Tobler's mathematical model can produce an estimation of a continuous vector field based on the actual migration flows (see Fig. 3.22). Like in meteorological wind maps Tobler used small arrows of varying sizes to portray varying magnitudes of the field vectors in different locations (another approach is to use arrows of the same size with varying density in the map [Lavin and Cerveny, 1987]).

Using a similar but discrete approach Thompson and Lavin [1996] created a computer program capable of producing an animated sequence depicting an estimated vector field of migrations (see Fig. 3.23). Each of the dots in this animation represents a vector of the field. During the looping animation each dot is moving a small distance in the direction of the vector it represents.

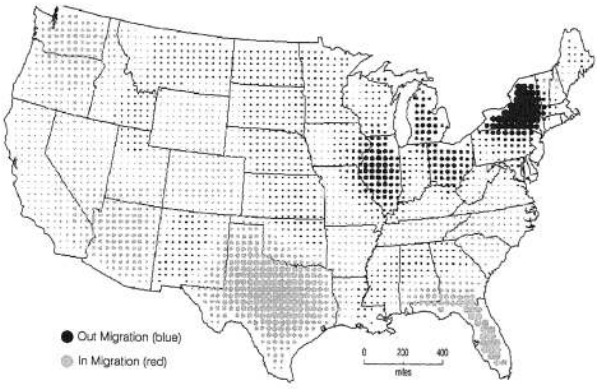


Figure 3.23: A frame from the animated US migration map represented as an estimated vector field of small migration movements. ©1996 University of Toronto Press. Reprinted, with permission from [Thompson and Lavin, 1996].

This method of representing discrete OD-data has not met widespread acceptance. One of the reasons for that is, probably, the different nature of OD-data compared to what vector fields are aimed to represent. Reading and comparing actual magnitudes of the flows between arbitrary origins and destinations in these maps is hardly possible unless the locations are adjacent. As with segmenting into flows between adjoining regions (see Section 3.5.6) it is very difficult to answer questions like "How many people migrated from A to B?" or "What are the major origins of migrants coming to A?".

3.6.3 Density plots

Using density plots is a way to avoid edge clutter in graph visualizations by not explicitly rendering edges [Van Liere and De Leeuw, 2003]. Instead of drawing edges as lines connecting pairs of nodes, the density of the line intersections is depicted in each pixel of the view. Rae [2009] advocates the use of this approach for representing OD-data (see Fig. 3.24).

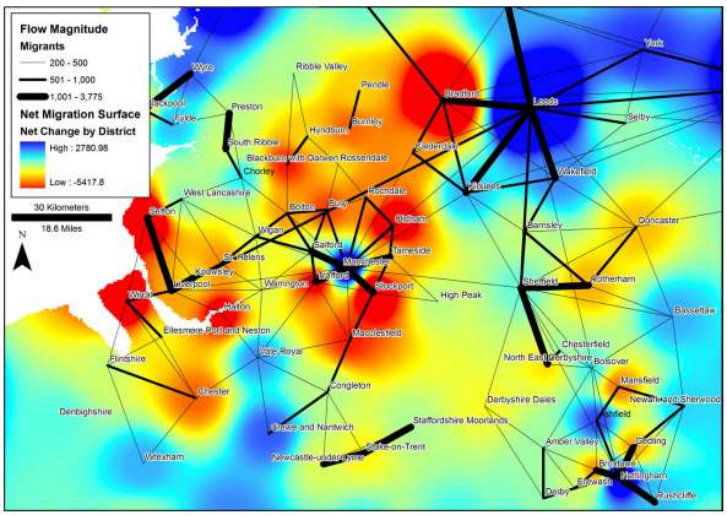


Figure 3.24: Density plot showing net migration of UK districts overlaid with an undirected flow map showing flows of a magnitude above a certain threshold. ©2009 Elsevier. Reprinted, with permission, from Rae [2009].

When used alone this technique has similar disadvantages as vector fields: it is hardly possible to see what the destinations of the flows of specific origins are and to read and compare actual magnitudes of the flows. Overlaying a density plot with the largest flows drawn as lines makes it easier to understand the main flow patterns. In this case the added value of the density plot comes down mostly to representing the net migrations of the locations and highlighting the parts of the map which are crossed by the most of the flows. In addition, as noted by [Wood et al., 2011], this technique assumes that the lines the density of which such plots depict are representative of the actual routes between the origins and destinations. In case of OD-data these routes are usually not known, therefore making this assumption can result in a misleading visualization.

3.7 Conclusion

In this chapter we discussed flow maps, the most widely used representation of OD-data. We considered various visualization techniques related to flow maps, tried to identify their advantages and disadvantages, analyzed the current state of research regarding the efficiency of these techniques, and talked about the problems of flow maps and the ways of addressing them.

It must be noted that in this chapter we did not discuss how temporal changes can be represented in flow maps. There are several ways of introducing the temporal dimension into an OD-data visualization and in Chapter 5 we discuss them. In Chapter 7 we present a user study in which we made a detailed comparison of the two most basic ways of representing temporal changes in flow maps: animation and small multiples.

Flow maps are arguably the most natural and easy-to-understand visualizations of OD-data. For this reason we believe that they will remain widely used with the advent of other "less natural" representations even if the latter prove to be more effective for certain tasks. However, not all of the shortcomings of flow maps which we discussed in this chapter have always been fully understood. Knowing the limitations

and ways of contending with them, understanding the tasks which flow maps address well and not so well and having a palette of techniques and alternative representations which can deal with the weaknesses by better addressing certain tasks can immensely help developers to avoid producing tools which leave users with cluttered, hard to interpret or even misleading visualizations.

On our way to this goal we introduce a taxonomy of tasks for OD-data analysis in the next chapter and in Chapter 5 we systematically describe the design space of OD-data visualizations including alternative representations.

Chapter 4

Analysis tasks

The goal of this chapter is to introduce a taxonomy of tasks which can be supported by tools for visual exploration of temporal OD-data. The taxonomy will help us to understand what kinds of questions can possibly be answered by analyzing temporal OD-data. We build this taxonomy by deriving the tasks from the components of the data. This way to identify the tasks is complementary to the user-centered approach which we take in Chapter 8.

4.1 Introduction

In the context of exploratory data visualization tasks are always concerned with information which needs to be obtained from the data. Essentially, tasks are questions which data analysts need to answer and visualization tools are built to support the analysts in finding answers to these questions.

Any tool which is intended to be useful for certain tasks must be deliberately designed in a way to support them. Understanding the tasks of the users derived from the problems which they address allows the developers of data analysis tools to acquire key requirements. However, acquiring this understanding is not always an easy task. Especially when it comes to data analysis and exploration, the task definitions which can be elicited from the users are often vague and imprecise and provide much space for interpretation.

Paradoxically, it is the vagueness of the tasks which actually makes it worth using exploratory visualization. Visualization might simply be the wrong approach when the tasks are crisply defined and all the necessary information is available and computerized. In such situations there is no real need for exploration and fully automatic solutions should be preferred [Sedlmair et al., 2012].

Despite that, tool creators must still explicitly consider tasks. The reason for this is that there are always real analysis tasks behind what explorers find to be useful information. Andrienko and Andrienko [2006] argue that an explorer usually does not look at data just for the purpose of looking at it, but *looks for something interesting* and this *interestingness* defines a relevance to the actual research questions that the explorer addresses by doing the analysis, even if the explorer is not consciously aware of the tasks and is guided by pure intuition in their exploration.

But how can data analysis tasks be discerned when the users are not fully aware of what they are looking for? The answer is that we can start from the data themselves. When the structure and the information content of the data in consideration is known we can build a taxonomy of tasks based on it. In this case the tasks do not necessarily describe what information the users *really need* to find, nor what they *think they need* to find, but what information *can* be found in the data, and this is a crucial step. Knowing the questions which *can* be answered by analyzing a specific kind of data makes it possible to deliberately choose those which *need* to be supported and design the tools accordingly. Making this deliberate choice of specific tasks to support is vital for the tool creators, because it is practically impossible to find one single visualization approach which works equally well for all kinds of tasks.

It must be clarified that what is said above does not mean that attempting to elicit the requirements for the tools from the users and domain experts is useless. In the contrary, it is indispensable for any tool designer to learn as much as possible about the requirements of the domain experts, in order to characterize and abstract the problem in the very early stages of the development. However, typically it is not sufficient to just talk to the target users, because what they say about their activities is only an incomplete match with what they actually do and most users do not accurately introspect about their data analysis and visualization needs [Sedlmair et al., 2012; Ericsson and Simon, 1980]. User-centered design addresses this problem by combining interviews with user observations [Sharp et al., 2007]. Having an understanding of the whole range of questions which can be answered through the analysis of a specific data type can help to further mitigate the problem. Combining this understanding with the user input makes it possible to choose out of all the possibilities those tasks which represent the best match with the needs expressed by the users, and thus, to develop tools which are as close as possible to fulfilling their real needs.

For our work, the main purpose of considering tasks for the analysis of temporal OD-data is to help to define criteria upon which the effectiveness of different exploratory visualization approaches for this specific kind of data can be evaluated which in the end can help data analysis tool creators to choose and design better visualizations for the tasks.

In the following sections we discuss task taxonomies for exploratory visualization proposed by other researchers and introduce a task taxonomy built for temporal OD-data.

4.2 Existing task taxonomies

Many different task taxonomies for visualization tools have been proposed so far. The researchers who develop such tools use different approaches and address different goals. In the next sections we will discuss some of them.

4.2.1 Shneiderman's task by type taxonomy

Shneiderman [1996] proposed a general taxonomy with the following tasks:

- Overview: Gain an overview of the entire collection
- Zoom: Zoom in on items of interest.
- Filter: Filter out uninteresting items.
- Details-on-demand: Select an item or group and get details when needed.
- Relate: View relations hips among items.
- History: Keep a history of actions to support undo, replay, and progressive refinement.
- Extract: Allow extraction of sub-collections and of the query parameters.

Shneiderman describes seven basic data types (1-D, 2-D, 3-D, multi-dimensional, time series, network, tree) and discusses techniques which can be applied to support each of the tasks depending on the type of the data under analysis. Despite its relation to the data types this taxonomy cannot be used to describe the kinds of information extracted from data as it only recounts the features which visualization tools should support.

4.2.2 Amar's low-level analysis tasks

Amar et al. [2005] introduce a low-level analysis task taxonomy with the goal of describing the analytic activity of the users. Amar's taxonomy encompasses the following tasks:

- Retrieve Value
- Filter
- Compute Derived Value
- Find Extremum
- Sort
- Determine Range
- Characterize Distribution
- Find Anomalies
- Cluster
- Correlate

The taxonomy was built by asking people how they would approach the analysis of specific datasets. Hence, it shows the variety of analytic questions typically posed by users when employing information visualization systems. This taxonomy is closer to describing the information needs of the users than Shneiderman's taxonomy, but it also says too little about the data under analysis.

4.2.3 ESRI common geographic analysis tasks

Mitchell and ESRI [1999] make a list of geographic analysis tasks claiming these to be the most common tasks which people involved in geographic analysis do every day in their jobs:

- Mapping where things are
- Mapping the most and least
- Mapping density
- Finding what is inside
- Finding what is nearby
- Mapping change

This taxonomy provides a bird's-eye view of the spatial exploratory analysis which also applies to OD-data, albeit a more detailed and structured approach can be used to systematize tasks.

4.2.4 Bertin's reading levels and question types

Bertin [1967] proposed a general approach for task systematization which has been used and refined by many researchers. Bertin distinguishes tasks by the *question types* and by the *reading levels*. The *question type* describes the kind of information sought, for instance:

Question	Question type
What is the origin of the largest flow of migrants to California?	Origin
How much money did World Bank donate to India in 2009?	Flow magnitude
In which year did the number of scientific collaborations between Norway and India start to grow?	Time

Bertin argues that as many types of questions can be asked about the data as there are components in the information. Hence, the question types should correspond to the structural components of the data under analysis, thus, encompassing questions concerning all the aspects of the data.

The *reading level* describes how many elements are concerned in the task. Bertin distinguishes three reading levels:

Elementary – Questions concerning one single element

Intermediate – Questions concerning a group of elements

Overall – Questions concerning all the elements together.

The virtue of Bertin's approach is that the task taxonomy is systematically derived from the data. Thus, the derived tasks better represent the data under analysis and allow us to argue that the taxonomy is complete (in the sense that the questions it defines encompass all the components of the data).

4.2.5 Peuquet's typology of queries for spatio-temporal data

Peuquet [1994] proposes a typology of queries for spatio-temporal data based on their triad representation framework consisting of the three basic components: "what", "where" and "when". The types of questions in this typology are the following:

> **when + where → what** : Describe the objects or set of objects (what) that are present at a given location or set of locations (where) at a given time or set of times (when).
>
> **when + what → where** : Describe the location or set of locations (where) occupied by a given object or set of objects (what) at a given time or set of times (when).
>
> **where + what → when** : Describe the times (when) that a given object or set of objects (what) occupied a given location or set of locations (where).

Peuquet also distinguishes types of spatio-temporal questions concerning time and changes by their level of analysis:

- Questions addressing changes in an object or feature
- Questions addressing changes in the spatial distribution or set of objects
- Questions addressing the temporal relationships among multiple geographic phenomena.

This typology is very much in line with Bertin's question types and reading levels. Hence, it can be considered an application of Bertin's approach to spatio-temporal data.

4.2.6 MacEachren's aspects of time

MacEachren [2004] more thoroughly describes various aspects of time which can be depicted on a geographic map and associates them with analysis questions:

- Existence of an entity: *if* the entity exists at the specified time?
- Temporal location: *when* the entity exists?
- Time interval: *how long* is the time span from beginning to end of the entity?
- Temporal texture: *how often* does an entity occur?
- Rate of change: *how fast* is an entity changing or *how much difference* is there from entity to entity over time?
- Sequence: in *what order* do entities appear?
- Synchronization: do entities occur *together*?

4.2.7 Blok's change analysis tasks

Blok [2000] discusses change monitoring in geo-spatial context and differentiates between two basic types of tasks which are supported by visual exploration: *identification* and *comparison*. The latter can actually be concerned with finding any kind of relationships including causal ones. Blok also argues that different questions can be asked depending on the length of the time series under analysis (this

is analogous to Bertin's reading levels). Taking these two dimensions into account she proposes the following taxonomy:

Tasks \Time series	Short	Longer
Identification	Change? What developments?	What process? Trend?
Comparison	Changes? Anomalies?	Relationships/causes? Anomalies?

4.2.8 Andrienko's approach to task systematization

Andrienko and Andrienko [2006] propose an approach for systematizing tasks for exploratory analysis of spatio-temporal data which is based on Bertin's question types and reading levels, but with some important differences.

The type of the question is defined in terms of its *focus* and *target*. As the approach primarily addresses spatio-temporal data, the *focus* is one of their three fundamental components: objects, space and time [Andrienko et al., 2011]. This is the same idea as in Peuquet's triad framework [Peuquet, 1994]. Hence, there are three possible focuses:

- *focus on objects* (movers, events, trajectories): characteristics of objects in terms of space and time; relations to locations, times, and other objects;
- *focus on space*: characteristics of locations in terms of objects and time; relations to objects, times, and other locations;
- *focus on time*: characteristics of time units in terms of objects and space; relations to objects, places, and other time units.

The *target* is a characteristic or relation of the focus which the question addresses. For instance, for the question "What is the origin of the largest flow of migrants to California?" the focus is "Flow event" and the target is the "Origin" of this flow event.

Instead of distinguishing between three levels of analysis as Bertin does (*elementary*, *intermediate*, and *overall*), Andrienko define only two analysis levels (*intermediate* and *overall* are combined together into *synoptic* tasks):

- *Elementary tasks*: Focus on one or more elements of a set with their particular characteristics and relations.
- *Synoptic tasks*: Focus on a set of elements as a whole or its subsets as wholes, disregarding individual elements (dealing with sets as wholes implies making a sort of synopsis concerning these sets, hence the name *synoptic*).

Andrienko further detail the task systematization by making distinctions between several *classes* of elementary and synoptic tasks:

Elementary tasks

- *Lookup*: Find values of some data components that correspond to given values of other components.
- *Comparison*: Compare characteristics to specific values or between different references.
- *Relation seeking*: Find references (e.g. dates) for which specific relations exist between the attributes corresponding to the references.

Synoptic tasks

- *Pattern definition*: Assigning a pattern to a particular type, summarization of characteristics.
- *Pattern search*: For a specified pattern find subsets of references such that the behavior over those subsets corresponds to this pattern.
- *Pattern comparison*: Differentiation between behavior fragments.
- *Relation seeking*: Looking for major contrasts, changes, and discontinuities; detection of outliers and deviations from the major trend.

This approach allows building a systematic and exhaustive taxonomy of tasks for a specific type of datasets as long as the focuses and the targets enclose all characteristics and relations of the data [Andrienko et al., 2011]. Having an exhaustive task taxonomy makes it possible to analyze visualization tools judging their effectiveness based on the tasks they support.

4.3 Taxonomy of tasks for temporal OD-data analysis

We decided to apply the approach proposed by Andrienko and Andrienko [2006] to building a taxonomy of tasks for the analysis of temporal OD-data. There are three main reasons for choosing this approach:

- First, it is the most systematic and detailed to the date.
- Second, it allows building a taxonomy by deriving it directly from the data model.
- Last, but not the least, it results in a complete taxonomy which encompasses questions concerning all of the components of the chosen dataset (or a type of datasets).

In this section we present the task taxonomy for temporal OD-data analysis which we built by applying the approach and discuss the analysis questions which it encompasses.

4.3.1 Scopes of elementary and synoptic tasks

From the data model defined in Chapter 2.1.1 we can discern the following main structural components of temporal OD-data which constitute the set of possible focuses for the analysis tasks:

- Flow event
- Origin
- Destination
- Time

For each of these components there are several targets representing the characteristic of the component which the task addresses. A combination of the focus and target defines the *scope* of questions addressed by the tasks. The scopes of temporal OD-data analysis questions which can be asked for each particular focus-target combination are shown in Table 4.1.

Flows in OD-data are usually directed, hence, there is an important difference between the origins and the destinations. Therefore, the tasks focusing on "origin/destination" are actually two separate groups of tasks. However, we combined them together, as most of them are essentially the same. Where we used the word "location" it should be replaced with either "origin" or "destination" depending on the focus of the task. The "Opposite location" for tasks focusing on "origin" is "destination" and the other way around. The tasks which concern the assymetry of the flows of the opposite directionality between particular locations (e.g. "How does the magnitude of A→B compare to the magnitude of B→A?") are still covered by the scope focusing on multiple "origin/destination" locations with the target "total magnitude".

Focus	Target	Elementary task	Synoptic task
Flow event	Time	Positions of flow events in time	Temporal distribution of flow events
	Origin	Origins of particular flow events; their locations	Spatio-temporal distribution of origins of flow events
	Destination	Destinations of particular flow events; their locations	Spatio-temporal distribution of destinations of flow events
	Magnitude	Magnitudes of particular flows	Spatio-temporal distribution of flow magnitudes
	Distance	Distances between origins and destinations of particular flow events	Spatio-temporal distribution of flow distances
	Context	Spatio-temporal context of this event	Relation of the spatio-temporal distribution of flow events to the context
	Relations between flow events	Relations between particular flows events	Relations between flow events; spatio-temporal distribution
Origin/destination	Flow events	Flow events of particular locations	Patterns of the distribution of the flow events of locations
	Time	Time of presence of flows from/to particular locations	Temporal patterns of the flow presence for locations
	Opposite location	Opposite locations (destinations/origins) of the flows of particular locations	Distribution of the opposite locations of flows (e.g. clusters of destinations for some origins); its change over time
	Distance	Distances of the flows of particular locations	Spatio-temporal distribution of the flow distances
	Total magnitude	Total magnitude of the flows for particular locations	Spatio-temporal distribution of the total magnitudes of the flows of locations
	Number of flows	Number of the flows of particular locations	Spatio-temporal distribution of the numbers of the flows
	Context	Relations to the spatio-temporal context	Spatio-temporal distribution of locations relate to context
	Relations between locations	Relations between particular locations (e.g. same origins/destinations, correlated changes of total magnitudes)	Relations between locations; their change over space and time
Time	Flow events	Flow events of particular moments in time; their spatial configuration	Changes over time of the flow events and their spatial configurations
	Origins/destinations	Origins/destinations of the flows of particular moments in time	Patterns of change over time of the origins/destinations of flow events
	Context	Context of particular moments in time; relation of flow events of particular moments in time to their contexts	Relation of spatio-temporal distribution of flow events to their temporal contexts
	Relations between moments in time	Relations between particular moments in time (e.g. how do the flows differ, what are the origins/destinations)	Specific relations between different moments in time (e.g. moments which are characterized by significant changes of flow magnitudes); their distribution over space and time

Table 4.1: Scopes of elementary and synoptic tasks for temporal OD-data analysis depending on the tasks' focus and target. Based on the types of tasks table from [Andrienko et al., 2011], but adapted to temporal OD-data.

Classes of elementary tasks	Classes of synoptic tasks
Lookup	**Pattern identification**
Direct: "What was the magnitude of the flow A→B in year N?"	*Pattern definition*: "What was the trend of the magnitude of A→B in years N to M?"
Inverse: "When did the magnitude of the flow A→B exceed X?"	*Pattern search*: "In what time intervals was the magnitude of A→B growing?"
Comparison	**Pattern comparison**
Direct comparison	*Direct comparison*
With attribute value: "Did the magnitude of A→B exceed X in year N?"	With pattern: "Was the magnitude of A→B growing in years N to M?"
Between references: "How did the magnitude of A→B change from year N to M?"	Between references: "What was the distribution of the magnitudes of the flows originating in A in year N compared to year M?"
Between attributes: "What was the difference between the total magnitudes of incoming and outgoing flows of A in year N?"	Between attributes: "How does the distribution of destinations of flows from A differ from the distribution of the origins of flows to A?"
Inverse comparison	*Inverse comparison*
With reference: "Did the total magnitude of the flows from A exceed X before or after year N?"	With reference sets: "How is the trend of growing outgoing flows related to the political situation?"
Between references: "How do dates on which the total magnitude of flows from A exceeds X and goes below Y compare (which was before, what's the interval)?"	Between reference sets: "How are periods of growing outgoing flows related to periods when they were decreasing?"
Relation-seeking	**Relation-seeking**
Between references and between attributes: "Are there any two locations which had large incoming flows from A simultaneously?"	Between reference sets and between attribute patterns: "Are there any two locations which during different time intervals had the same trend of total magnitude change?"
Between reference characteristics: "When did the total magnitude of outgoing flows from A exceed the total magnitude of year N?"	Between attribute patterns over subsets of a reference set: "In what time interval was the trend of the total outgoing flow magnitude opposite to that in the given interval?".
Between references: "In which locations did the total magnitude of the outgoing flows grow from year N to year M?"	Between attribute patterns over different reference sets: "In which regions did the configuration of the flow origins-destinations change significantly during years N and M?"
Between attributes: "In which locations did the total magnitude of outgoing flows exceed the total magnitude of incoming flows."	Between patterns of different attributes: "In which time periods were the trends of the total magnitudes of the outgoing and incoming flows the same?"

Table 4.2: Example tasks illustrating the classes of elementary and synoptic tasks for the analysis of temporal OD-data. Based on the table showing correspondence between classes of elementary and synoptic tasks from [Andrienko and Andrienko, 2006], but adapted to temporal OD-data. The notation A→B should be read as "Flow from A to B".

It has to be mentioned that it is possible for a task to be elementary with regard to one data component and synoptic with regard to another component. For instance, when looking at the distribution over time of the destinations of the flows from a specific origin, the task is elementary with respect to "origin" and synoptic with respect to "time" and "destination". To distinguish between them in the rest of the thesis we will use the following terms:

Spatio-temporal synoptic task
Tasks which are synoptic in respect to "time" and one or both of "origin" and "destination".

Temporal synoptic task
Tasks which are synoptic in respect to "time" and elementary in respect to both "origin" and "destination".

Spatial synoptic task
Tasks which are synoptic in respect to "origin" and/or "destination" and elementary in respect to "time".

4.3.2 Classes of elementary and synoptic tasks

Each of the task scopes shown in Table 4.1 represents a number of questions of different *classes* (see Section 4.2.8). Providing a complete taxonomy of tasks for each focus-target-class combination would take too much space, therefore in Table 4.2 we show example questions of different scopes which illustrate the classes of elementary and synoptic tasks. We will presume, though, that the complete taxonomy incorporates tasks for all valid focus-target-class combinations.

To illustrate the relation between the two tables, let us consider the first question in Table 4.2: "What was the magnitude of the flow A→B in year N?" which belongs to the class of elementary direct lookup tasks. To find the matching cell in Table 4.1 for this task we identify its focus as "Flow event" and the target as "Magnitude", therefore the scope of this elementary task is "Magnitudes of particular flows".

The task scopes describe the data components and their characteristics which the questions consider and address while the classes outline the nature of the tasks from their operational perspective.

4.4 Conclusion

In this chapter we discussed various approaches to building taxonomies of tasks for exploratory data analysis, and presented our own task taxonomy for temporal OD-data.

We applied the approach proposed by Andrienko and Andrienko [2006] which is based on the ideas of Bertin [1967] and influenced by Peuquet [1994] and Blok [2000]. To build the taxonomy we started by taking the main components of the data model which we discussed in Chapter 2.1.1, then discerning the basic question types and targets, identifying the elementary tasks, which deal with individual elements, and the synoptic tasks dealing with sets of elements as wholes. The resulting taxonomy covers the questions which can be answered by analyzing temporal OD-data.

We did not provide the complete list of the analysis questions the taxonomy encompasses because it would require too much space and would not be very useful anyway. Instead, we provided the full list of the scopes of these questions and their classes. The combination of the two defines the structure of the full taxonomy. This structure is defined in accordance with the methodology described by Andrienko et al. [2011], and our contribution is merely in applying this methodology to temporal OD-data.

This taxonomy allows making judgments about the effectiveness of different approaches to the visualization of temporal OD-data. Besides, it can help to design new visualization tools by m"aking tool developers aware of the tasks which can possibly be supported and by letting them to make a deliberate choice of tasks to support. We will refer to this taxonomy in the rest of the thesis when appropriate.

Chapter 5

Design space exploration

Flow maps are by far not the only way of representing OD-data. There are several alternatives which we talk about in this chapter. We use a systematic approach to describe the design space of temporal OD-data visualizations and identify the analysis tasks for which each of the alternatives may be suited best. This allows us to give a number of recommendations for the choice of design alternatives depending on the tasks which must be supported in the first place.

In this chapter we investigate the main design alternatives for the visualization of OD-data and temporal OD-data. We attempt to systematically explore the space of the possibilities and identify some of the main principles upon which the existing approaches and techniques rely. We discern aspects of OD-data important for the analysis and find the approaches which are better suited for representing each of these aspects. This also lets us to identify the analysis tasks which are better supported by the techniques depending on the way they are fundamentally built and allows us to formulate design choice recommendations. The results of this exploration can be of use for developers of analytic tools for choosing the most appropriate technique for the tasks their tools must support as well as for researchers devising new visualization techniques.

5.1 Visualizing OD-data

In this section we provide a summary of the existing techniques for visualizing OD-data in order to make a detailed comparison of their characteristics and the particular aspects of OD-data they represent. First, we consider the existing techniques for non-temporal OD-data and then discuss the ways of adding the temporal dimension to them.

Flow map

Flow maps are node-link diagrams in which the flows are represented as lines connecting locations on a geographic map. The varying magnitudes are shown with the widths of the flow lines. Flow maps were thoroughly discussed in Chapter 3, therefore, we only provide another example of a flow map in Fig. 5.1.

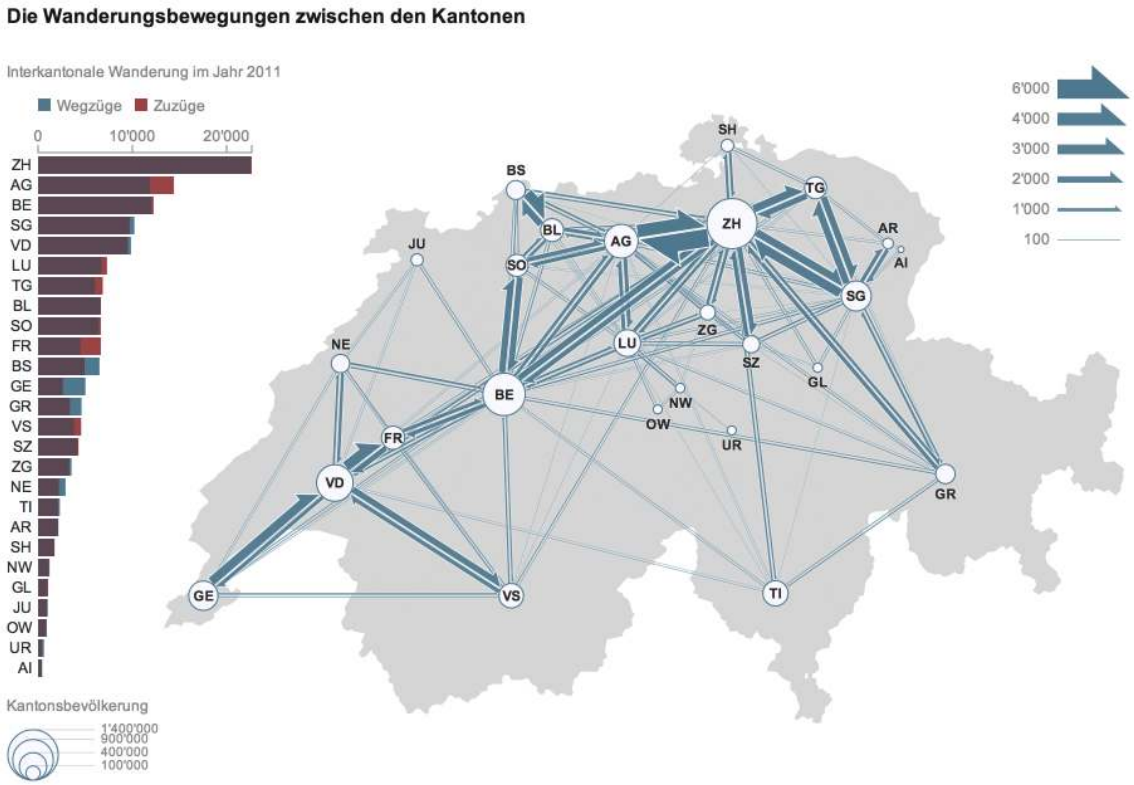


Figure 5.1: Flow map showing migrations between the Swiss cantons. The bar chart to the left shows the total incoming and outgoing flow magnitudes for the cantons, and the circle sizes represent the populations of the cantons. (Image produced by the author of this thesis for *Neue Zürcher Zeitung* during his work at *Interactive Things*).

Chord diagram

Chord diagrams discard the geographic positioning and use radial layout for representing locations, i.e. they are placed on a circle. Flows are shown as lines of varying widths connecting the nodes. Each flow line serves to show the magnitudes of the flow in both directions, as the line thickness on the two ends can be different. The flows of the same nodes are grouped in a way which allows us to see the total magnitudes of the flows of each node. A reordering of the nodes along the circle can be applied to correspond with geography in some way, for example, by grouping locations which are in a neighborhood as in Fig. 5.2. Chord diagrams are often easier to read than flow maps because they use screen space more efficiently by distributing the nodes along a circle.

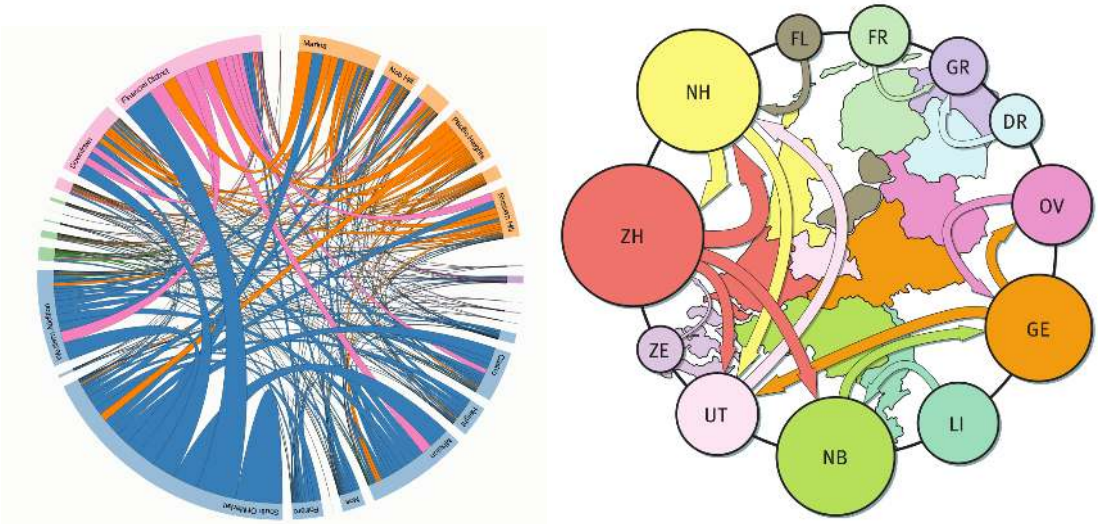


Figure 5.2: (left) Chord diagram showing migration between San Fransisco districts (interactive visualization by Bostock [2012]); (right) Necklace maps showing population of the provinces in the Netherlands and the relocation flows between the provinces ©2011 IEEE. Reprinted, with permission, from [Speckmann and Verbeek, 2010].

Speckmann and Verbeek [2010] proposed a variation of chord diagram based on the use of Necklace maps which attempt to maintain a better connection with the geography (Fig. 5.2). However, because of the use of color coding, this approach only works well for a relatively small number of nodes.

Arc diagram

In arc diagrams nodes are placed on a straight line and are connected by circular arcs representing connections between the nodes. As in flow maps varying the thicknesses of the arcs can be used to represent the magnitudes of the flows. The nodes are usually sorted in some meaningful way, e.g. to minimize crossings or to represent geographical distances from one selected location.

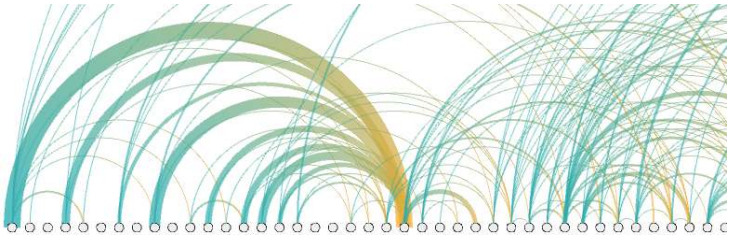


Figure 5.3: A fragment of an arc diagram visualizing rides between bus stops in Singapore using color gradient to show flow directions. (Image by the courtesy of Till Nagel).

Sankey arcs

This is a modification of the arc diagram recently proposed by Nagel et al. [2012] using a similar approach as Sankey flow maps which we discussed in Section 3.3.3. Instead of drawing all of the arcs directly from the node centers, their ends are placed adjacent to each other. As a consequence, the total magnitudes of the outgoing and incoming flows of a node can be easily compared across the nodes (see Fig. 5.4). Note that the chord diagram in Fig. 5.2 also uses this approach to enable comparison of the node totals.

Figure 5.4: Sankey arc diagram visualizing rides between bus stops in Singapore. (Image by the courtesy of Till Nagel).

OD-matrix

In Section 2.1 we discussed OD-matrix as an approach to represent OD-data in the memory of a computer. The data is represented in a matrix, the rows of which correspond to the origins, the columns – to the destinations, and the cells – to the magnitudes of the flows between the respective origins and destinations. The same approach to structuring OD-data can be used to visualize them. Fig. 5.5 shows a heatmap representation of an OD-matrix (the colors of the matrix cells represent the flow magnitudes). An important advantage of this way of depiction of OD-data is that it does not involve drawing flows lines, hence, it is completely devoid of clutter cased by the line crossings in node-link diagrams. This is less of an advantage, however, if the matrix is sparse, that is, if it is populated primarily with zero magnitudes. In this situation a flow map might be more effective, as it will show the geographic distribution while not having too much clutter.

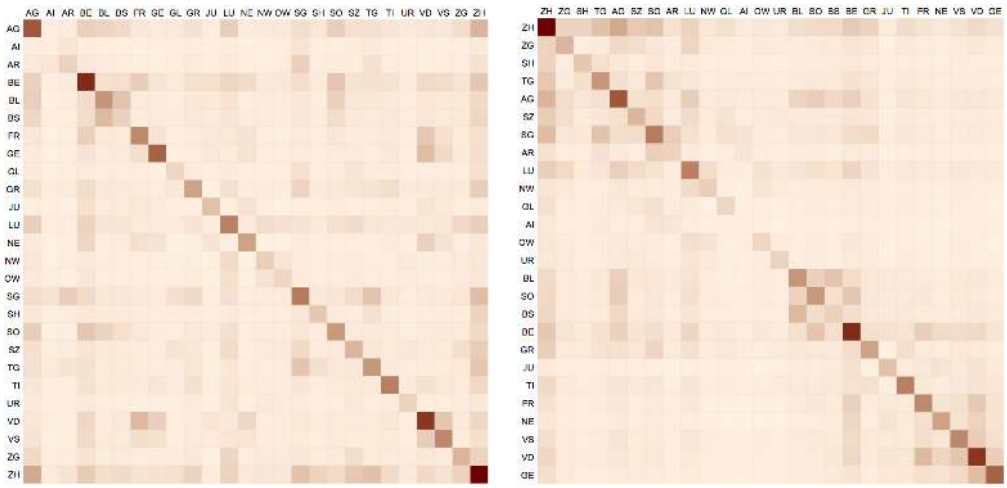


Figure 5.5: OD-matrix showing migrations between the cantons in Switzerland. In the left matrix the rows and columns are ordered alphabetically, in the right one – by the distances to the canton centroids from Zurich. (Image produced by the author of this thesis).

Bertin [1967] introduced the notion of *reorderable matrix* representation. Either manually or using an algorithmic approach rows and columns of a matrix can be reordered in various ways to better represent structural patterns of the data. Automatic approaches often involve algorithms allowing to generate permutations placing similar rows and columns closer together [Chen et al., 2004]. This requires introducing proximity measures defining the similarities between rows and columns. The process of finding good row and column permutations which reveal clusters in the data is sometimes called *seriation*, and there is a large number of methods described in the literature for solving this problem (e.g. Robinson matrix, elliptical seriation [Wu et al., 2008; Hahsler et al., 2007]).

A specific ordering can also be used to represent an intrinsic property of the nodes. For example, Becker et al. [1995] discuss an OD-matrix representation of network traffic in which the rows and columns are ordered according to the West-to-East geographic positions of the corresponding origins and destinations. In Fig. 5.5 (the right matrix) the rows and columns are ordered by their distances from a selected location, namely, from the centroid of canton Zurich. Even this simple ordering reveals several clusters which are not visible with the alphabetic sorting (the left matrix in Fig. 5.5). A good matrix reordering can reveal high-level structure in the same matrix representation which provides a very detailed view of the data [Fekete and Henry, 2009].

OD-treemap

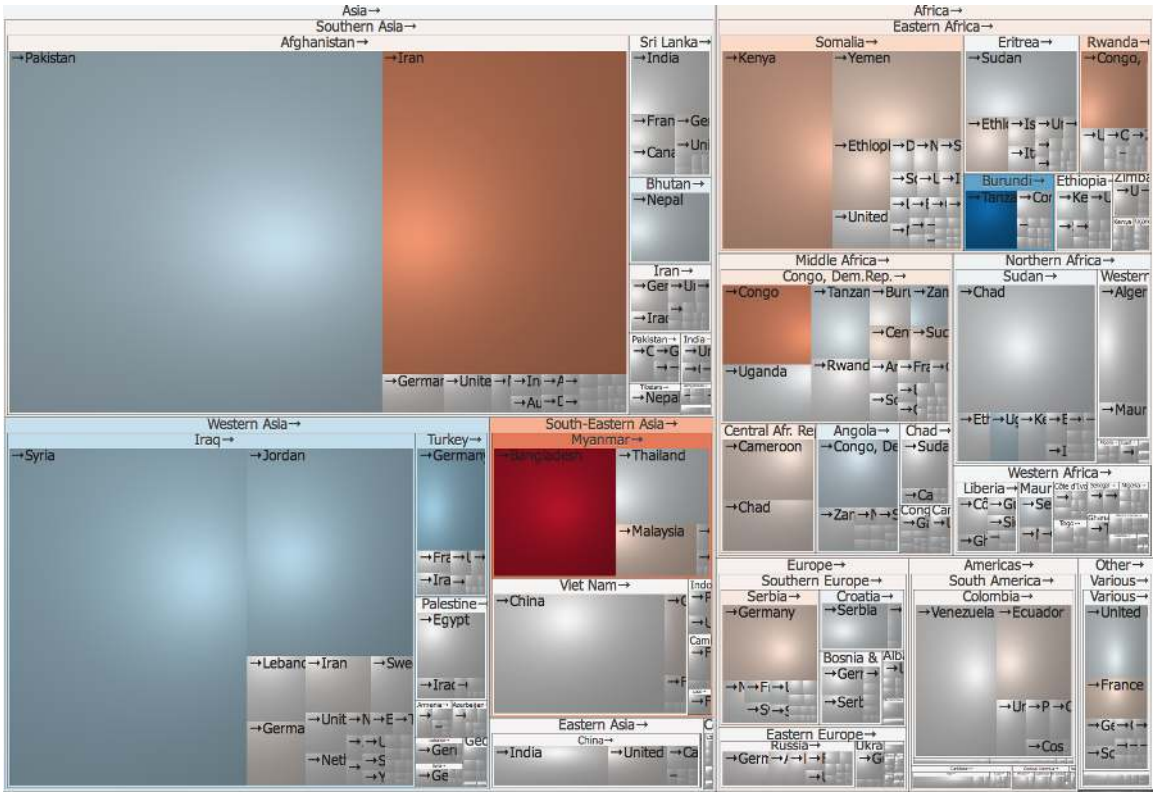


Figure 5.6: OD-treemap showing the magnitudes of the refugee flows between the world's countries. The sizes of the rectangles represent the magnitudes of the flows in 2009, the color shows the changes of the numbers of refugees between 2008 and 2009 (red is positive, blue is negative). The position of the arrow in the labels of the rectangles indicates whether the location is an origin or a destination. (Image produced with the help of Luc Girardin and TreeMap software [Brodbeck and Girardin, 2012]).

OD-treemap is a kind of a treemap [Shneiderman and Wattenberg, 2001] in which the hierarchy is defined by the relation between the origins and destinations, that is, the origins are the parents, the

destinations of the flows from a specific origin are the children of this origin in the hierarchy. Thus, in OD-treemap any rectangle corresponding to an origin contains all the rectangles corresponding to its destinations. A reverse hierarchy in which the destinations are the parents is also possible. Additional levels of hierarchy can be easily introduced into the treemap by spatially grouping the locations, for example, into regions as in Fig. 5.6.

Map^2

Map^2 proposed by Guo et al. [2006] employs the idea of OD-treemap nesting, but instead of using the area of the rectangles to represent the total flow magnitudes of the locations it uses equally sized rectangles, but arranges them in a way so that their positions resemble the actual geographic placement. It displays separate maps corresponding to each of the origins, and each of these small maps shows the magnitudes of the flows from the origin the map represents to different destinations. This is an example of a very effective use of small multiple display.

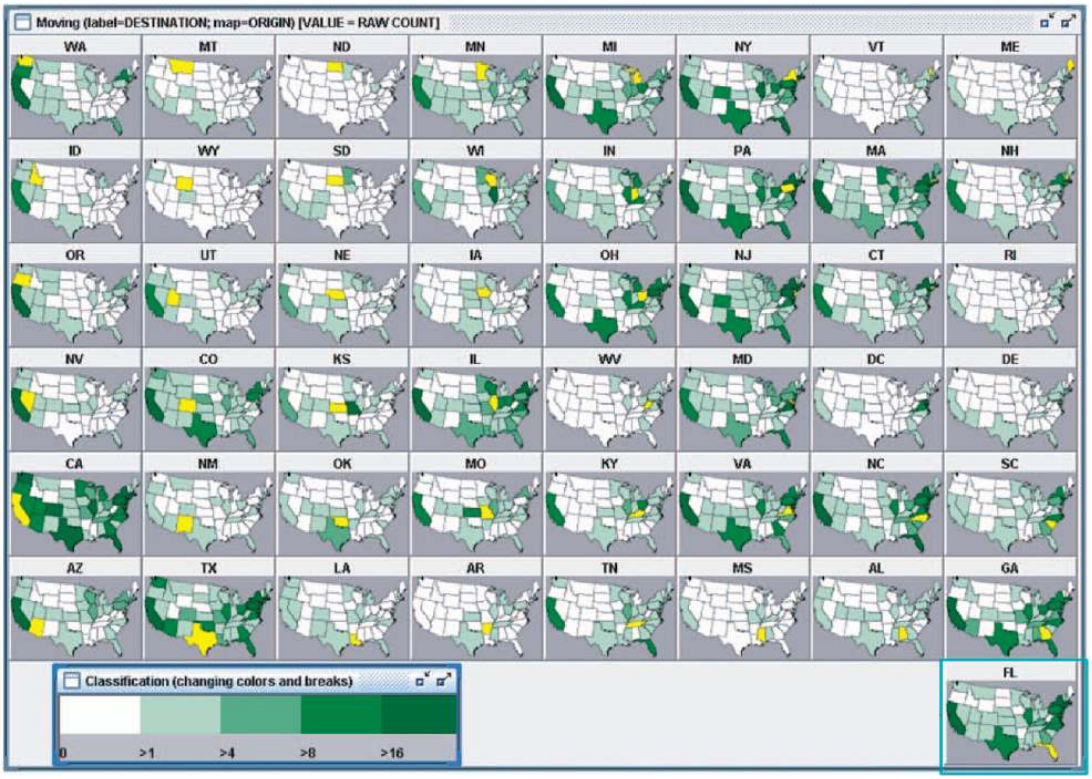


Figure 5.7: Map^2 visualization of US companies which relocated from one state to another. ©2006 IEEE. Reprinted, with permission, from [Guo et al., 2006].

OD-map

Wood et al. [2010] bring the idea of nesting employed in Map^2 and OD-treemap even further. As in Map^2 they use a small multiple display consisting of small maps representing each of the origins and each small map shows with the color the magnitudes of the flows from the origin it represents to different destinations. However, in OD-maps it is essential that the large map is divided into cells with a regular grid in exactly the same way as the small maps nested in it. Each grid cell becomes a new location and if multiple locations of the original data fall into the same cell, the flows of these will be aggregated. This way the layout of the rectangles on the both levels of the hierarchy strictly corresponds to the actual geographic layout. The rectangles are actually cells of a grid drawn on a geographic map.

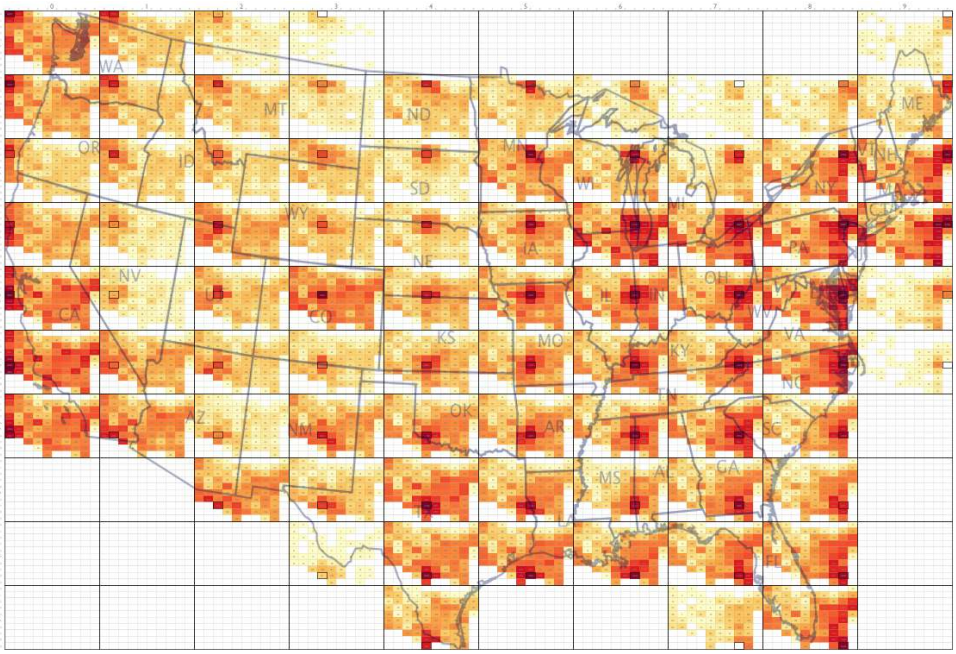


Figure 5.8: OD-map representing about one million migration flows between the US counties. The approach is described in [Wood et al., 2010].

However, grouping together locations which fall into the same grid cells makes it impossible to see the flows of the individual locations in each group. This can be detrimental in situations when locations which should be considered separately fall into the same grid cell. To avoid such situations an approximate geographic layout can be used where each location of interest has a dedicated grid cell, and the cells are arranged so that the final layout resembles the actual geographic layout. To produce such a layout an automatic algorithm can be used, for instance, spatial treemap layout which we discussed in Section 3.6.1.

Hive plot

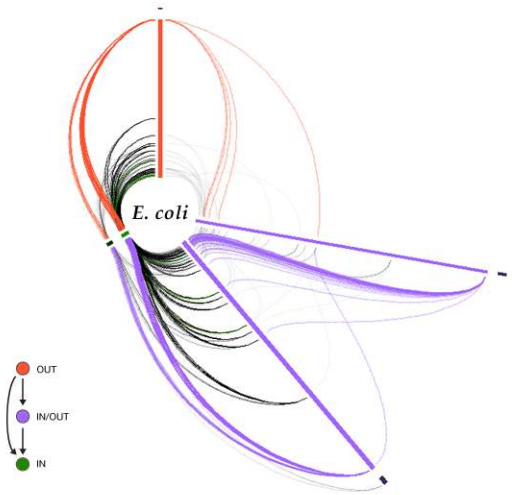


Figure 5.9: Hive plot showing *E. coli* gene regulatory network. (Image by the courtesy of Martin Krzywinski).

Hive plot proposed by Krzywinski et al. [2011] is an attempt to find a better alternative for cluttered node-link diagrams. It has multiple axes on which nodes are placed depending on a category they belong to. Same nodes can be placed on multiple axes at the same time if they belong to multiple categories. The connections between the nodes are still drawn as lines as in node-link diagrams, but because of the node placement on the axes, the resulting representations are often less cluttered. The number of the axes can vary and each axis has a special meaning chosen by the chart designer, so that positioning the nodes on the axes conveys important information about the nodes. For instance, for OD-data it could make sense to use three axes: one for the nodes which have only outgoing flows, for the nodes which have only incoming flows, and for the rest. The nodes on each of the axes can be grouped by the geographic proximity, or can be sorted by the total magnitudes of their flows. Flows can then be drawn as lines of varying widths connecting the locations.

Symbol map

A map with symbols (e.g. bubbles as in Fig. 5.10) plotted over it showing the total incoming and outgoing flow magnitudes for each of the locations. The bubbles showing incoming and outgoing flows may be distinguished by their color: one color corresponds to the incoming flows, another to the outgoing ones. The actual flows are not shown, therefore there is no cluttering problem. To see the actual flows between a specific origin/destination pair the user has to interact with the view by selecting a node. When a node is selected the flows of this node can be either shown as lines in a flowmap, or "filtering" is applied the view, that is, the circle sizes are updated to only represent the flows to and from the selected node.

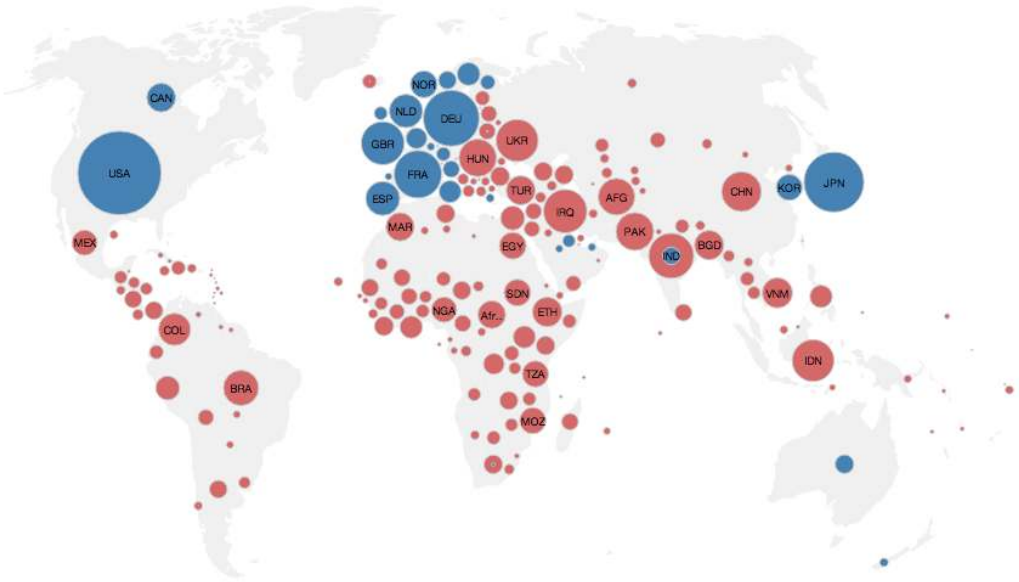


Figure 5.10: Symbol map view showing AidData donors (blue) and receivers (red). (Image produced by the author of this thesis).

O and D symbol maps

Same as symbol map, but using two separate maps: one for the origins, another for the destinations with bubbles representing the totals for each of the locations. Thus, in the origins map the totals for the outgoing flows are shown, in the destinations map – for the incoming. As bubble map this view also requires user interaction to see the individual flows. When the user selects a node, the opposite map can be updated to show only the flows from or to the selection.

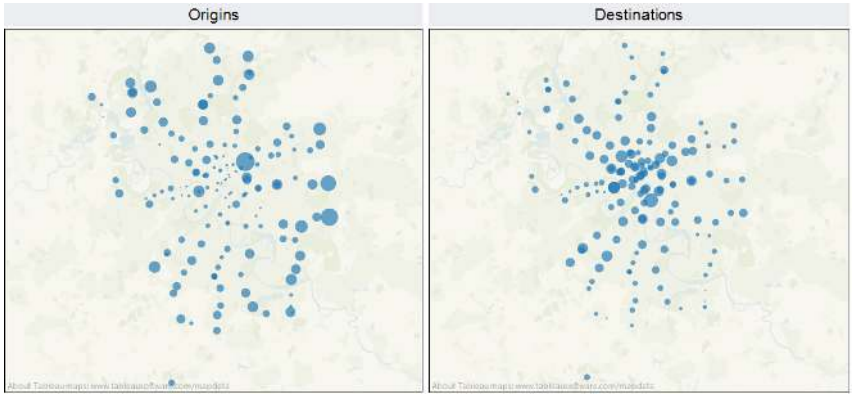


Figure 5.11: O and D symbol maps representing Moscow metro passenger rides at 8 a.m. (Image produced by the author of this thesis with *Tableau*).

5.1.1 Classification of OD-data representation techniques

As the next step, we consider different aspects of OD-data representations which are important for the analysis and classify the visualization techniques according to these aspects. The aspects which we selected for the classification are closely tied with the basic components and characteristics of OD-data discussed in Chapter 2 as well as to the analysis tasks taxonomy introduced in Chapter 4. At the same time we chose them in a way so that it is possible to make meaningful distinctions between the techniques in terms of the ways these techniques represent the components of the data.

For the classification we considered the following aspects:

– **Layout**
The way the nodes representing the locations are laid out on the screen.

– **OD**
The way the origins and the destinations are positioned and whether and how the distinction between the origins and the destinations is made.

– **Flow**
How flows between the origins and the destinations are represented.

– **Direction**
How the directions of the flows are represented.

– **Magnitude**
The way the flow magnitudes are represented.

– **Distance**
Whether and how the distances between the origins and destinations are represented.

– **OD total**
Whether and how the outgoing and incoming totals for the nodes are represented.

– **OD degree**
Whether the numbers of the outgoing and incoming flows of each node are represented.

Table 5.1 shows how the techniques listed in Section 5.1 compare in respect to the above aspects. This classification will allow us to access the suitability of the techniques for the analysis tasks. The possibilities for “Layout” and “OD” require an additional explanation, hence, they are discussed in the next subsections.

Table 5.1: OD-data visualization techniques. The table shows how the techniques compare in respect to the ways in which they represent the components and characteristics of OD-data.

Technique	Layout	OD	Flow	Direction	Magnitude	Distance	OD total	OD degree
Flow map	geo	same	explicit line	directed line	line thickness	yes[1]	optional[2]	yes
Chord diagram	circular	same	explicit line	directed line	line thickness	no	yes[3]	yes
Arc diagram	linear	same	explicit line	directed line	line thickness	no	optional[4]	yes
Sankey arcs	linear	same	explicit line	directed line	line thickness	no	yes[5]	yes
OD-matrix	linear	matrix	row→column	row→column	cell fill color	no	optional[6]	yes
OD-treemap	fit[7]	nesting	parent→child	parent→child	cell fill color	no	yes[8]	yes
Map[2]	geo	nesting	parent→child	parent→child	area fill color	yes	yes[9]	yes
OD-map	geo	nesting	parent→child	parent→child	cell fill color	yes	yes[10]	yes[11]
Hive plot	linear	separate	explicit line	O-axis→D-axis[12]	line thickness	no	optional[13]	yes
Symbol map	geo	same	on-demand[14]	O-color→D-color	on-demand[15]	yes	yes[16]	on-demand
O and D symbol maps	geo	separate	on-demand	O-map→D-map	on-demand[17]	yes[18]	yes[19]	on-demand

[1]The easiness and the accuracy of the distance estimation depends on the projection used.
[2]"Optional" means that the view has to be augmented with additional graphical objects to represent the aspect.
[3]segment size
[4]overlay with circles or bars, circle sizes or bar heights represent the totals
[5]total outgoing lines thickness of the node
[6]column/row "overall color"; rather hard to estimate
[7]The layout utilizes all available space, but can optionally produce an arrangement which is approximately geographic.
[8]Either for origins or for destinations: full size of the rectangle representing the location.
[9]Either for origins or for destinations: overall color of the cell composed of the subcells of different colors. Rather hard to estimate.
[10]Same as Map[2]
[11]But due to aggregation within the grid cells, the degrees of the original nodes are not necessarily shown.
[12]or directed line
[13]overlay with circles, circle size represents the total
[14]"On-demand" means that the user is required to interact with the view in order for the property to be represented graphically.
[15]circle size or line thickness
[16]circle size
[17]circle size
[18]Hard to estimate, because the origins and the destinations are shown in different maps.
[19]circle size

Layout

The way the origins and the destinations of OD-data are laid out in the visualization has a great influence on its readability as well as on the tasks which the visualization can support. There are the following options in our classification:

- **Geographic**
  The nodes are placed according to their locations on a geographic map, thus, supporting tasks concerning the spatial distribution of the flows.

- **Linear**
  The nodes are placed on a straight line (or several straight lines depending on the type of nodes as in case of the hive plot). A specific ordering might be applied. As a result, the number of crossings between the lines connecting the nodes can be significantly reduced in comparison to a flow map.

- **Circular**
  The nodes are arranged radially, that is, they are placed on a circle. In this case the number of line crossings between the lines connecting the nodes is not reduced, however, there is usually less clutter than in a flow map, because the nodes are not occluded by the flows and vice versa.

- **Fit**
  The node layout algorithm uses a space-filling technique attempting to fit the nodes in the available space. Treemap is a notable example of such a layout. Often, in the end the node positions do not have intrinsic meanings and do not represent the spatial configuration of the locations. However, algorithms exist which attempt to produce layouts resembling the geographic placement of the nodes and utilizing the whole available space at the same time [Mansmann et al., 2007; Wood and Dykes, 2008].

OD arrangement

This aspect concerns the way the origins and the destinations are positioned and whether and how the distinction between the origins and the destinations is made in the view. In Fig. 5.12 we show pictographs of the following four design alternatives for representing OD-data without the temporal dimension:

A. Same
No distinction is made between the origins and the destinations when they are positioned. Flows are shown either explicitly as lines connecting the nodes (as in flow maps), or "on-demand", that is, by interactively filtering the view or drawing flow lines when a specific node is selected. When used with a geographic representation, this way of arranging the origins and the destinations is the best of the four alternatives in terms of supporting tasks related to the spatial components of the data. This arrangement realistically portrays the orientation of the flows in space and their distances. However, it can result in producing a lot of visual clutter when flows are explicitly drawn as lines.

B. Separate
The origins and the destinations are placed in two separate parts of the view. These two separate parts might show the total outgoing and incoming flows for the origins and the destinations respectively. The flows between specific origins and destinations might be either represented as lines drawn across the "sub-views" or shown on-demand. Despite the separation the positioning in each of the sub-views can still be geographic. However, because of the separation, spatially relating origins to destinations, estimating distances and seeing other patterns is more difficult.

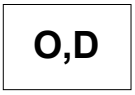
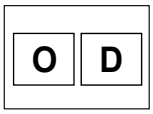
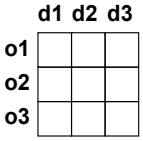
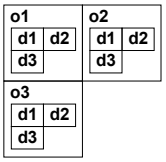
OD arrangement	Pictorgraph	Explanation	Spatiality
A. same		No distinction is made between the origins and the destinations when they are positioned.	The representation can be geographic.
B. separate		The origins and the destinations are placed in two separate parts of the view.	The placement can still be geographic. However, because of the separation, spatially relating origins to destinations is more difficult and some patterns may be harder to see.
C. matrix		The origins and the destinations are represented as rows and columns of a matrix.	The positions cannot fully represent geographic locations. However, the ordering of the rows and columns can represent a specific geographic property e.g. proximity of the locations.
D. nesting		The destinations are nested within the origins (or the contrary).	The positions of the nodes on each of the two nesting levels can be geographic. However, the level of detail is limited and spatially relating origins to destinations is difficult.

Figure 5.12: Design alternatives in terms of the OD arrangement for representing OD-data without the temporal dimension.

The clear separation between the origins and the destinations makes it possible to produce visualizations in which it is easier to compare the totals of the origins or of the destinations and also to reduce the amount of visual clutter.

C. Matrix

The origins correspond to the rows and the destinations to the columns of a matrix (the opposite is also possible). The flow magnitudes are shown by coloring the cells of the matrix. Matrix rows and columns can be freely reordered, hence, this representation can be very effective for revealing clusters of origins and destinations with similar flows, thus, supporting synoptic pattern identification tasks focusing on the flow events with the targets of their origins and destinations. The representation is not geographic, but an ordering of the rows and columns according to a geographic property of the locations can be applied. No lines have to be drawn to depict the flows, thus, the view is free from clutter caused by overlapping lines.

D. Nesting

The relations between the origins and the destinations are shown by nesting. Either the destinations are nested in the origins or the reverse. The placement of the nodes can be non-geographic (as in OD-treemap), approximately geographic or strictly geographic (as in OD-maps). Spatially relating origins to destinations is difficult because of the nesting. The level of detail is limited as the nested elements must be fit within a smaller available space. Despite the disadvantages this arrangement allows presenting a very good overview without any clutter, because, as in the matrix approach, no lines have to be drawn to depict the flows. Thus, the approach scales very well to the number of flow events. When used with a geographic layout such a representation can very effectively portray the spatial distribution of destinations for each of the origins. This means that it can provide a very good support for the spatial synoptic tasks focusing on the origins and targeting the destinations.

In contrast to the matrix arrangement with nesting the distribution of the totals for the locations (either for the origins or for the destinations) can also be shown and compared across the locations. Besides, this approach permits the use of a fully geographic placement (see Section 5.1) which is not possible with the matrix arrangement.

5.2 Visualizing temporal OD-data

In the previous section we considered OD-data without the time dimension. However, the major challenge is in bringing together the spatial and the temporal dimensions of OD-data in a way which makes it possible to explore the relationships between their temporal and non-temporal aspects. In this section we discuss visualization design alternatives which can help to address this challenge.

5.2.1 Related work

Gleicher et al. [2011] propose a taxonomy of visualization designs supporting visual comparison and suggest that any such design can be classified as one of (or a combination of) the three basic categories:

- *Juxtaposition*: The compared objects are represented separately either in space or in time (this can correspond either to small multiples or to animation which we consider later in this section). This approach requires the viewer to memorize details about separate objects in order to make comparisons.
- *Superposition*: Representation of the compared objects are shown at the same time on top of one another. This approach takes the most advantage of the human vision for making comparisons.
- *Explicit encodings*: The differences (or relationships) between the compared objects are computed and shown explicitly. Visualizations of this category rely on computation to determine the differences.

This taxonomy is very general and is not limited to the representations of changes over time. It identifies the fundamental approaches which visualizations for comparison can use, and is therefore very much related to what we discuss in the rest of the chapter. However, it primarily focuses on how visualizations can be organized to enable comparison, whereas we want to narrow the focus to the ways temporal changes can be represented.

Peuquet [1994] propose a high-level framework for representing temporal dynamics in geographic information systems enumerating the following cartographic approaches for representing time:

- Manipulating the symbology within a single visual cartographic display;
- Using a sequence of static maps that represent “snapshots” at sequential moments in time;
- Presenting a sequence of discrete displays at various speeds or dynamically modify display elements via a program or interactive control;
- Augment static maps display with supplementary graphs depicting the change in a specific variable in specific locations or over the entire region.

These approaches fully apply to visualizing temporal changes in OD-data, therefore, we will refer to them in our design space exploration.

Finally, Andrienko et al. [2011] introduce a taxonomy of generic approaches for the analysis of spatio-temporal data related to movement. The authors distinguish visualizations according to the ways the fundamental components of spatio-temporal data are represented in them and discuss the tasks which

the different types of visualizations address. While OD-data is one of the data types considered in this taxonomy, it is not its main focus. For this reason, some of the important aspects of OD-data as well as some analysis tasks and visualization approaches were not included in this general taxonomy. Being inspired by the methodology, we therefore, decided to apply it to temporal OD-data and systematize the ways in which they can be visualized.

5.2.2 Representation of time

As we just mentioned, Peuquet [1994] described four basic cartographic approaches for representing time. In addition, it is possible to use a three-dimensional representation in which the temporal changes are depicted in one of the three dimensions and the other two are used for a spatial or an abstract two-dimensional representation of the flows of specific moments in time [MacEachren, 2004; Kraak, 2003; Adali et al., 2012]. All these approaches apply to visualizing temporal changes in OD-data, so we can make the following list of the alternatives for representing time:

Small multiples

A static sequence of images each depicting a certain moment in time. The images are presented to the user at once. The approach has a limited scalability to the number of represented time periods, because the more time periods are shown the smaller each individual image has to be, and the more difficult it is to see the details. But given that the number of time periods is relatively small it can provide a very good support for synoptic spatio-temporal tasks allowing the analysis of the changes of the overall distribution of the flows over time. It is much less effective, though, for tracking the changes of the individual elements over time, that is, for temporal synoptic tasks which focus on individual flows or locations. The reason for this is that it is difficult to locate a specific element visually in each of the small multiples and track the changes of a specific element across the multiple views ignoring all the other elements. We discuss this in somewhat more detail in Chapter 7.

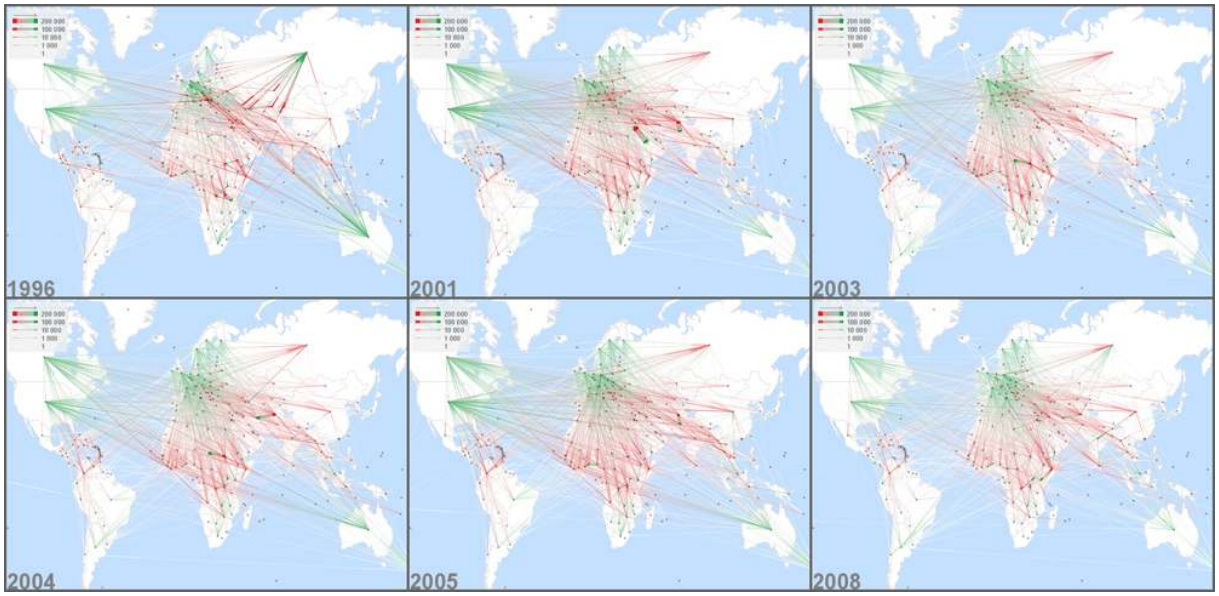


Figure 5.13: Small multiples view of the refugee dataset showing data for several years.

Animation

A sequence of images each depicting a certain moment in time, but presented one after another in an animation or with the use of an interactive control for switching between specific moments in time. Having one large view gives an important advantage for the analysis of the individual elements in a

specific moment in time. Thus, in case if a spatial layout is used, the spatial tasks, both elementary and semantic, can be very well supported by allowing the user to observe a specific moment in time. Animating over time is especially effective for detecting the appearance and disappearance of elements (again, Chapter 7 provides more insight on that), that is, for the elementary tasks in respect to the temporal component focusing on flows or locations and targeting time. More generally, animation is usually more effective for comparing changes between subsequent time periods than for seeing the overall temporal patterns, thus, temporal elementary tasks are better supported by this approach than temporal synoptic tasks.

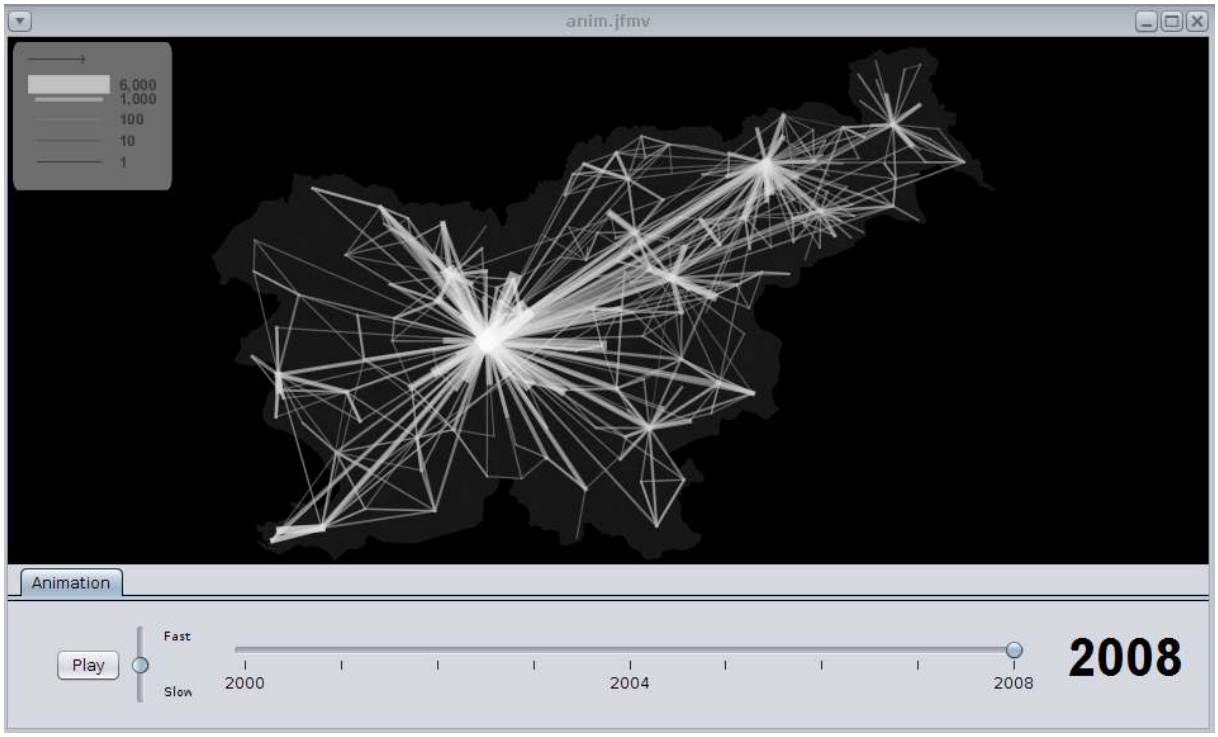


Figure 5.14: Animated view showing commuter flows in Slovenia. When the play button is pressed, the temporal changes of the flow magnitudes represented by the flow line thicknesses are animated.

Embedding

Temporal data is embedded into a non-temporal static view by introducing separate graphical elements to represent the flows in each of the moments in time in consideration (Fig. 5.15) or by providing each of the elements portraying flows or nodes with mini time series, glyphs or any other kind of graphical representation of the temporal changes of the flow magnitudes (Fig. 5.16). In other words, the changes of the data corresponding to each of the elements can be seen at once in a static view providing support for temporal synoptic tasks. In case if a geographic layout is used this means that spatio-temporal synoptic tasks can be supported as well. However, it is often quite difficult or next to impossible to visualize all the flows and the temporal changes of their magnitudes in one static view in a readable way avoiding cluttering and cognitive overload unless a small number of flows and moments in time is visualized.

3rd dimension as time

One of the dimensions of a 3D representation is used to show the temporal changes and the two others for a 2D representation of the data at each specific moment in time. When used with a geographic layout the approach has basically the same idea as space-time cube [Kraak, 2003].

Potentially, such an approach could provide support for spatio-temporal synoptic tasks in one static view. However, as with embedding it is quite difficult to create such a representation so that it is free of occlusion unless the number of the portrayed flows is very small. Fig. 5.17 shows how this approach can be used for a dataset with a very small number of flows [Proulx et al., 2010]. As with embedding, in the third dimension it is easier to depict only the node totals (as opposed to the flows) and their changes over time. Similar approaches were used in [Adali et al., 2012; Tominski et al., 2005].

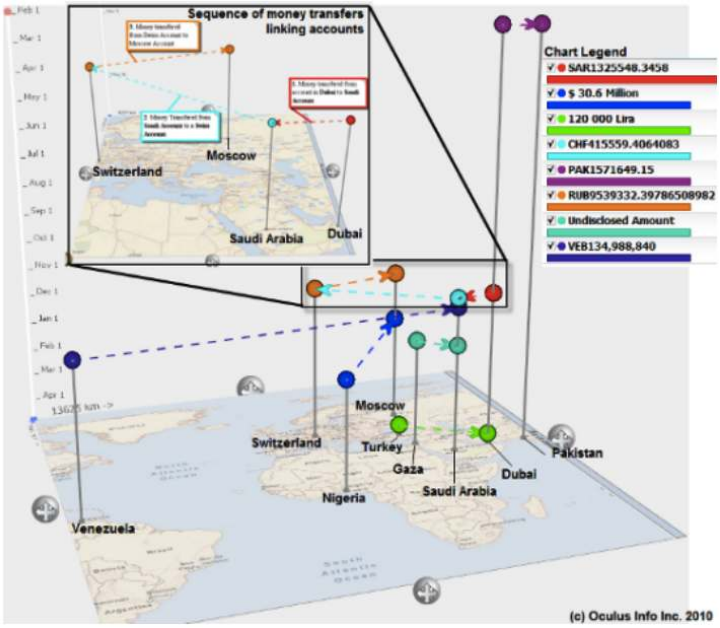


Figure 5.17: Temporal OD-data visualization of money transfers using the 3rd dimension to represent time made with the Oculus GeoTime software. ©2010 IEEE. Reprinted, with permission, from [Proulx et al., 2010].

Supplementary view

The temporal data is shown in a separate view which is connected to the representation of the origins and the destinations either by explicitly showing the relationships between elements (visual linking) or by synchronized interaction (brushing). Using this approach gives the visualization designer the freedom of choice of the temporal representation. Hence, the designer can choose a representation with the best support for temporal tasks, both elementary and synoptic. However, relating the spatial and temporal components can be more difficult with this approach requiring additional interaction or visual cues. An example of such representation is presented in Chapter 6.

In principle, each of the above approaches can be applied to any of the non-temporal OD-data visualization techniques shown in Figure 5.1. The resulting visualization will be capable of representing temporal OD-data. However, not all of the visualizations which can be created this way are equally useful for all of the tasks. In order to better understand the differences between the visualizations in Figure 5.18 we classify them according to the approaches they use for *OD arrangement* and *Representation of time*. These two aspects correspond to the three main components of temporal OD-data: origins, destinations and time. Hence, they describe the way the visualizations are fundamentally built and allow us to argue about the tasks which they can support.

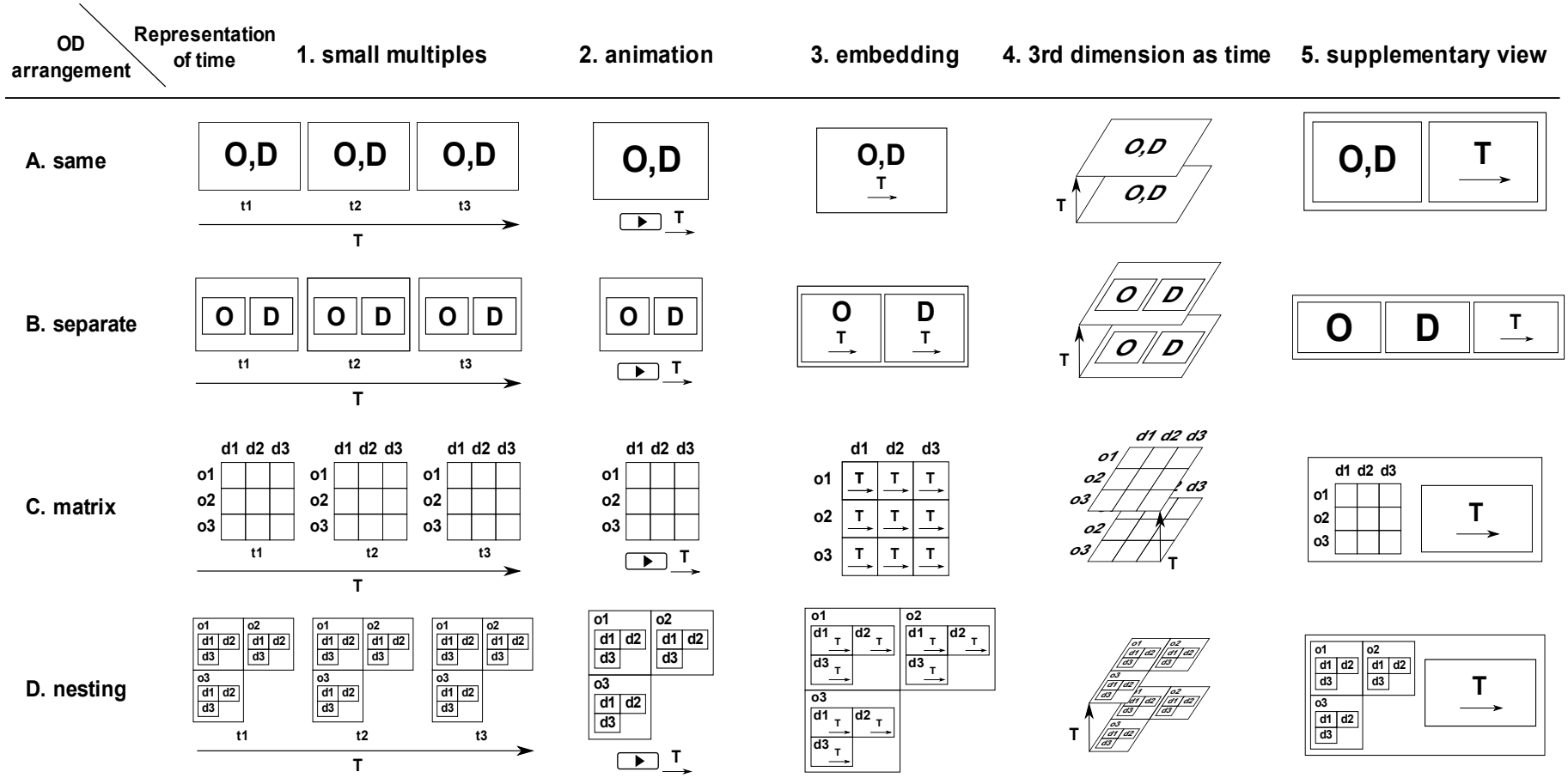


Figure 5.18: Design alternatives for representing OD-data with the temporal dimension.

5.3 Recommendations for design

Considering the design alternatives presented in Figure 5.18 and the analysis tasks discussed in Chapter 4 we can give a few general recommendations on the choice of the representation of temporal OD-data depending on the tasks which have to be addressed. These recommendations are based on the analysis of the strengths and weaknesses of the alternatives.

Addressing synoptic tasks is usually more challenging than elementary, thus, we can argue that the former should have a higher priority when choosing a visual representation. In Fig. 5.19 we schematically represent a decision tree which can help visualization designers to choose an alternative depending on the synoptic tasks support for which is most important.

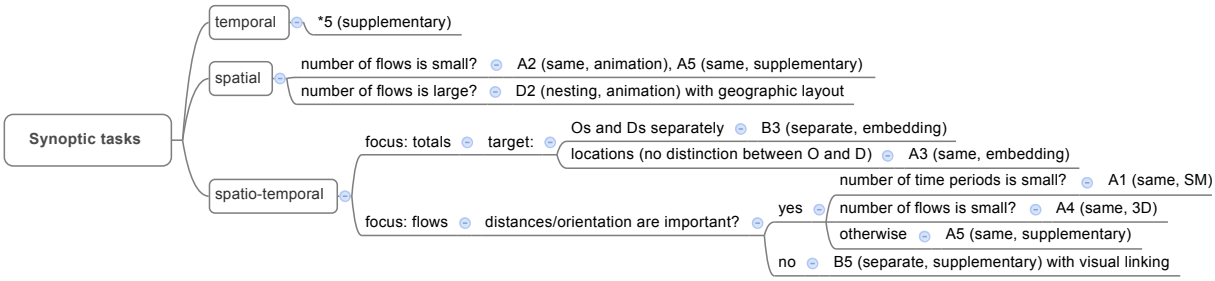


Figure 5.19: Recommendations for the choice of a temporal OD-data visualization design in the form of a decision tree based on the synoptic tasks which need to be supported in the first place.

First question to ask is whether tasks which are synoptic in respect to both spatial and temporal components at the same time have to be supported. If it is not the case, it is often better to choose a visualization which is very effective for either spatial or synoptic tasks, not both at the same time, but which avoids making a compromise between the tasks it supports.

Hence, when supporting the temporal synoptic tasks is more critical than the spatial tasks, the "supplementary view" alternative should be preferred for the representation of time (column 5 in Figure 5.18). This way the visualization designer is not constrained by the need to fit the temporal data into a geographic representation, and therefore, better support for the temporal tasks can be achieved.

In situations when spatial synoptic tasks are essential and temporal are not, obviously, geographic representations are the most useful ones. If the number of flows is relatively small and can be visualized as a flow map without too much clutter, the alternatives A2 ("same", "animation") or A5 ("same", "supplementary") should be preferred. Otherwise, using animated OD-maps can achieve very good aggregated representations showing the spatial distributions of the destinations of the flows of each of the origins. This corresponds to the alternative D2 ("nesting", "animation") with a geographic layout. The representation of time which we call "animation" implies that the user can stop the animation selecting any particular moment in time and focus on the analysis of the flows of this specific moment in time.

Compromises have to be made in cases when the most important tasks which have to be addressed are synoptic in respect to both space and time. If the focus of the tasks is on the totals for locations, not the flow magnitudes, then B3 ("separate", "embedding") or A3 ("same", "embedding") is recommended, depending on whether the totals between the origins and the destinations must be compared separately.

When the focus is on flow magnitudes, but flow distances and their spatial orientation are not important, then B5 ("separate", "supplementary") with visual linking should be used. Otherwise, if the flow distances and spatial orientation are important the choice of the design alternative is determined by the size of the dataset: A1 ("same", "small multiples") when the number of time periods is small; A4 ("same", "3D") when the number of flows is very small; and A4 ("same", "supplementary") otherwise.

5.4 Conclusion

In this chapter we discussed the design alternatives for visualizing temporal OD-data. First, we considered the alternatives for OD-data without the temporal dimension and then discussed the ways of introducing the temporal dimension and built a taxonomy of the design alternatives. Finally, we provided recommendations for making design choices depending on the tasks which need to be supported in the first place.

We cannot claim that the list of the design alternatives presented in Fig. 5.18 is complete. It is possible that some novel designs which do not fit into this taxonomy will be developed in the future. Despite that we believe that our design space exploration is useful. It presents a systematic view on the design alternatives and helps to understand the fundamental differences between them, underpinning their advantages and disadvantages. This allows us to argue about the tasks the design alternatives provide the best support for.

The design choice recommendations which we give in Section 5.3 are intended to help developers of tools for visual exploration of temporal OD-data. These recommendations have a limitation that we only give them for some of the most important types of tasks whereas in most cases multiple tasks need to be supported by visualization tools. Nevertheless, these recommendations can help visualization tool developers to better understand the trade-offs of the design choices.

In the next chapter we take advantage of this design space investigation to develop an exploration tool for temporal OD-data which supports a subset of the tasks focusing on the temporal changes of the flow magnitudes.

Chapter 6

Flowstrates

In this chapter we present Flowstrates, our technique for the visualization of temporal OD-data, which brings together a geographic and a time-oriented representation, overcomes some of the deficiencies of other approaches and provides means for identifying and analyzing spatio-temporal patterns in temporal OD-data.

The content of this chapter is based on paper [Boyandin et al., 2011].

6.1 Introduction

For any dataset of a considerable complexity and size a single static view can rarely portray the whole information contained in the data and give an overview while showing substantial details at the same time. Most temporal OD-datasets are no exception, because of their intrinsic complexity. We believe that interactive exploration can help to address this challenge. For this the users must be provided with an integrated visualization which gives a good overview of the whole dataset and supports various interactive exploration techniques which allow them to analyze the data in every detail. This is what we tried to achieve with Flowstrates, the technique which we developed for the analysis of temporal OD-data. In this chapter we present a detailed description of this technique.

We start the chapter by discussing the tasks from the taxonomy introduced in Chapter 4.3 which we chose for Flowstrates to address in the first place, then we present the technique itself and demonstrate on several usage scenarios how it supports the selected tasks.

6.2 Tasks

Flowstrates were originally developed to address the challenge of finding an effective representation for the UNHCR Refugee Dataset (see Section 2.3.1). Based on our experience with the analysis of this dataset [Boyandin et al., 2010], studying analytic reports investigating the dataset which we targeted [UNHCR, 2010], having discussions with practitioners, and reviewing the tasks which existing related techniques support [Harris, 1999], we came with the following list of tasks to address:

T1. Finding the flow events of the largest magnitude:

   A. Find when the flows of the largest magnitude took place and what are their origins and destinations (synoptic pattern search with respect to both spatial and temporal components; the focus is on “Flow event” the targets are “Magnitude”, “Time”, “Origin” and “Destination”)

   B. For a given origin and/or destination find when did the largest flows take place (synoptic pattern search with respect to the temporal component, elementary lookup with respect to the spatial components; the focus is on “Origin” and “Destination” the targets are “Magnitude” and “Time”)

   C. For a given time period find the origins and destinations of the largest flows (synoptic pattern search with respect to the spatial components, elementary lookup with respect to the temporal component; the focus is on “Time” the targets are “Magnitude”, “Origin” and “Destination”)

T2. Locating the origins and the destinations of specific flows for a specific time period“ (elementary lookup task focusing on “Flow event” with the target on “Origin” and “Destination”)

T3. Examining the flows in the neighborhood of a specific location (synoptic pattern definition in respect to the temporal component focusing on “Origin” or “Destination” and targeting on “Flow event”)

T4. Examining the “big picture”, an overview of the flow events of the whole dataset or of a region considering the flows of the region as a whole (synoptic pattern identification in respect to both the temporal and spatial components; the focus is on “Flow event” the targets can be any of the possiblities)

T5. Comparing temporal changes of the magnitudes of the flows of two locations (synoptic direct pattern comparison with respect to the temporal component; focus is on “Flow event” and the target is “Magnitude”)

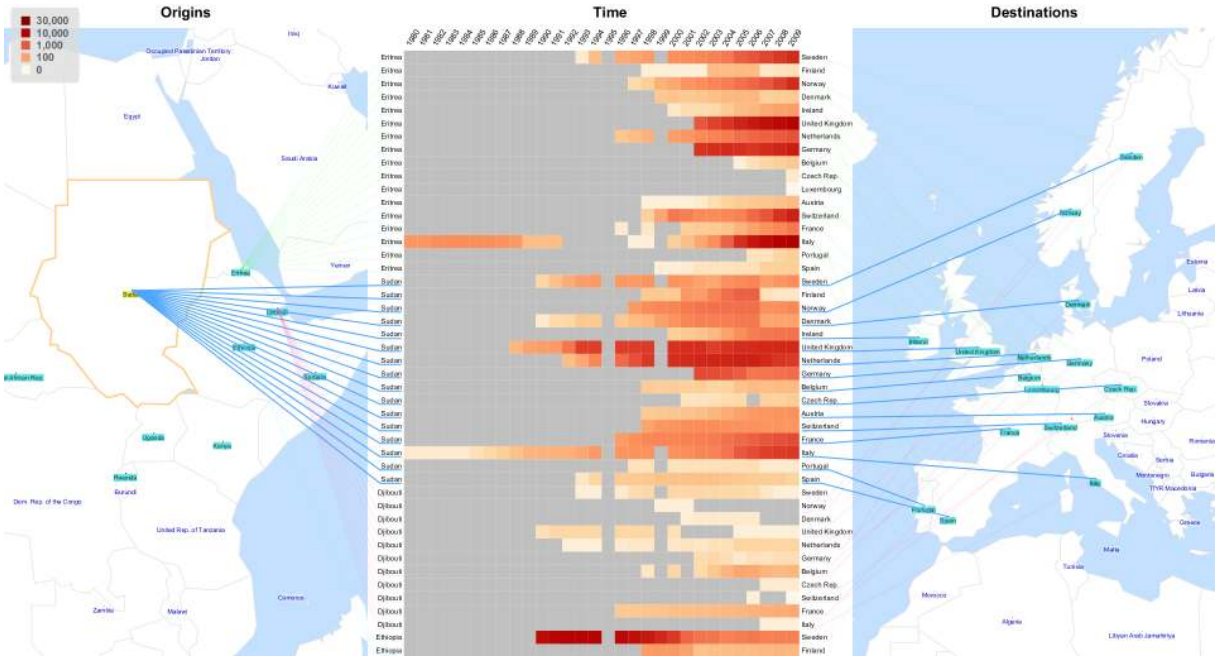


Figure 6.1: Flows of refugees are shown between East Africa and Western Europe. Flows having their origin in Sudan are highlighted. The heatmap shows the flow magnitudes by year and origin-destination. By following the lines of the heatmap it is possible to see the flows' origins, destinations and the changes of the magnitudes over time. Several temporal patterns are visually salient, such as a consistently high number of refugees from Sudan to the United Kingdom and the Netherlands, a marginal decrease to Denmark, Norway and Germany, and an increase to Ireland and Italy.

T6. Examining changes over time of the flow presence and magnitudes for specific locations (synoptic pattern identification with respect to the temporal component, elementary lookup with respect to the spatial components; the focus is on "Origin" and "Destination" the targets are "Magnitude" and "Time"; this task is similar to T1.B, but is a "pattern identification", and not "pattern search").

T7. Examining the distribution of the total incoming and outgoing flow magnitudes of locations

   A. spatial distribution for a specific time period (synoptic pattern identification with respect to the spatial component; focus is on "Origin/destination" and the target is "Total magnitude")

   B. temporal distribution for specific locations (synoptic pattern identification with respect to the temporal component; focus is on "Time" and the target is "Origins/destinations" and "Total magnitude")

Focusing on these tasks does not necessarily mean that the other tasks from the taxonomy are not supported by the technique, but that these tasks were selected to be supported in the first place. Knowing the most important tasks allowed us to make certain design decisions on which we elaborate in the following sections.

6.3 The Flowstrates

In Flowstrates the origins and the destinations of the flows are displayed in two separate maps (see Fig. 6.1). Analyzing the flow distances and orientation is not one of the tasks we want to address in the first place, besides, the exact flow routes are usually not known in origin-destination datasets. Therefore,

we have the freedom to reroute the flow lines in any way. So we represent the temporal information in an abstract view, that is, a heatmap in which the columns correspond to different time periods, and draw the flow lines so that they connect the flow origins and destinations with the corresponding rows of the heatmap, as if the flows were going through it. By an analogy with OD-matrix the way of representing flows in a matrix so that each row corresponds to a specific origin-destination pair and portrays flows of different types (or different time periods) of this OD-pair is sometimes called *dyadic OD-matrix* [Marble et al., 1997; Berry, 1966; Black, 1973].

In other words, considering the problem, the tasks we want to support (see Section 6.2), and the available design alternatives, we made the following design choices:

- Represent locations on geographical maps
- Use two separate maps for origins and destinations
- Show the temporal information in a separate abstract view
- Visually link the geographic and temporal views.

This design corresponds to the pictograph B5 in our taxonomy of the design alternatives presented in Fig. 5.18. That is, the origins and the destinations are arranged in two separate views, and a supplementary view is used for the representation of time. In the next subsections, we give a detailed discussion of the reasons for making these design choices and of their implications.

Why maps?

Maps are well familiar to everybody. They allow reasoning about the geographic patterns of the movement as no other representation by naturally providing answers to questions such as: “What is the spatial distribution of the locations?”, “How far are they from each other?”, “What are the neighbors of a location/area?”, “Which areas constitute a region?”, thus, supporting the tasks T1.A, T1.C, T2, T3, and T4.

Why two separate maps?

Displaying the flow origins and destinations in two different maps gives the following advantages:

- visually separate the origins from the destinations, thus, clearly depicting the flow directions: origin $\rightarrow$ destination (tasks T2, T1.A, T1.C)
- use any appropriate representation for the temporal data without being constrained by having to fit it into a map (tasks T1.A, T1.C, T5, T6)
- focus on different regions for the origins and destinations and perform visual queries for them in two separate maps to support tasks T3 and T4 (see 6.3.1 for details)
- augment the two maps with aggregated information for both origins and destinations at the same time, that is, showing the outgoing and incoming totals by coloring the countries (task T7.A).

However, these advantages come at a certain price. Compared to a conventional flow map, the distances between the origins and destinations, the flow routes and actual orientations in space cannot be naturally visualized in Flowstrates (see section 6.7), thus, tasks related to these properties of the flows are not supported by the technique. Despite that, the two-map solution is advantageous in situations when these properties of the flows are less important for the analysis than the temporal changes of their magnitudes. We discuss this and other limitations of the approach in Section 6.7.

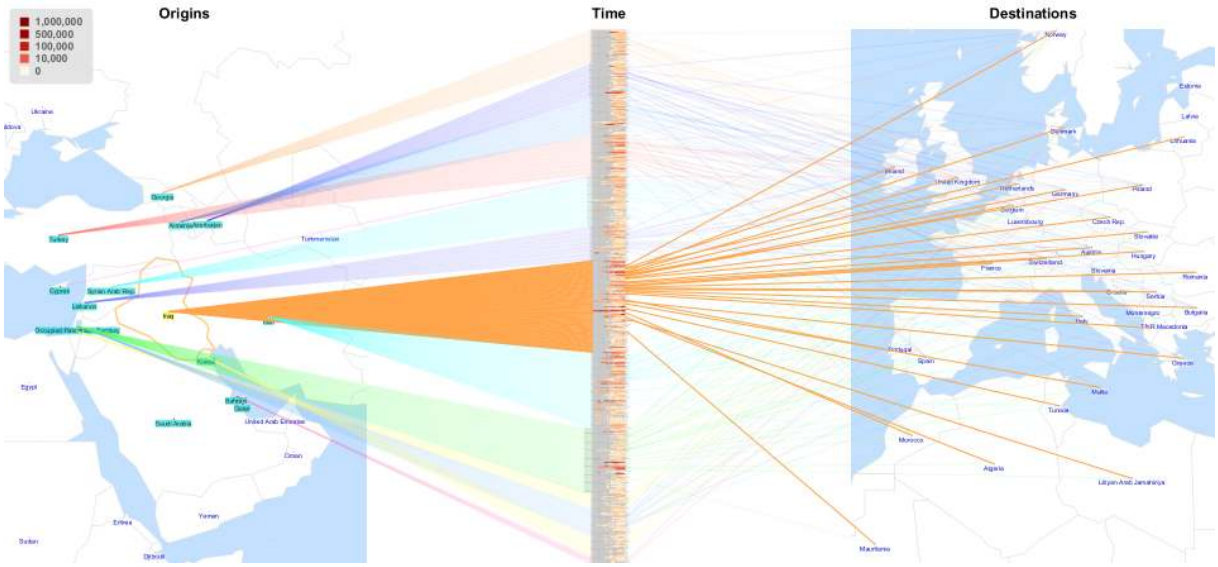


Figure 6.2: Flow line coloring: The flow lines of a selection of countries are colored by the flow origins, using a qualitative color map. The heatmap rows are sorted by the vertical positions of the origins, so that flows from the same origins are grouped together. This makes it easy to see the parts of the heatmap which represent the flows originated in the selected locations. Iran is selected in the origins map, therefore the lines from Iran are highlighted and are more opaque than the others.

Why links?

The idea to show the locations and the temporal changes of the flow magnitudes in separate views and to visually link the corresponding origins and destinations across the views was inspired by the *semantic substrates* approach for the interactive exploration of complex graphs [Shneiderman and Aris, 2006; Aris and Shneiderman, 2007].

The visual linking can be very useful in some situations. For instance, in Fig. 6.2 without the links, only with the ability to highlight a row in the heatmap or a country in the maps, it would be only possible to see flows from one origin at a time (when a country is selected in the origins map) or the origin of one flow at a time (when a row is selected in the heatmap). With the links we can clearly see what the origins of a several hundred flows in the heatmap are. Color-coding and coloring countries in the maps and highlighting the corresponding segments in the heatmap instead of drawing lines would also be possible, but then we would be limited in the ability to use coloring to show country totals in the maps (Fig. 6.4).

Why a separate temporal view?

Separating the geographic and the temporal views allows presenting the changes over time of the flow magnitudes in a way which is most suitable for the analysis of temporal patterns (this is the alternative 5, supplementary view, from our taxonomy of representations of time discussed in Section 5.2.2). The temporal view can be manipulated by the user, e.g. it can be filtered, reordered, aggregated. Still, the connections between the geographical locations and the rows of the temporal view representing flows are maintained, so that the analysts can track down the relationships between the spatial and temporal aspects of the data. In addition, this clear separation between the spatial and the temporal representations provides flexibility in terms of the initiation of the task. The analyst can begin the exploration from the temporal view and then use the spatial representation to understand where the events took place. Conversely, the user can begin from a specific region of interest and then isolate the temporal patterns

pertaining to the region of interest. Refer to section 6.4 for more details on the exploration strategies.

Why heatmap?

We chose the heatmap as the temporal data representation for two main reasons. First, it can seamlessly represent the temporal changes of the flow magnitudes at different zoom levels, thus, allowing the exploration on different analysis levels and providing support for both elementary and synoptic tasks. Second, the same color scheme as in the heatmap can be used to show the totals of the outgoing and incoming flow magnitudes in the origin and destination maps. Hence, the totals in the geographic maps can be compared to the individual values in the heatmap. The values in the heatmap represent the magnitudes of the individual flows, hence, they are usually much smaller compared to the totals in the choropleth maps. Therefore, the color scale is adjusted when a time period is selected and unselected, so that the whole available range of colors can be used in both situations.

It must be noted that the use of the heatmap requires a careful choice of the color scheme, because it has a huge impact on the interpretation of the visualization. We used a scheme based on ColorBrewer's OrRd from [Brewer and Harrower, 2009] (and RdBu for the "difference mode" shown in Fig. 6.7). Unlike in color schemes based on trilinear interpolation in the RGB color space, ColorBrewer is designed so that the intervals between the colors are perceptually uniform. Hence, there is a smaller chance that the colors are misinterpreted by the user. When a color range is chosen, the values must be mapped to the colors in the range, and there are many different approaches for doing that [MacEachren et al., 1994]. The optimal choice of the color mapping depends on the distribution of the values in the data sample. For instance, for the refugees dataset we chose a logarithmic scale, because we wanted to make the values of several different orders of magnitude well distinguishable. In addition, in Flowstrates it is possible to switch between a classified and an unclassed color scheme [Andrienko and Andrienko, 1999]. In the classified scheme the value domain is split into several intervals each represented by exactly one color, and only these several colors are actually used in the visualization. This makes it much easier for the user to tell which interval a color in the visualization belongs to. In the unclassed color scheme the colors are interpolated, hence, different values within the same interval are still represented with different colors. Switching between the two modes allows the user to decide what is more important: ability to tell easily which interval a value belongs to or fine-grained differences between individual values.

Beside the heatmap the design of Flowstrates can accommodate a number of alternative temporal views, e.g. multiple time series. Lam et al. [2007] compared the effectiveness of using multiple line graphs and heatmaps for analyzing overviews over large datasets and found that heatmap was more efficient for finding the maximum values and comparison, but less efficient for finding the graph with the maximum number of peaks. Horizon graphs [Heer et al., 2009], which are more space-efficient than time series, could be also used in place of the heatmap. It is not clear, though, how well they would support changing the zoom level. Another alternative would be to plot all the time series of the changing flow magnitudes in a single row as in TimeSearcher [Keogh et al., 2002]. The temporal view would then require much less space vertically, but it would not give a good overview and would make linking it to the geographic maps more difficult.

6.3.1 Interaction techniques

Flowstrates are meant for interactive exploration. Unlike OD-matrices which represent exactly one flow in each heatmap cell, in Flowstrates every flow takes the whole row of the heatmap. Thus, much more screen real estate is used to represent the same number of flows. Hence, for many datasets it is impossible to display all the flows simultaneously without filtering or aggregating them. If we want the analysts to still be able to explore the data in every bit of detail, then we need to provide means of interaction for controlling filtering, zooming and aggregation. Flowstrates can provide support the following techniques:

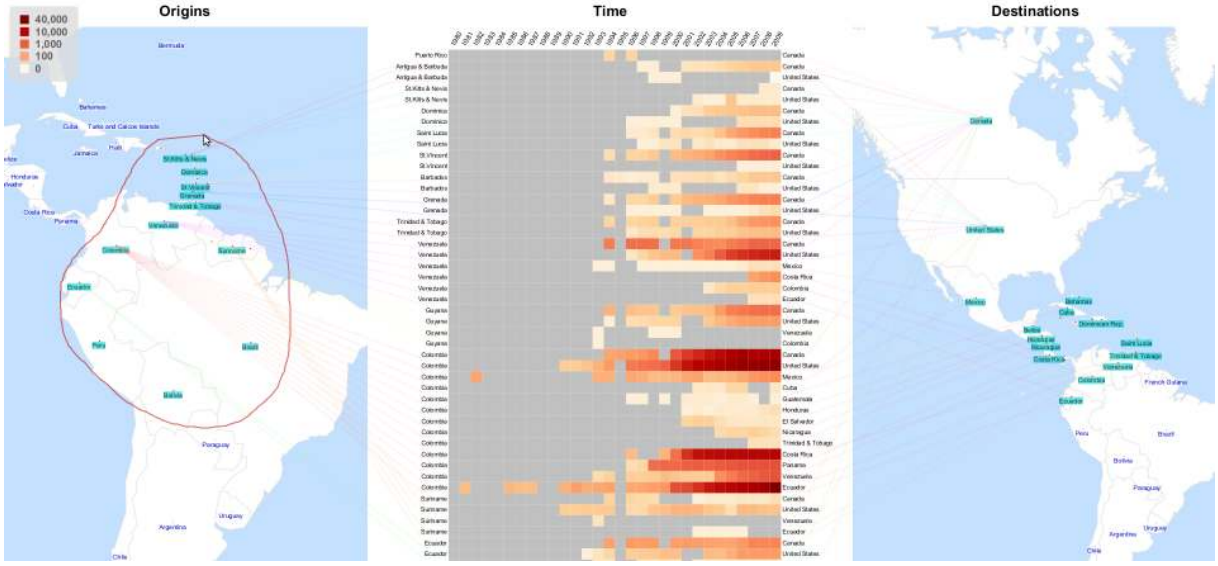


Figure 6.3: Selecting origins using lasso: When a selection is made, the heatmap is updated, so that only the flows between the selected origins and destinations are displayed.

Zooming and panning

In Flowstrates all the three views (the origin map, the heatmap and the destination map) can be zoomed and panned independently. Hence, the user can focus on different regions for the origins and the destinations and select the most relevant part of the heatmap, which is necessary for supporting the task T4. This is especially useful when considering flows between two specific regions, e.g. Asia and Europe.

Visual querying and filtering

Filtering is the most straightforward way to make a visualization more comprehensible by reducing the amount of data represented in the view (see Section 3.5.2). By allowing the user to filter the flows by their magnitude (by the maximum or average magnitude in a row) the user can reduce the number of rows in the matrix concentrating on the most important flows. Also, a subset of locations can be selected in the origin and destination maps (either filtering by name or using the lasso tool, as shown in Fig. 6.3). When a selection is made, the heatmap is updated, so that only the flows between the selected locations are displayed. Had we used only one map, making a separate selection of origins or destinations directly on the map would probably be more difficult for the user. Due to the separation of the origins and destinations in Flowstrates, we can provide support for such queries in a straightforward way. This is necessary for supporting the tasks T1.B, T3, T5 and T6.

The user can also select a time period. In this case, the outgoing and incoming totals of the regions for this time period are displayed in the maps (Fig. 6.4). To enable comparisons between the totals in the choropleth maps and the heatmap we use a common color scale for the two representations.

Finally, the user can switch to the "difference mode" in which the differences of the flow magnitudes to those of the preceding years are shown in the heatmap instead of the absolute magnitudes (see Fig. 6.7). This makes it easy to see when and how much the flow magnitudes were increasing and decreasing.

Heatmap matrix reordering

In 5.1 we discussed reorderable matrices in relation to OD-matrix representation. The heatmap in Flowstrates is also a matrix, hence, reordering can also be applied to its rows (and potentially, columns). For

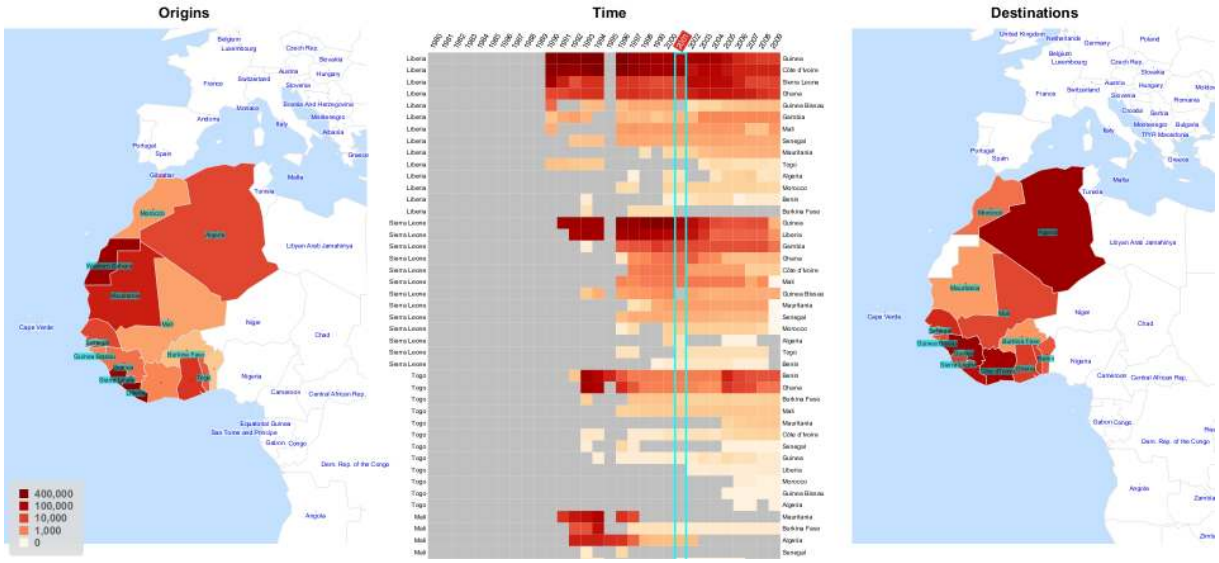


Figure 6.4: Selecting a year: Here the year 2001 is selected in the heatmap header, so the countries in the geographic maps are colored according to the total magnitudes of the outgoing and incoming flows in 2001. The heatmap rows are sorted by the maximum (over time) total magnitudes for the origin countries, and by the max magnitude in each row within the same origin country.

instance, the rows can be ordered by the maximum/average flow magnitudes, by similarity to a selected row, or they can be clustered by similarity (the latter has not been implemented in the current prototype yet). Various distance measures can be used to define the similarity between rows. When dealing with a discrete set of time periods, the simplest approach is using Euclidean distance and treating the rows as vectors in which the time periods are dimensions. There is a number of more sophisticated approaches to time series comparison [Ding et al., 2008; Morse and Patel, 2007] which can be applied as well. Alternatively, the heatmap rows can be ordered by the geographic positions of the flow origins and destinations as are the origins in Fig. 6.2 (only the y-coordinate of the projected locations is taken into account). When using the latter ordering approach, the flows sharing the same origins or destinations are grouped together in the heatmap, so the colored flow lines form “bundles” which are easy to follow.

Columns representing time periods are normally sorted in an ascending order from left to right. But the design of Flowstrates does not restrict this ordering. As the rows of the heatmap, its columns can be potentially ordered or clustered in various ways (e.g. by the similarity of the flows of different time periods). In other words, the use of the matrix representation of temporal data presents a number of possibilities for ordering and grouping the flows and the time periods. The most appropriate ordering can be chosen interactively depending on the concrete dataset and the analysis tasks.

Heatmap row aggregation

The flows represented in the heatmap can be aggregated using different grouping functions. They can be grouped, for example, by their origins so that each heatmap row represents the total magnitudes of the outgoing flows of each of the origin (see Fig. 6.5), by destinations, by the geographical regions of the origins or the destinations, or by any other flow attribute. This way we can analyze the data on different aggregation levels, or in other words, change the spatial resolution. Grouping by origins or destinations provides support for T7.B.

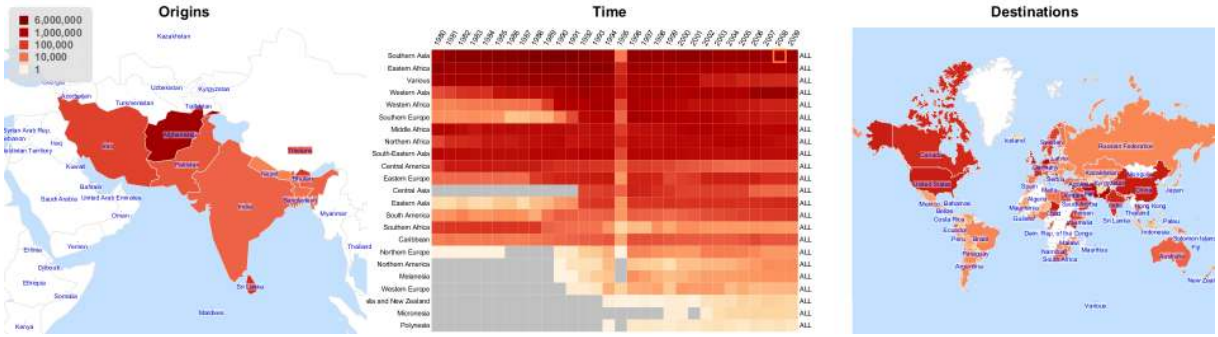


Figure 6.5: Flow aggregation: Here all the individual flows between the world's countries were aggregated by the geographic region of the origin country, so that we could see the totals of the magnitudes of the flows originated in each region. We selected Southern Asia in 2008 in the heatmap, thus the maps are colorized showing the outgoing totals for the countries of Southern Asia and the incoming totals for the countries the flows from Southern Asia went to in 2008. Here we also sorted the heatmap rows by the average magnitude in each row. The column corresponding to the year 1995 is noticeably lighter than the others. There was apparently a problem with the data acquisition, because flows for many countries are missing.

6.4 Exploration strategies

Flowstrates supports three basic exploration strategies which address the user tasks described in section 6.2. The first two strategies are both concerned with the observation of the patterns in the heatmap and differ in the initiation of the task: from location to time, or from time to location. The last one is about the comparison of either locations or time periods.

S1. **Location $\longrightarrow$ Spatial or temporal pattern.** Select a location or a region in the origins map, then find out what is going on in the heatmap or in the destinations map. This strategy relates to tasks T1.B, and T6 (described in 6.2).

S2. **Temporal pattern $\longrightarrow$ Location.** Find something interesting in the heatmap, focus on the time period when it was happening and find out in the geographic maps what the origins and the destinations were. This strategy relates to tasks T1.A and T1.C, T2, T3, and T4.

S3. **Comparison of two locations.** Select two locations or regions in the origins map and compare the temporal changes of the flows of these locations in the heatmap and their respective destinations in the destinations map. This strategy relates to task T5.

These strategies are stereotyped on purpose. In a real-life scenario analysts would generally combine them in order to solve their exploration tasks and gain new knowledge about the data. We will show how these three strategies can be applied in the usage scenarios in the next section.

6.5 Usage scenarios

In this section we illustrate how Flowstrates can be used to analyze real-world temporal origin-destination datasets applying the exploration strategies presented in section 6.4.

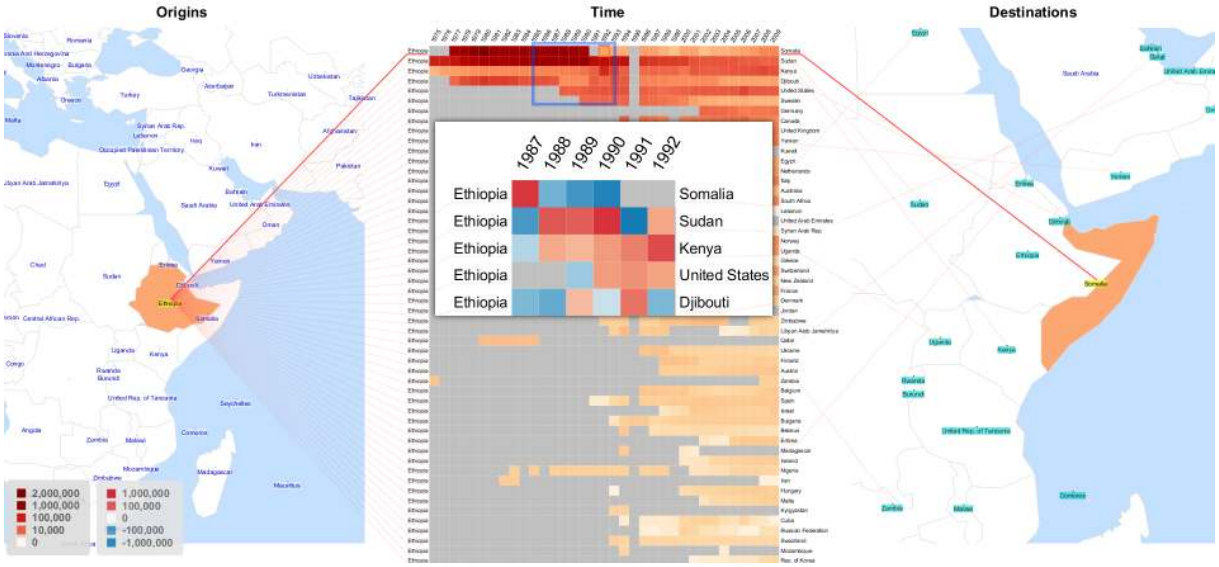


Figure 6.6: Refugees from Ethiopia: The enlarged heatmap shows the differences between the pairs of subsequent years (red shows an increase and blue a decrease in the number of refugees). The rows are sorted by the average number of refugees (over the whole range of years). Ethiopia is selected so that we only see flows from this country. Somalia is highlighted, therefore the flow Ethiopia→Somalia is more opaque than the others. The blue rectangle drawn over the heatmap highlights an interesting "staircase" pattern. Here we first used the geographical maps making visual queries to select countries. Then we used the heatmap to find the curious temporal pattern, referred back to the geographical maps to see where the other relevant locations are and found out that most of them were neighboring countries.

6.5.1 Analyzing refugee flows

With Flowstrates it is possible to see the dynamics of the changes in the refugee flows data described in Chapter 2.3.1. Suppose, we want to analyze the largest flows from Ethiopia. Let us begin the exploration with strategy S1 (in section 6.4). We zoom in to Ethiopia in the origins map, select the country (or simply use the name filter), then sort the heatmap rows by the flow magnitudes (Fig. 6.6). Now we look at the heatmap to find something curious that was happening and then where it was happening, thus switching to strategy S2. In 1987, most refugees from Ethiopia were in neighboring Somalia, but between 1988 and 1992 the number of Ethiopian refugees in Somalia drastically decreased. We can better see it in the heatmap showing the differences between the years. From 1988 to 1990 it was increasing in Sudan (second row). In 1991, it decreased in Sudan and increased in Kenya, another neighboring country. This is the reason why we can see the "staircase". Apparently, many refugees from Ethiopia were leaving Somalia from 1988 to 1990 and most of them were going to Sudan. In 1991, many were forced to flee again, this time from Sudan (probably, because of the severe drought and food shortage), and went to Kenya.

6.5.2 Commuters in Slovenia

This dataset contains the numbers of people who commute to work between the towns and villages in Slovenia. There are about 17 thousand flows for the years from 2000 to 2008, so we can explore the temporal development of these flows with Flowstrates.

In Fig. 6.7 you can see the result of the comparison of the flows to Ljubljana and Maribor, the two largest cities in Slovenia. Here, we use strategy S3 (see section 6.4). First, we select Ljubljana and

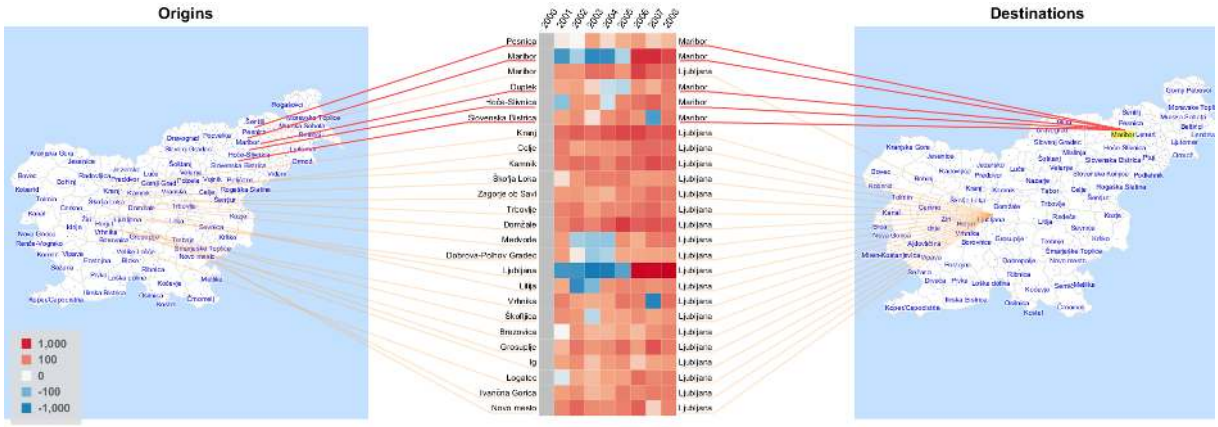


Figure 6.7: Commuters in Slovenia: to Ljubljana and Maribor, the two largest cities. Here we see the differences between the number of commuters in each pair of subsequent years (hence, there are no values for 2000). Red corresponds to an increase and blue to a decrease in the number of commuters. Only the most significant flows are shown here (filtered by the average magnitude). The numbers of non-commuters (the half-blue rows of the heatmap labeled as Maribor to Maribor, and Ljubljana to Ljubljana) were decreasing from 2001 to 2005 and increasing significantly from 2006 to 2008, whereas the number of commuters from almost all the other places were steadily increasing all the time. The flows to Ljubljana come from more distant locations than the flows to Maribor, which all come from nearby towns and villages.

Maribor in the map of destinations so that only the flows to these two cities are shown. Then we use a flow magnitude filter to show only the flows having the largest average magnitudes (over time). In the heatmap we chose to show the differences of the magnitudes between each pair of subsequent years and not the actual numbers of commuters, because we want to better see where they were increasing and decreasing.

What we can see in the heatmap in Fig. 6.7 is that for most places the numbers of people commuting from them to both Ljubljana and Maribor were steadily growing. However, the numbers of non-commuters (rows of the heatmap labeled as Maribor to Maribor, and Ljubljana to Ljubljana) were decreasing from 2001 to 2005 and increasing significantly from 2006 to 2008.

In Fig. 6.7 we also highlighted the flows to Maribor in order to see in the map of origins where the largest flows come from geographically (strategy S2). The largest flows to the capital Ljubljana come from more distant locations than the largest flows to Maribor which all come from nearby towns and villages.

6.6 Implementation

To implement Flowstrates we used the Java programming language and a number of open-source libraries. Most importantly, to build the visualizations, we used Piccolo [Bederson et al., 2004] which takes care of the rendering of the scene graph of 2D vector objects, does the input event processing for the objects in the scene, and has built-in support for zooming and panning. Besides, we used the graph data model of Prefuse [Heer et al., 2005], because it provides useful data querying capabilities treating the edges and nodes of a network as tuples in relational tables. The source code of the Flowstrates implementation is freely available as part of the JFlowMap project [Boyandin, 2010].

6.7 Limitations

One limitation of Flowstrates arises when too many lines are shown between the geographic maps and the heatmap. Too many intersecting lines become tedious to follow. This happens unless only a few locations are selected from either side of the heatmap. A partial solution is to use the heatmap row ordering by vertical positions of the origins (as in Fig. 6.2). But this only works on one side, either for the origins or for the destinations. A more sophisticated line crossings minimization algorithm could give better results, but would still require a specific ordering of the rows of the heatmap, thus limiting the possibilities for reordering. A radical solution is to only display the flow lines for a few selected or highlighted nodes.

Not being able to see the flows on *one* map as with conventional flow maps (i.e. a single map with the flow lines drawn between each origin and destination) can also be considered a limitation. With Flowstrates the orientation of the flows is not realistic and the distances between the origins and destinations cannot be estimated from the lengths of the flow lines as in a conventional flow map. To partially compensate this limitation, a conventional flow map could be displayed on demand showing the flows for a specific time period selected by the user. Alternatively, an additional column could be added to the heatmap showing the length of the flows.

6.8 Conclusion

In this chapter we presented Flowstrates, a technique for the visualization and exploration of temporal origin-destination data. The design of this technique is based on the task analysis and on a study of design alternatives. Flowstrates does not only represent spatial and temporal aspects of the data, it also highlights the relationships between these two dimensions and facilitates the interactive exploration by querying, filtering, various ordering and grouping techniques. We demonstrated how Flowstrates supports the tasks which we chose for it to address and illustrated how it can be used to analyze real-world datasets.

In Chapter 5.2.2 we discussed several alternatives for representing temporal changes in OD-data visualizations. All of them have their advantages and disadvantages, providing better support for some of the tasks than for the others. In Flowstrates we tried to combine the advantages of two of them, embedding and non-geographic view. Being able to see the flow origins and destinations on a map makes it possible to observe spatial patterns. Using an abstract temporal view allows visualizing the changes over time without having to fit the visualization in a map. Flowstrates, the solution which we presented in this chapter, takes advantage of these two alternatives bringing them together in a simple yet elegant way.

Chapter 7

User study on animation and small-multiples

Even with the development of novel and abstract visualizations flow maps will still be widely used as it is the most natural representation of OD-data. Of the multiple alternatives for representing temporal changes in flow maps small multiples and animation are the most basic ones. We analyzed the differences in the types of insights which can be gained with the use of these two representations. The chapter describes the qualitative user study we carried out to provide the basis for this analysis.

The content of this chapter is based on paper [Boyandin et al., 2012].

7.1 Introduction

As we have seen in Chapter 6 designing visualizations for the analysis of temporal changes in OD-data is a challenging task. The complex nature of the data makes it difficult to find the most suitable representation showing how the spatial relationships change over time while keeping the geographic metaphor intact. In Chapter 3 we discussed flow maps, a straightforward geographic representation of OD-data which represent flows as lines connecting pairs of locations on a map and their magnitudes by varying the thicknesses of the lines. Animation and small multiples are the two most natural of the alternatives, which we considered in 5.2.2, for adding the time component to an OD-data representation. They preserve the geographic metaphor and are, at the same time, easy to understand and interpret, compared to the other more abstract representations we considered. Flow map animation shows the changes of the flow magnitudes with interpolated transitions between the time periods. Small multiples represent discrete time periods as static images arranged next to each other in a grid format. Animation allows for a higher resolution at each time step, but, given its transitory nature, puts a high load on the user's short memory. Small multiples use space to represent time and, thus, provide only a limited resolution for each of the views.

Trying to better understand how animated and small multiple flow maps compare we studied the literature and found that:

- The comparison of animation and small multiples is a classic problem studied in conjunction with a large number of visualization techniques [Slocum et al., 2004; Fabrikant et al., 2008; Griffin et al., 2006];
- User studies carried out so far show that depending on the tasks and the exact settings of the experiment either animation or small multiples is more efficient or leads to fewer errors [Griffin et al., 2006; Robertson et al., 2008; Archambault et al., 2011; Farrugia and Quigley, 2011];
- No works comparing the use of animation and small multiples with flow maps have been published so far.

Furthermore, most studies on dynamic graph visualization consider graphs in which the positions of the nodes can change, whereas the edge weights are constant. In flow maps, on the contrary, the nodes remain stationary and only the edge weights, representing the magnitudes of the flows, change (flows may also disappear, when their magnitude becomes zero). Hence, changing flow maps represent an important special case of dynamic networks worth a separate investigation.

Our study takes a great deal of inspiration from the following research:

- The work by Griffin et al. [2006] on the perception of moving clusters, which shows that animated maps may reveal patterns that cannot be detected with static representations.
- The review of evaluation methods in visualization by Ellis and Dix [2006] in which they advocate for "explorative evaluation", an approach more focused on open-ended questions and with a higher chance of deriving new knowledge about a visualization.
- The broader line of research of insight-based evaluation [North, 2006; Saraiya et al., 2005; Yi et al., 2008] in which visualizations are evaluated in terms of the user-generated output rather than performance in the completion of benchmark tasks.

Inspired by these works, we decided to focus our research on the following question: "*Do animation and small multiples lead to the detection of different kinds of information? And if yes, how do they differ?*" To this end, we used an open-ended exploratory protocol focusing on the collection of findings, that is, any kind of information users extracted by using the tool (we prefer to use the word "finding"

instead of "insight" to allow for the inclusion of information at any level of complexity). We instructed the study participants to interact with the views and to document in the form of short sentences every piece of information they could find.

In the analysis phase we manually "coded" these findings using techniques drawn from grounded theory [Charmaz, 2006] and captured the emergent categories. The distribution of the findings across these categories forms the basis for the comparison of the two views. The methodology we chose implies a qualitative nature of the study, therefore, our analysis is not based on statistical tests. However, we provided, where appropriate, statistical numbers to document our analysis (see Section 7.4.2).

In summary, the main contribution of this work is the observation that using animation or small multiples may lead to different kinds of findings; without necessarily having one outperforming the other. Most notably, animation may promote findings of a smaller temporal and geographic scope than small multiples. Our qualitative analysis of the collected data also leads to several useful guidelines for practitioners and open questions to pursue in future research. A secondary and minor contribution is the methodology itself. The coding approach we describe in this chapter allows the comparison of visualizations in terms of the *types* of findings rather than just their *numbers*, thus, providing a deeper understanding of the information people extract from the visualizations.

7.2 Related work

Animated flow maps are discussed e.g. in [Becker et al., 1995; Thompson and Lavin, 1996], but to our knowledge no user studies have been carried out yet which analyzed the effectiveness of animation and small multiples for representing changes over time in flow maps.

MacEachren et al. [1998] discuss a user study of static and animated choropleth map representations of heart disease mortality rates. In this study a looping animation could reveal a specific pattern (a shift in location of high mortality rates) which was much more difficult to see with the use of discrete time stepping.

Slocum et al. [2004] present an evaluation comparing the use of animation and small multiples in the software package MapTime for exploring temporal changes in geographic data. The evaluation which involved interviews and discussion groups showed that animations have a more important role for examining general trends, small multiples for comparing arbitrary time periods, and change maps for explicitly depicting change.

Fabrikant et al. [2008] concluded based on the analysis of eye-movements of the participants of a controlled experiment that small multiple displays are generally not informationally equivalent to non-interactive animations and that making an animation equivalent to a small multiple display in order to achieve good experimental control for comparison may actually mean degrading its potential power.

In the user study by Robertson et al. [2008] the effectiveness of three alternative trend visualizations was evaluated by asking subjects to find answers to various analysis questions. Trend animation led to many participant errors and was the least effective form for analysis. The two static depictions of trends (small multiples and traces) were significantly faster than animation, and the small multiples display was more accurate. In this study each individual image in the small multiples view represented the trace of changes over time of one single data element, not all the elements for a particular time slice like in our case.

Griffin et al. [2006] compared the effectiveness of animated vs static small multiple maps for discovering space-time clusters and found that the subjects could more quickly and correctly identify clusters with animation than with small multiples.

In their taxonomy of techniques for visual comparison Gleicher et al. [2011] classified both animation and small multiples as juxtaposition: the former as juxtaposition in time, the latter as juxtaposition in space. According to them, both juxtapositions in space or time rely on memory for comparison, although

juxtaposition in time may be augmented by pre-attentive pattern or motion perception.

A lot of research has been carried out on the use of animation for representing changes in graphs [Purchase and Samra, 2008; Ware and Bobrow, 2004; Saffrey and Purchase, 2008; Farrugia and Quigley, 2011]. Archambault et al. [2011] evaluated the effectiveness of small multiples and animation for the comprehension of graphs changing over time. In their study small multiples gave significantly faster overall performance, whereas animation led to significantly fewer errors than small multiples for the tasks of determining nodes and edges added to the graph. Moreover, the study showed that preserving the mental map (roughly, the node positions) while showing changes both in animation and small multiples had hardly any effect on the subjects' performance.

The effectiveness of animation and small multiples is highly dependent on the information being visualized, the way it is presented to the users and the tasks they need to perform with it. It seems that very few general conclusions can be made about the usefulness of these two views for representing changes. Thus, their effectiveness has to be assessed explicitly for each particular situation.

7.3 Design of the study

The goal of our study is to find how animated (Fig. 7.1) and small multiple (Fig. 7.3) flow maps compare in terms of the types of observations people make with them. Thus, we analyzed the findings made by the study participants in these two conditions. In the rest of the chapter we will refer to the conditions as ANIM and SM respectively.

We performed an experiment with 16 subjects who were graduate and post-graduate students in computer science with no expert knowledge in migration flows or geographic visualization. The subjects were divided in two equally-sized groups. The first round of the study was designed as a between-subjects experiment. Each group was assigned one of the two views, either SM or ANIM (see Table 7.1). The participants were asked to explore the data by interacting with the view and to document their findings. The findings were collected in a database, and later, manually classified. We then performed a detailed comparison of the types of the first round findings between the views.

There was an additional second round in our experiment. The subjects were asked to continue making observations with the same dataset, but in the view which they *did not* use in the first round. Our goal was to see whether switching from one view to the other while still analyzing the same dataset would induce the subjects to make findings of different types compared to those which they made in the first view. Hence, we did not compare the findings made by the two subject groups in the second round. Instead, we compared the types of findings the subjects made in the second round with those which the same subjects made in the first round.

We preferred this approach over within-subject design because with a within-subject experiment, it would be much harder to analyze the effects of switching the view while still exploring the same dataset on the types of findings the subjects make. The main reason for designing the first round as a between-subjects experiment was also our desire to analyze the view change effects in the second round.

	Group 1	**Group 2**	**Time limit**
	Training		-
Round 1	SM	ANIM	20 min
(main experiment)	Rating findings		-
Round 2	ANIM	SM	10 min
(additional round)	Rating findings		-
	Questionnaire		-

Table 7.1: The protocol of our user study.

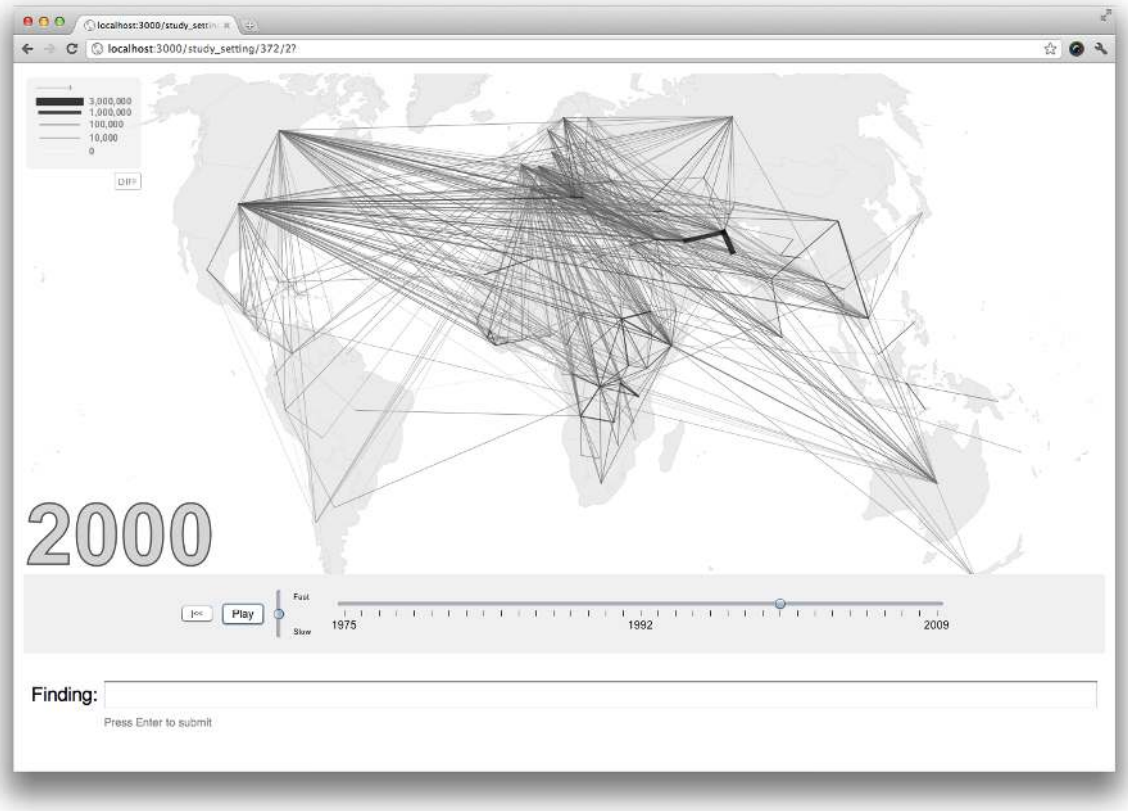


Figure 7.1: ANIM condition (the animated view) representing migration flows between the world's countries. The subjects were asked to interact with the view and make findings about the dataset entering them in the text field below the view.

At the beginning of each session the subjects were trained. Both views and their interactive facilities were presented and explained to them. The subjects could interact with the views for a while and ask questions about them. The training was performed with a different dataset from the one used during the experiment itself.

After the training phase which took between 5 and 10 minutes, the first round started. The subjects were asked to make findings in the data using the view automatically selected for them depending on the group they were in. More precisely, the subjects were given the following task:

The views you will see represent refugee migration flows. Explore these views and type down the important findings you make.

We also gave the subjects an idea of what an important finding is:

When deciding which findings are important, imagine, that you have to use them to present to somebody else what you have learned about these data.

We decided not to give examples of findings to be sure the subjects are not biased by their particular types. The goal was to see what types of findings the subjects would come up with on their own.

In the view which was shown to the subjects beside the visualization there was a text field in which they had to type short one-sentence descriptions of the findings they make. After a subject had typed a finding and pressed "Enter", the finding was stored in the database and the text field was cleared, so that a new finding could be submitted. If a subject felt that no more important findings could be made in the view, the round was finished before 20 minutes were over.

After each round, the subjects rated the findings they just made by their importance on a Likert scale with four choices between "Not important" and "Very important". They also had the possibility to mark

a finding as “Wrong” if they discovered an error, but they were not allowed to edit the findings.

In the second round the users were asked to continue the exploration of the same dataset during 10 additional minutes, but using the other visualization. In order to prevent too much fatigue, and taking into account the fact that the users would already be familiar with the data, we decided to make the second round shorter. During the analysis we did not compare the absolute numbers of findings made by the subjects, but the average percentages of the types of findings made in each of the rounds (see Section 7.4).

Finally, after completing the second round and rating the findings the subjects were asked these multiple-choice questions on the computer:

– Which of the two interfaces did you prefer overall?

– Which of the two interfaces was easier to use?

– Which of the two views allows for a higher number of discoveries?

– With which of the two views it is more likely to miss relevant information?

The possible answers were: “animation”, “small multiples”, “no difference”, “I don’t know”. The participants were also asked to rate their overall impressions of the two conditions on a Likert scale with five choices and to describe the strengths and weaknesses of the conditions. The questionnaire ended with an open question for any general comments or suggestions.

The study was performed in a closed room in front of the same computer with a 24 inch monitor. One organizer was sitting next to the subject during the whole session.

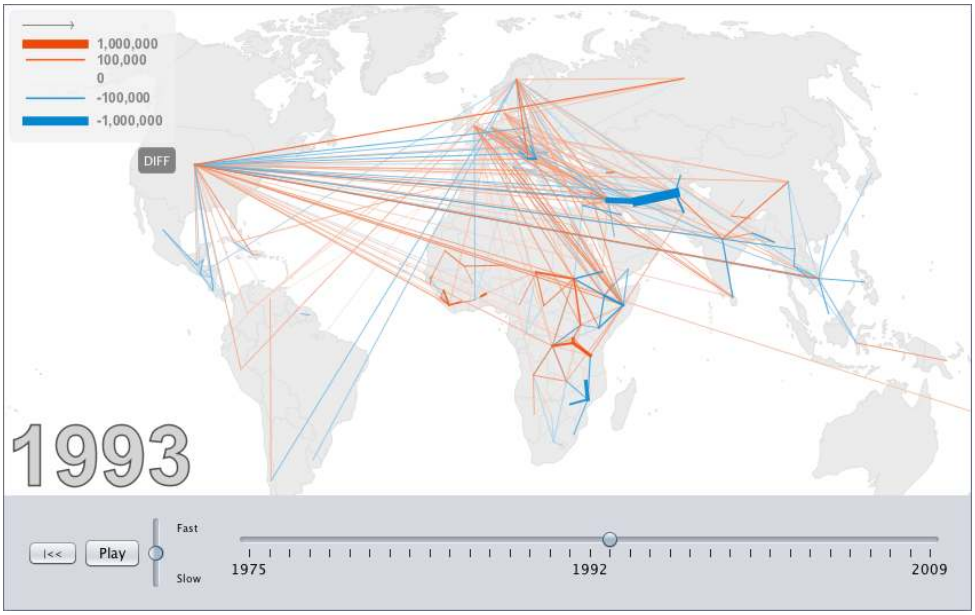


Figure 7.2: The “difference view” of the ANIM condition which shows positive and negative changes of the flow magnitudes between the currently selected and the previous years. The study participants had the possibility to switch between the original view and the difference view at any time in both ANIM and SM.

7.3.1 The conditions

The ANIM (Fig. 7.1) and SM (Fig. 7.3) conditions which we used in the experiment were based on the same flow map representation. In this representation flows of people migrating between the world’s

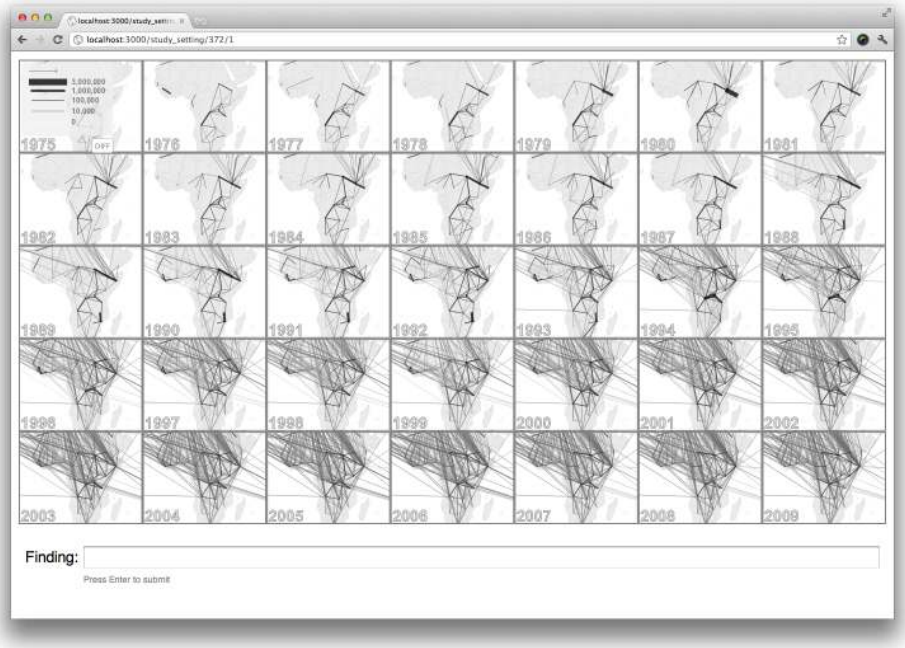


Figure 7.3: SM condition (the small multiples view). The condition supported zooming synchronously in all years' views (with the mouse wheel) and highlighting (by hovering mouse over a flow line). Here the user zoomed in to see Africa in more detail.

countries are shown with straight lines connecting the countries on a geographic map. The widths and the colors of each flow line represent the number of people migrating. We decided not to show the directions of the flows in order to simplify the views and avoid additional cluttering. For this experiment, we were more interested in analyzing findings concerning flows which changed their magnitudes over time, not in the flow directions.

The subjects had the possibility to highlight a flow by hovering over it with the mouse. When a flow was highlighted, detailed information about it was displayed, namely, the origin and the destination countries and the number of people migrating between them. In both conditions, ANIM and SM, it was possible to zoom and pan with the mouse. Zooming and panning in SM was applied simultaneously to each of the small multiples. The animated view provided the users with animation controls: a play/stop button, a small slider for changing the speed of the animation, and a large slider which allowed to select any of the 35 years presented in the dataset. The animation smoothly interpolated the data between the years.

In both views, SM and ANIM, the participants had the possibility to switch to the "difference view" (see Fig. 7.2) which showed only the differences between the selected year and the previous one. Flows which had an increased number of people moving compared to the previous year were colored red and the ones which decreased blue. The width of these flows represented the absolute values of the differences. This way participants could see what exactly changed compared to the previous year. This made it easier for them to make and document findings concerning changes between subsequent years.

The main dataset we used for the experiment represented migration flows for 35 years (available from data.un.org). Each year's data contained about 200 nodes and a few hundred flows. We had to filter the flows, showing only the few hundred largest ones, in order to guarantee that the animation runs smoothly. The dataset we used for the tutorial was different: it represented commuters in Slovenia and contained data for 9 years.

7.3.2 Data collected

During the experiment we collected the following data:

- short textual descriptions of the findings submitted by the subjects
- the importance of the findings as rated by the subjects
- screenshots of the views taken automatically when the findings were submitted
- interaction logs recorded during the sessions (all the users' actions supported by the views were logged, e.g. highlighting, zooming, starting animation)
- questionnaire submission
- videos with screen and audio recordings of the sessions.

Having such abundant data helped us during the analysis. Not only did it allow us to discern various aspects of the process of making findings, it was also useful for clarifying the meanings of those findings, which were not clearly formulated.

7.4 Analysis

The main goal of the analysis of the collected findings was to find out whether there were qualitative differences between the types of findings made in the animated view and in the small multiples. As our approach was based on grounded theory [Lazar, 2010], we did not have pre-formed hypotheses. Instead, we started from the analysis of the findings and the interaction logs developing a well-grounded theory from these data. To achieve this goal we identified coding categories, performed manual coding of the findings and carefully analyzed their distribution across the categories. In the rest of this section we discuss this process in detail.

7.4.1 Coding

As the findings concerned flows of people between geographic locations changing over time, we chose "geographic scope" and "temporal scope" as the main properties for the coding. Here are the definitions of the properties we used:

- **Temporal scope** - The time span the finding refers to.
- **Geographic scope** - The extent of the geographic entities mentioned in the finding.
- **Validity** - Whether the statement of the finding can be interpreted as a valid finding.

First, we applied a top-down approach and tried to identify the coding categories for each of the properties before the coding. Then, manual coding was performed with the predefined categories by three people (the three authors of the paper [Boyandin et al., 2012]). Each of the coders had to go through the whole list of findings and assign the property categories by choosing one of the predefined values listed for each of them. After that we calculated inter-annotator agreement rates to ensure the reliability of the coding. They were as follows:

Property	Initial agreement	Final agreement
Temporal scope	0.675	(1.0)
Geo scope	0.811	1.0
Validity	0.888	1.0

The initial agreement rates for the “temporal scope” was obviously way too low. Hence, we developed and used a simple web-based tool which helped us to find and resolve the disagreements by seeing the answers given by each of us and negotiating (see Fig. 7.4). It helped us to improve all the agreement rates, but still, the “temporal scope” agreement rate was not satisfactory.

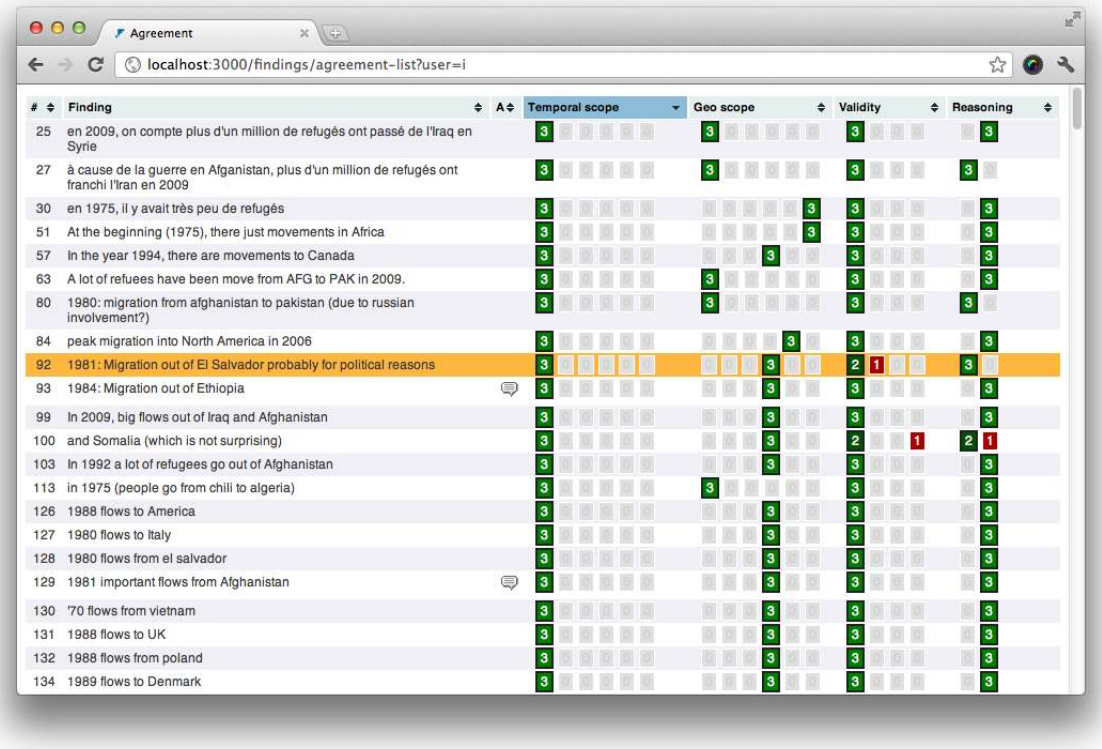


Figure 7.4: Resolving disagreements during manual coding of the findings. Each row corresponds to a finding, each column with squares to a property. The square positions represent different classes which could be chosen for each property. A green square shows an agreement indicating the number of those of us who agreed on a class, a red square is shown when there was a disagreement.

The approach of choosing predefined categories and then categorizing the findings according to them led to ambiguous coding in many cases. Hence, after several failed efforts to improve the agreement rates we decided to apply a bottom-up approach instead and manually grouped the findings without predefined categories. The approach we used for this was based on card sorting [Spencer and Warfel, 2004]. We placed all the findings written on small cards on a virtual desktop and kept arranging and grouping them on the screen until they finally formed meaningful categories (see Fig. 7.5). We did not calculate the final agreement rates for the two properties we categorized this way, because we performed this final categorization collaboratively (therefore, the final agreement for “temporal scope” is put in parenthesis in the table above).

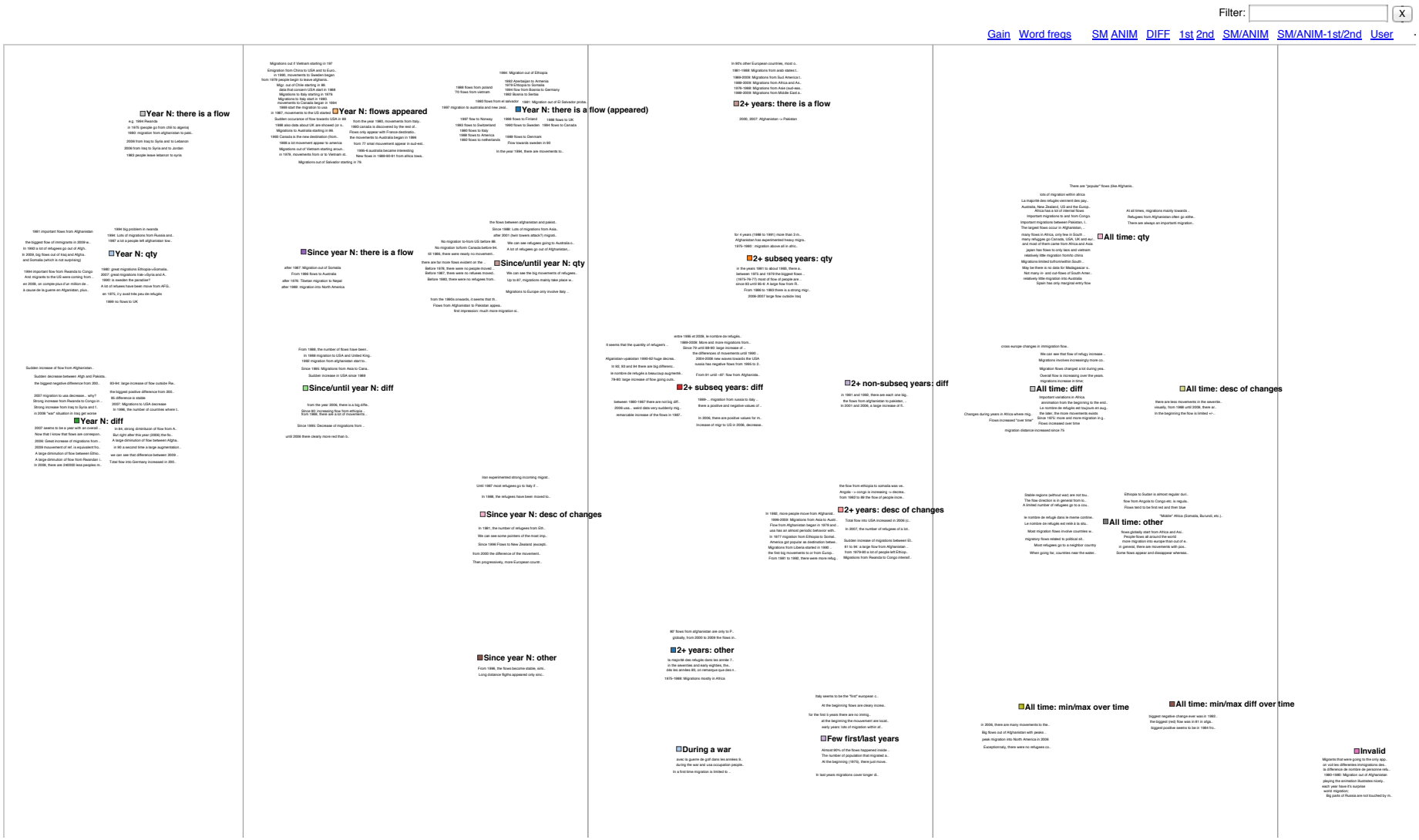


Figure 7.5: Our web-based tool which we collaboratively used for establishing the coding categories by manually grouping the findings. Each finding is a small label which can be drag-and-dropped between categories. Categories are not predefined, but can be added and removed during the process. Here the categories are arranged in columns by their temporal scope, so that "one year" findings are placed in the leftmost column and "all time" findings in the rightmost column.

The categories which we finally used for the "temporal scope" were as follows (the bottom-up approach):

- **One year** - Describes what was happening in one specific year (e.g. "1994 important flow from Rwanda to Congo").
- **Until or since** - Describes a pattern which was apparent for a time period before or after a specific year (e.g. "in 1979 movements from or to Vietnam started").
- **Interval** - Describes what was happening in a time span of several years.
- **All time** - Applies to the whole time period for which the data was available (e.g. "Migration involves increasingly more countries over time").

And for the "geographic scope" (the top-down approach):

- **Country** - Describes flows specifying only the country in which they originate or which they have as their destination, not both (e.g. "Large flows from Italy in 1992").
- **Country - Country** - Describes a flow between two specific countries (e.g. "Large flow from Russia to Italy in 1992").
- **Region** - Describes flows originating in a specific region or having the region as their destination.
- **Region - Country** - Describes flows between a country and a region.
- **Region - Region** - Describes flows between two regions.
- **Global** - Describes a global (geographically) pattern (e.g. "When going far, countries near the water are more popular destinations").

The values of the "reasoning" and "validity" properties were just "yes" or "no". In the end, the categories we came up with turned out to be useful for achieving our goal: pinpointing the differences between the types of findings made in the animated view and the small multiples.

7.4.2 Results

In all the sessions with 16 users we collected 285 findings (17.8 findings per user on average with stdev of 4.65). There were 8 findings which were not formulated clearly enough, so that it would be possible to interpret them. They were marked as "invalid" and were not considered anymore. Out of the valid findings 173 were made in the first round (ANIM: 86, SM: 87), and 104 in the second round (ANIM: 55, SM: 49).

Main experiment

In the main experiment we only compared the types of the findings made in the first round. Concerning the temporal scope (Fig. 7.6), we observed that more findings of the types "one year" and "until or since" were made in ANIM, and more findings of the type "all time" were made in SM.

Notably, 93% of "one-year" findings in SM were made in the "difference view" (see 7.3.1). It was apparently too difficult to see the differences between subsequent years in the original view in SM. In contrast to that only 44% of "one-year" findings were made in the "difference view" in ANIM.

Concerning the geographic scope (Fig. 7.7), in the first round the subjects made more local observations ("country-country") in ANIM and more global observations ("region", or "global") in SM.

Figure 7.6: Temporal scope of the findings made in the first round in each of the views.

Figure 7.7: Geographic scope of the findings made in the first round in each of the views.

Additional round

In the additional round we explored how different were the types of observations the subjects made after switching from one view to the other.

Fig. 7.8 illustrates the distribution of the findings by their geographic scope when switching between the views. When switching from ANIM to SM the proportion of "country" findings decreased while the proportion of "global" findings increased. When switching from SM to ANIM the proportion of "country" findings increased, while the proportion of "global" stayed the same.

We observed a similar effect with the temporal scope (Fig. 7.9). After switching from ANIM to SM the proportion of "one year" and "until or since" findings decreased and "interval" and "all time" increased. Switching from SM to ANIM had mostly an opposite effect: proportion of the "until or since" findings greatly increased, whereas "interval" and "all time" decreased.

Figure 7.8: Comparison of the geographic scopes of the findings made in the first and the second round depending on the subject group (i.e. the order in which the subjects were using the views).

Figure 7.9: Comparison of the temporal scopes of the findings made in the first and the second round depending on the subject group (i.e. the order in which the subjects were using the views).

7.4.3 User feedback

In addition to the findings we analyzed the questionnaire submissions made by the study participants. Asked about their overall impressions of the two conditions, the participants favored ANIM: 15 of the 16 participants rated ANIM as “good” or “very good” and 10 of them rated SM in the same way. 44% of the subjects found ANIM easier to use, 13% SM, the rest had no preference. Further, we asked the participants to provide feedback concerning the strengths and weaknesses of each view. Several of them mentioned having one large view as the main advantage of ANIM over SM. It was also much easier for them to identify appearing or disappearing flows in ANIM and compare subsequent years. SM was credited with giving an overview over the whole dataset at once, providing better support for making quick comparisons and finding differences between non-subsequent years. The main weaknesses of SM

mentioned by the study participants were that the views were too small and that there were too many of them. Thus, despite having a good overview in SM, it was difficult to focus on single elements and see how they were changing over time.

7.5 Discussion

7.5.1 On making findings

The results of the study show that alternative visualization techniques can generate or promote different types of findings and that switching from one view to the other might actually accentuate this effect. These results have a number of implications we discuss in the following.

- **Animation should be preferred for sudden change detection tasks.** Our results corroborate the outcome of the study on cluster detection by Griffin et al. [2006], which showed that certain changes might be better perceived when using animation. The higher level of small temporal scope findings and the users' feedback we received suggest that when the main task for a visualization is to be able to detect a sudden change, then animation is the preferable solution. This is also consistent with the results reported in [Ware and Bobrow, 2004] and in [Archambault et al., 2011] for the node/edge appearance tasks.

- **Using only a single technique might lead to the loss of findings.** The complementarity of the two techniques is evident from the analysis of the findings and reinforced by the feedback we received from the participants. One corollary is that if only one technique is used important findings might be lost. Animation allows for higher resolution and easier detection of sudden changes, small multiples reduces the load on short memory and allows for comparisons across many or arbitrary years. One possible solution is to provide both techniques in the same environment. Another one would be to find a way of integrating them which would allow us to overcome their limitations.

- **Switching between the views can have a beneficial effect on producing findings.** Our results suggest that switching from one view to another can lead to a boost in the number of produced findings of certain types. Thus, switching between the views can be explicitly used as a way to promote insights. Research on coordinated and multiple-view visualizations, e.g. [Keefe et al., 2009], does also show the usefulness of working with different representations. As well as the past work on investigative analysis which suggests that such an approach can help to avoid bias in judgments [Richards J. Heuer, 1999]. A systematic analysis of the effect of switching between views can be an interesting line of research to pursue in the future.

7.6 Interaction patterns and the use of animation

It appears that people use different sequences of interactions to make findings, but they often stick to their initial strategy and repeatedly use the same sequence to make new findings (see Fig. 7.10). We also noticed that the study participants tended to rate findings requiring more time and interactions as more important and those which were easier to make as less important.

In our analysis of the interaction logs we examined in particular how people use and perceive animation. First of all, we observed a disparity between what people like and what research suggests. Similarly to other studies on the use of animation, subjective user feedback is very favorable of animation; even when performance measures do not support it. Animation has definitely a special appeal on people which cannot be neglected and should be studied in more depth.

– People use the "change year" slider a lot (see Fig. 7.10), instead of playing the animation, and find it very helpful. Thus, providing animation controls without support for direct manipulation is clearly suboptimal.

– People use different strategies and might get stuck in a single one (see Fig. 7.10). Some use only the slider to control the animation manually, others only use the play/start button, some use it only at the beginning, and some combine the two. Users should be instructed on using more strategies.

– Additional research is needed to make interaction with animation smoother and more productive.

When comparing the interaction logs of ANIM and SM, we observed that people interact much more with the view in ANIM than in SM, not only because the ANIM view provides more interaction capabilities with the time slider. For instance, the highlight action was used about 12 thousand times with ANIM, and 6 thousand times for SM for about the same number of findings produced in both views in the first round (many of these highlight actions were triggered unintentionally, though, by just moving the mouse around). We made a similar observation for the panning action (542 times with ANIM, 293 times with SM). Zoom-in and zoom-out were used roughly equally numbers of times in both views. The total number of interactions is roughly two times larger with ANIM than with SM. Our interpretation of this is that ANIM favors interactions to observe local patterns and to detect sudden changes in time. On the other hand, SM favors reflection and requires less interactions to come up with findings concerning longer time periods.

Figure 7.10: Gantt chart showing a timeline of the findings (large bars) made by the 8 participants who used ANIM in the 1st round. Along with the findings we see the history of the animation-related actions (red and thin green bars) and the "change-year" slider action (thicker green bars) which the participants used to make these findings.

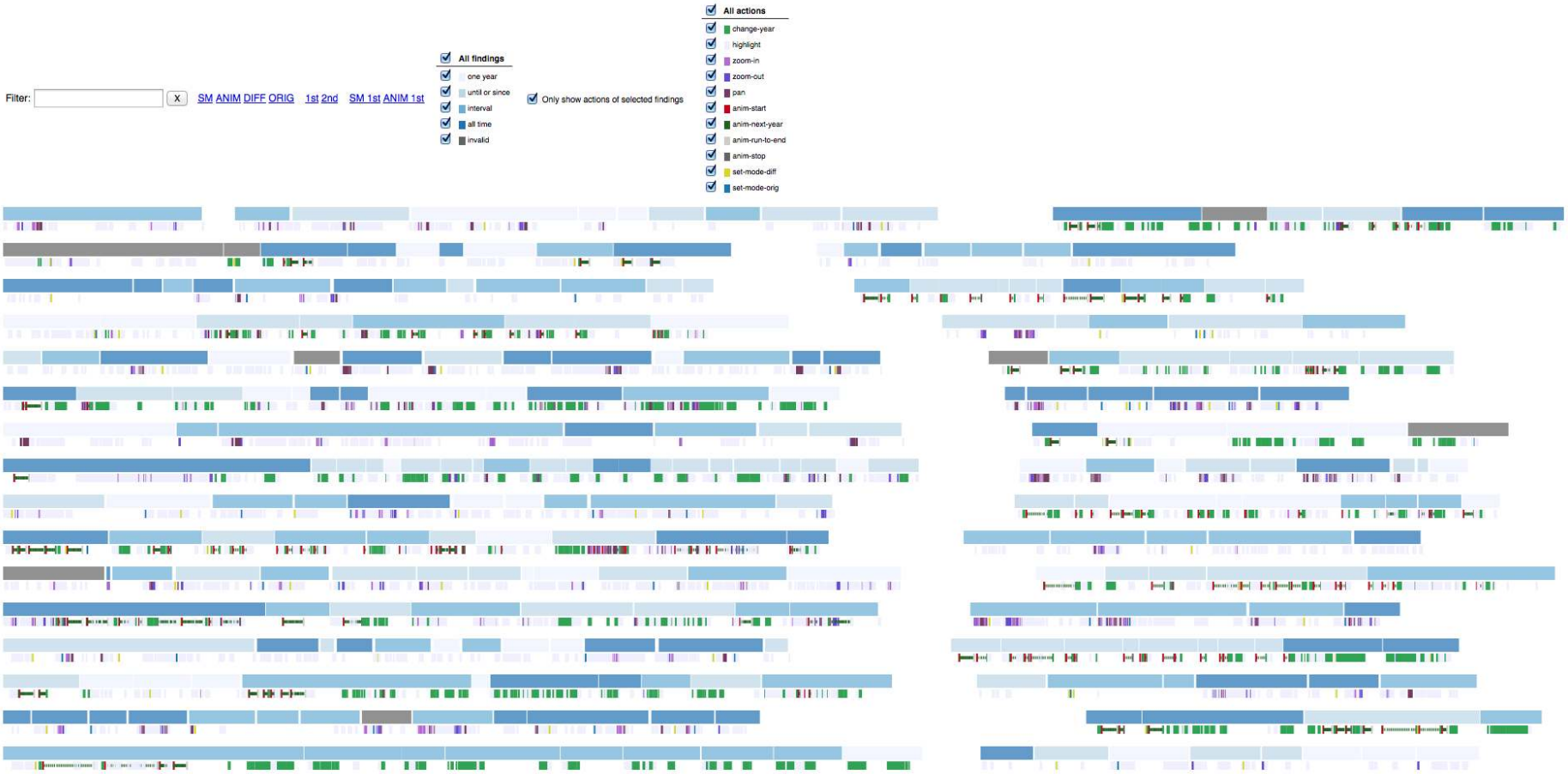


Figure 7.11: Gantt chart showing a timeline of the findings and the subjects' interactions with the views.

7.7 Conclusion

The study presented in this chapter gives partial answers to the questions we raised in the thesis introduction (see Fig. 1.1), namely: "What insights can be gained by using temporal OD-data visualizations?", "How are insights gained?", and "What interactions are used to gain insights?". The answers are partial, because we only considered two of the many possible representations: animated and small multiple flow maps. However, we find these two representations important, because, compared to the other more abstract representations, they are the most straightforward and natural for this kind of data. Besides, the choice of the representations in our user study made it possible to contribute to the long line of research comparing animation and small multiples.

More precisely, in the study we tried to find out whether the use of animated and small multiple flow maps leads to making different kinds of findings and to see how exactly the findings differ depending on the representation used. We observed that with animation the study participants made more findings concerning geographically local events or changes between subsequent years (especially, events in which flows appeared or disappeared in a specific year). With small multiples more findings concerning longer time periods were made.

Besides, our results suggest that switching from one view to another might have beneficial effects in terms of covering a larger spectrum of types of observations made. Thus, developing a smooth mechanism for integrating the two views in one exploration tool presents a great opportunity for future research (a study by Keefe et al. [2009] discusses such an approach, but applied in a very different context). It must be noted, though, that we did not make a comparison of those who switched views with those who did not. Hence, we cannot completely rule out the possibility that just the extra 10 minutes of analysis would have made the difference in the finding types. However, this is unlikely, which is corroborated by the fact that the effect mirrored depending on the order in which the views were used.

Finally, we observed that while different people used different sequences of interaction techniques to make similar findings, they often tended to stick to one strategy once they had learned how to make findings of a specific type.

One important limitation of the study is that, because of the qualitative nature of the study, the results were not statistically validated, and therefore, they cannot be easily extrapolated to a general case and have to be taken as suggestions for future research. Despite that, some of the results were confirmed by the feedback from the users or were consistent with previous research. In the future, a formal quantitative study must be performed for obtaining more generalizable results. The findings of the present study could be turned into questions which users would have to find answers to. This would make it possible to quantitatively compare different visualizations by measuring the users' performance with them.

A demonstration of some of the tools which we presented in this chapter is available online[1].

[1] http://bit.ly/flowmap-changes

Chapter 8

Visualizing AidData

This chapter presents a design study in which we try to shed light on the real challenges and on the process of temporal OD-data visualization taking a user-centered approach. For this we address a real-world problem of the analysis of financial aid allocated to countries. In the chapter we discuss the interviews we conducted with domain experts which let us characterize the problem and identify the important analysis tasks. We present visualizations we developed to address these tasks, consider the user feedback and talk about the lessons learned during this project.

In this chapter we try to learn more about the problem of temporal OD-data visualization involving real users in the process. We discuss our experience developing interactive visualizations of a dataset representing flows of financial aid between the world's countries. By taking a user-centered approach to the visualization development we tried to understand which tasks were important for the researchers working with these data and to implement the tools according to their needs and preferences. In the chapter we talk about the interviews we conducted with the researchers, the analysis tasks we derived from them, and present the prototype visualizations which we developed to address these tasks. Finally, we consider the lessons which we learned during this project and mention the challenges yet to be addressed.

8.1 Introduction

AidData.org is an initiative which supports research on aid allocation and aid effectiveness by providing scientists with the most up-to-date information about financial aid given to the world's countries by other countries and international organizations. The dataset maintained by the initiative combines information from different sources and contains more than one million flows of aid for the time span between 1949 and 2011. It includes detailed information about every individual flow and the flows are classified by their purposes.

Our collaboration with AidData.org started after one of the researchers working with AidData.org who attended a presentation about Flowstrates (which we discussed in Chapter 6) approached us and proposed to develop a specialized visualization of financial aid based on AidData, because he saw a great potential in this idea. After having remote meetings and discussions with the members of the AidData initiative the goals of the project were clarified. For AidData they were threefold:

- Providing policy-makers, development practitioners, civil society organizations, journalists, researchers, and the general public with tools that make it easy to analyse AidData and find, document and share answers to foreign aid-related questions;
- Stimulating open discussion and critical reflection on aid-related questions based on analysis of AidData, and providing an online platform for this discussion to take place;
- Raising public awareness of and interest in questions related to foreign aid, and empowering people with knowledge about the factors and processes that shape foreign aid distribution in order to improve aid transparency and accountability.

In particular the high-level questions about the aid allocation which are of interest for AidData researchers are the following:

- How is aid allocated?
- Does aid reach those who need it most?
- What factors influence the decision to provide or not provide aid?
- What is the impact of aid?

Our motivation to participate was based on the desire to apply what we had learned about the visualization of temporal OD-data to a real-world dataset working in a collaboration with domain experts. We saw this as an opportunity to put some of our ideas to test and to learn new things. Besides, the project presented a few interesting challenges:

- The dataset was by two orders of magnitude larger, than all the other OD-datasets we had worked with before;

- The questions which the researchers wanted to address were quite unobvious. It was interesting which of them we could address with a generic visualization;
- The data included additional dimensions which were important for the analysts. In particular, the flows were grouped by their *purpose* and it was important to be able to differentiate between flows of different purpose; and it was important to eventually be able to compare changes of the magnitudes of the flows of aid to other time series, e.g. specific indicators of the donor or the recipient countries;
- The target group of the visualization had to be researchers and the broad public at the same time. Hence, the visualization had to be very informative and allow researchers to find answers to complex questions, and at the same time, easy to use and interpret by the broad public.

8.2 Interviews

We conducted interviews with five political scientists working with AidData in order to understand how they analyzed these data, what were the questions they addressed and how they saw the potential role of visualization in aiding this analysis. The main findings from these interviews were the following:

- Visualization tools like what we planned to develop had not existed at that time. Researchers mostly applied statistical methods to different pieces of the data without "seeing" them as a whole. Therefore, they were very optimistic about how our tools could be useful for them.
- The potential uses of visualization of AidData which they could envision were the following:
  - for researchers
    - overview of the data, explore, drill down
    - generate new hypothesis which they could then test with statistical methods
    - quickly check that a hypothesis is feasible and worth investing more time
  - for the public: raising awareness in questions related to aid
  - for education (especially, students in the field of political science): to motivate them doing research in this field.
- Researchers themselves do statistical analysis building models and checking hypothesis. They use simple visualizations (e.g. scatter plot or a histogram) to illustrate the results. More sophisticated visualizations would probably not be of use here. Choosing a specific model and building a tool allowing the prediction of the future aid magnitudes based on various parameters could also be interesting, but would be limited to a single model.
- The most interesting questions for the researchers were concerned with showing causation between the amount of aid donated or received and some other indicators. But finding automatically which indicators correlate would not be the way it is done. Normally researchers approach this by forming a hypothesis based on some logical considerations which they then statistically test with empirical data to show causation.
- The following are examples of questions which AidData researchers address in their work:
  - Do donors manage money in ways that are responsive to past performance?
  - How does the political economy of aid (that is, influence of political interests on donors' decisions where aid goes) impacts aid effectiveness?

  - What is the impact of trade flows on aid? Will a donor give more funding to an important trade partner than to a less important one?
  - Does giving money to dictators provide economic growth?
  - How well education aid helps to reduce human capital constraints to firms?
  - Is aid used to reduce the flows of migrants coming from specific countries? Is it effective?

  However, the interviewees admitted that addressing all these detailed questions in one visualization tool would probably be too difficult and hardly even possible. They all involve different additional data to be included in the analysis which can be difficult to integrate into a single visualization. In we end we agreed that developing a visualization which would allow the users to quickly see the aid allocations related to these questions would already be quite useful.

- According to the interviewees adding the following features to the visualization would be useful:
  - Drilling down, isolating donors/recipients, filtering by purpose, filtering/querying by various indicators (from other datasets);
  - Mashing AidData with other datasets: especially, with World Development Indicators, but also with Worldwide Governance Indicators, migration data, trade data etc. This should allow comparisons of time series coming from different dataset;
  - Exporting the raw data from a visualization (once a researcher has formed a hypothesis he would want to analyze the data further with statistical tools).
- There are many issues with AidData which we and the users have to be aware of (different definitions of what aid is and what it is not, missing data, discrepancies with OECD data etc).

8.3 Requirements and tasks to support

Based on our conversations with the AidData researchers we came with the following list of requirements for the first prototypes (in which we only addressed the visualization of AidData and not the comparisons to other related datasets):

- The visualization must, first, present an overview of the total donations and receipts for all the countries and organizations;
- The country totals must be shown either for the whole time span or for a specific year, and in the latter case it must be possible for the user to select a year for which the flows are shown;
- The user must be able to select a combination of donor/recipient/purpose to see only the flows of aid for the selected combination;
- It must be possible to see how the aid amount was changing over time for a selection of donor/recipient/purpose;
- The visualization must be zero-installation and web-based to enable access by the broad public, and it must be easy to navigate and interpret.

The main tasks from our taxonomy discussed in Chapter 4 which needed to be supported were the following:

- Elementary and synoptic tasks focusing on time and targeting flow events and origin/destination ("What were the aid flows in a particular moment in time?", "Where were their origins and destinations?", "What were the changes over time of the flows of aid and their spatial configuration?")

– Elementary and synoptic tasks focusing on origin/destination and targeting the opposite locations ("Where did the flows of a specific location go and what were the patterns of their distributions over time and space?")

Adding support for comparisons to country indicators from other datasets would correspond to the following task in the taxonomy:

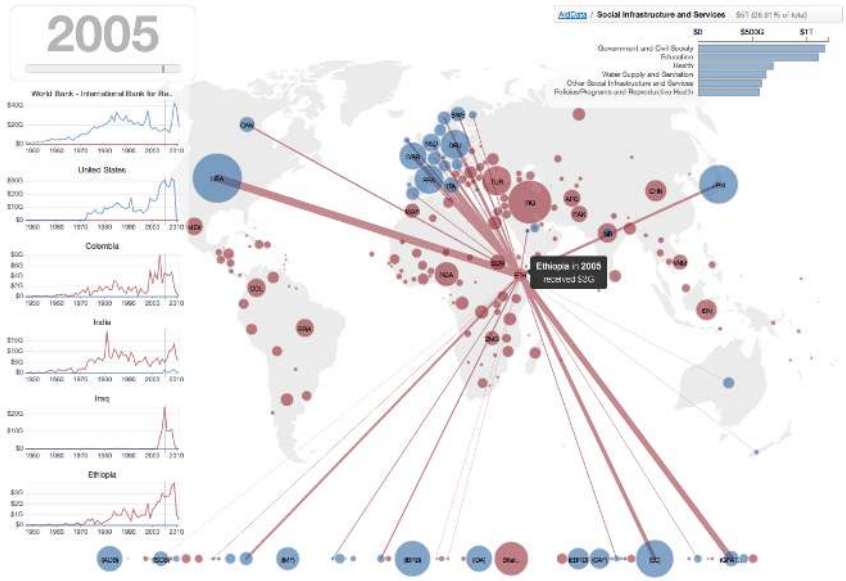
– Elementary and synoptic tasks focusing on flow events and targeting the relations of the aid flows to the context, especially, direct and inverse pattern comparison ("What was the spatio-temporal context of a specific aid flows?", 'What was the relation of the spatio-temporal of the flows to the context?")

However, we realized that adding support for these tasks and including the required additional datasets in the visualization would make it too complicated for the broad public. Therefore, we decided not address it in our first prototypes and concentrate on the visualization of AidData alone and that we would later create a more specialized tool for researchers which would address these tasks.

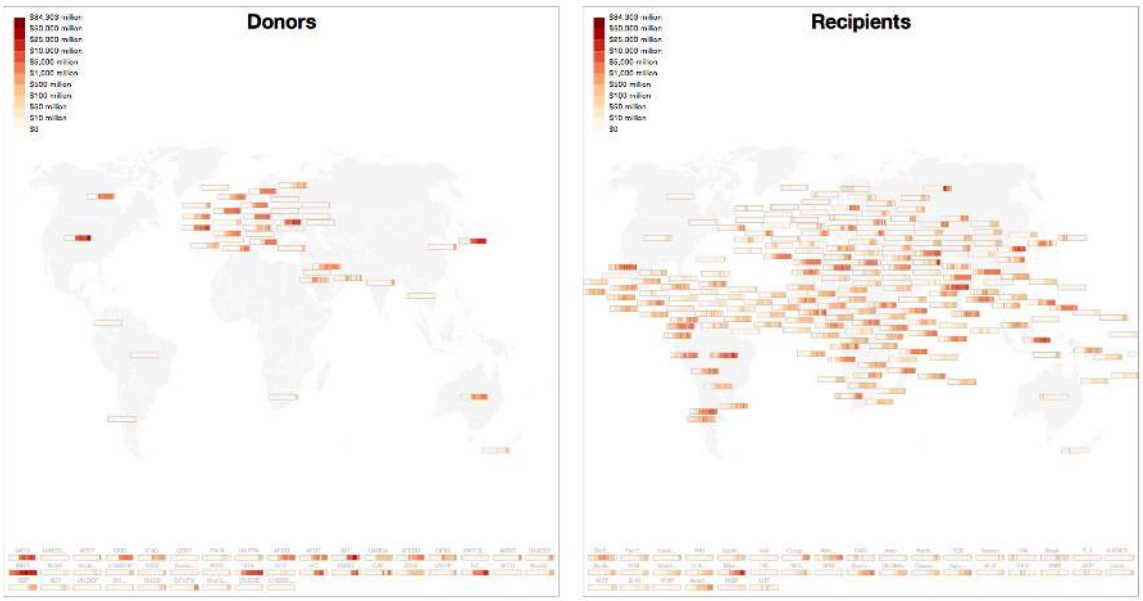
8.4 Prototypes

In order to get some initial feedback from the AidData researchers we developed three prototypes showing overviews of the data and supported basic interactions. In the prototypes we used quite different visual encodings. Below we briefly describe the three views:

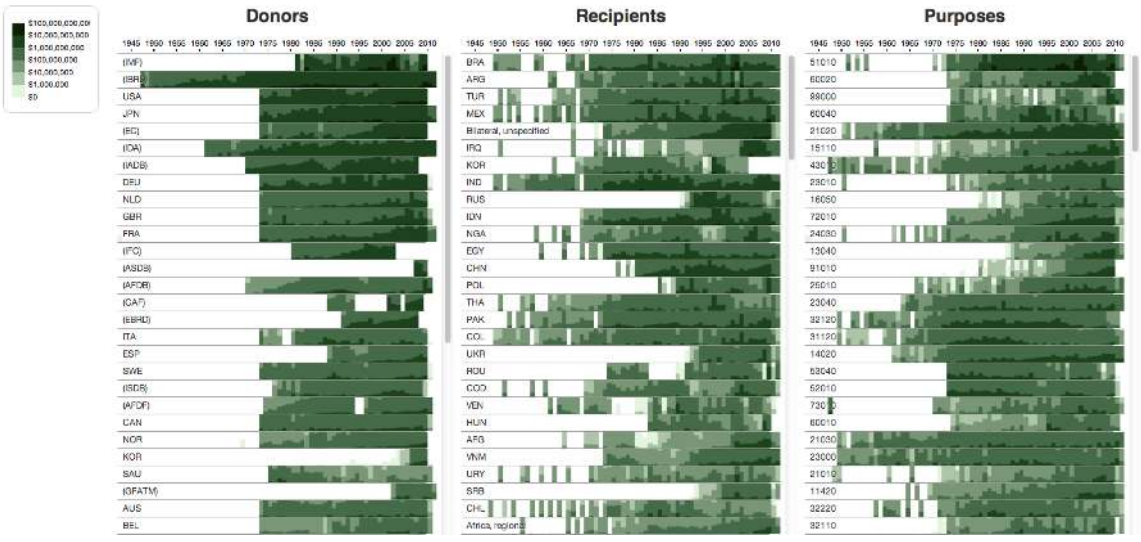
A. This view is a symbol map showing the total magnitudes of the aid flows for countries and using animation to show changes over time. It can be classified as A2 ("same", "animation") in our taxonomy of the design alternatives (see Fig. 5.1). The donors and recipients can be distinguished by color. A force-directed algorithm is used to prevent circles from overlapping, hence, producing a Dorling cartogram-like layout. To see the individual flows interaction is required: the user must hover on a location. Clicking on a location updates the view so that only the totals for flows of the selected locations are shown. It is also possible to filter the flows by selecting a specific purpose of aid. Hovering or selecting a country triggers the appearance of a small time series plot showing the temporal changes of the total amount of aid of the country enabling support for the temporal synoptic tasks.



B. In this view we decided to show the temporal changes of all the countries' total magnitudes at once without the need to use animation. To avoid confusion between the in- and outflows we used separate views for the origins and for the destinations. Every country and organization has a small bar with a color-based time series showing the temporal changes of the total outgoing or incoming (depending on the view) amount of aid of the country. A force-directed algorithm is used to prevent time series bars from overlapping. This view can be classified as B3 ("separate", "embedding") in our taxonomy. To see the actual flows instead of the totals the user has to select a country in one of the maps, then the opposite map is updated so that only the flows of the selected country are shown.



C. The locations in AidData are countries (and a few international organizations which we do not put on the map). There are not so many of them and data analysts usually know where they are located. Hence, we can argue that support for spatial tasks is less important than support for temporal synoptic tasks at least for the analysts. In the prototype C we took this argument into account and developed a visualization focusing on the temporal changes of the aid flows. It also classifies as B3 in our taxonomy, but contrary to the prototype B it does not use a geographic representation and attempts to use all the available space to better represent the temporal changes.



In this view we used horizon graphs [Heer et al., 2009] to allow more efficient comparison of large numbers of time series. Due to the large differences between the amounts of aid across countries we split the horizon into bands logarithmically. The advantages of putting small multiples in the same row is that it is easier to compare one to another because the years are aligned. Besides they can be sorted and grouped as necessary. The user can select one or multiple origins, destinations or purposes to filter the flows, then all the three columns are updated and only the flows corresponding to the selection are shown.

The feedback from the AidData researchers concerning these prototypes was unanimous. All of them preferred the prototype A as the most legible and most useful. They also saw potential in the prototype C as it let one see the whole time series for the top donors and recipients at once and compare them.

We on our side saw more potential in the prototype B, because it showed both the temporal changes and the spatial distribution at the same time. However, our partners found that the view was not very informative, but too difficult to read and to interpret. The main reason for that was probably the fact that this visualization did not utilize the space very efficiently. The bars showing the temporal changes of the aid in this view are too small despite the abundance of whitespace on the map. Using a space-filling technique which produces a layout resembling the actual spatial arrangement of the countries like spatial treemap layout [Wood and Dykes, 2008] or HistoMap [Mansmann et al., 2007] could have improved the situation with the prototype B, but at that time it was decided to finalize the prototype A and to make it available for the broad public.

8.5 The deployed solution for the broad public

The solution we finally deployed on a website for the broad public[1] is shown in Fig. 8.1. It did not differ much from the alternative A except for the additional support for interactive queries, drilling-down and filtering by different attributes and the possibility to see the original flows of the represented selection in a table view.

Besides, we tried to increase the role of interactivity in controlling the animation. Instead of the play button to animate over time we introduced a special control (in the top left corner of Fig. 8.1) which is used by just moving the mouse over the dedicated area without holding the mouse button to change the displayed year. This way it is much easier for the user to control the speed, the direction and the range of the years for which the animation is played. The mouse can also be moved over the small time series to change the current year which serves as a kind of a dynamic temporal legend advocated by Kraak et al. [1997].

8.5.1 Technical details

In this section we briefly talk about several technical details of our implementation which we find worth mentioning.

Choice of technology

One of the requirements for the visualization was that it had to work in the browser with zero installation on most computers. Hence, as the implementation technology we chose JavaScript (compiled from CoffeeScript sources) and the visualization library D^3 [Bostock et al., 2011]. This choice appeared to be the most promising in terms of the browser and device support.

The visualization required running some code on the server. For that we used Node.js as the server-side environment, and again, CoffeeScript as the implementation language. Being able to use the same

[1] The visualization can be accessed online at `http://diuf.unifr.ch/viz/aid`

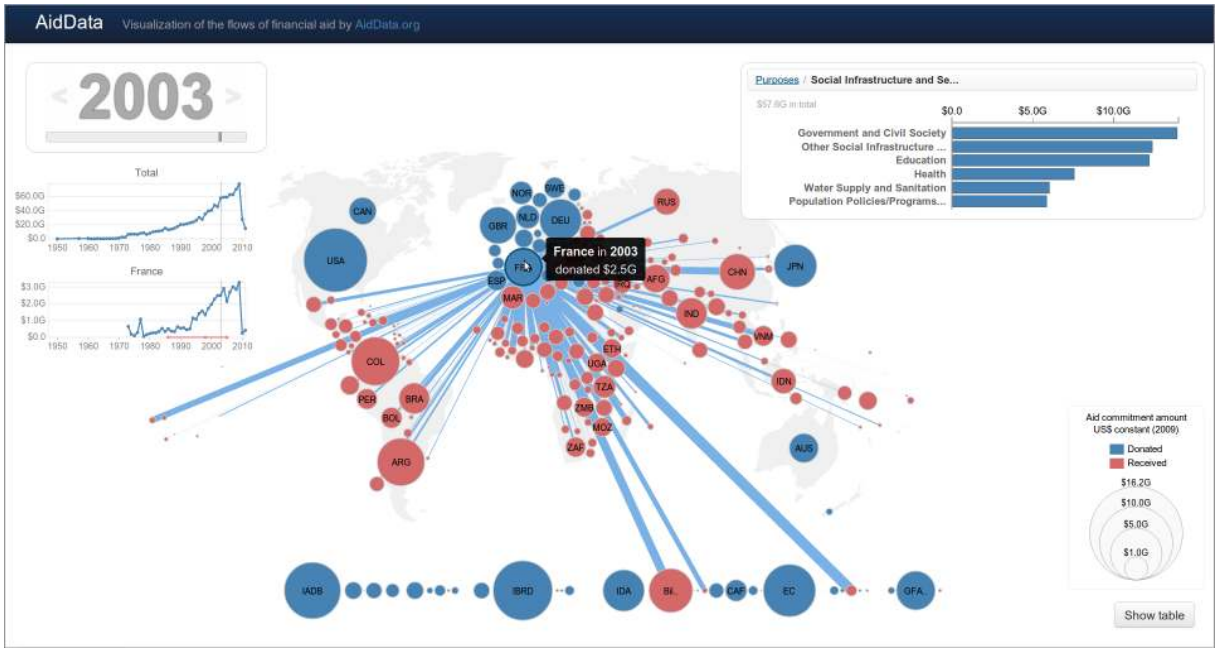


Figure 8.1: The final visualization of AidData which we developed for the broad public and deployed on the web. Here in the purpose hierarchy "Social Infrastructure and Services" is selected, hence, we only see the flows of this purpose and its sub-purposes. France is highlighted, and the flows of aid from this country are displayed as lines.

programming language on the server-side and the client-side was an advantage, as it significantly reduced the cognitive cost of switching between the two environments when programming. Besides, parts of the code could be used across the environments.

Data storage organization

The whole AidData dataset was way too large to load in the browser. For that reason we decided to request aggregated views of the data asynchronously from the visualization as they were needed. When the user interacted with the view and selected a location or a purpose a request was sent to the server to filter and aggregate the aid flows in accordance with the user selection. Then, the aggregation results were received by the application running in the browser and the visualization was updated.

This approach required a very high responsiveness of the server-side part which was responsible for data filtering and aggregation. We tried several database solutions, but they turned out to be too slow in performing the aggregation. Finally, we went with an in-memory column-based storage solution which allowed the server to swiftly perform the necessary aggregation operations and send the requested data to the client. For this we used Datavore [Heer, 2012], a small database engine written entirely in JavaScript which was in fact intended for use in the browser, but we used it successfully on the server for the in-memory storage of the basic information required to perform the necessary filter and aggregation operations on AidData.

8.6 Addressing the advanced analysis tasks

To address the advanced analysis tasks involving comparisons between the time series of aid flows and time series of country indicators from other datasets we started developing a more sophisticated tool for researchers. In the usage scenario of this tool illustrated in Fig. 8.6 we selected a subset of recipient countries and broke the time series data down by recipient to display separate small time series for each

of the recipients. Then, one of about 7000 World Bank country indicators was selected and added to the visualizaton, namely: “Poverty headcount ratio”, in the small time series it is show in red. The time series for Argentina as one of the recipients is highlighted. Flows of aid received by Argentina appear to correlate in a peculiar way with the poverty ratio in this country. This might be related to the economic crisis Argentina was suffering in 1999-2002.

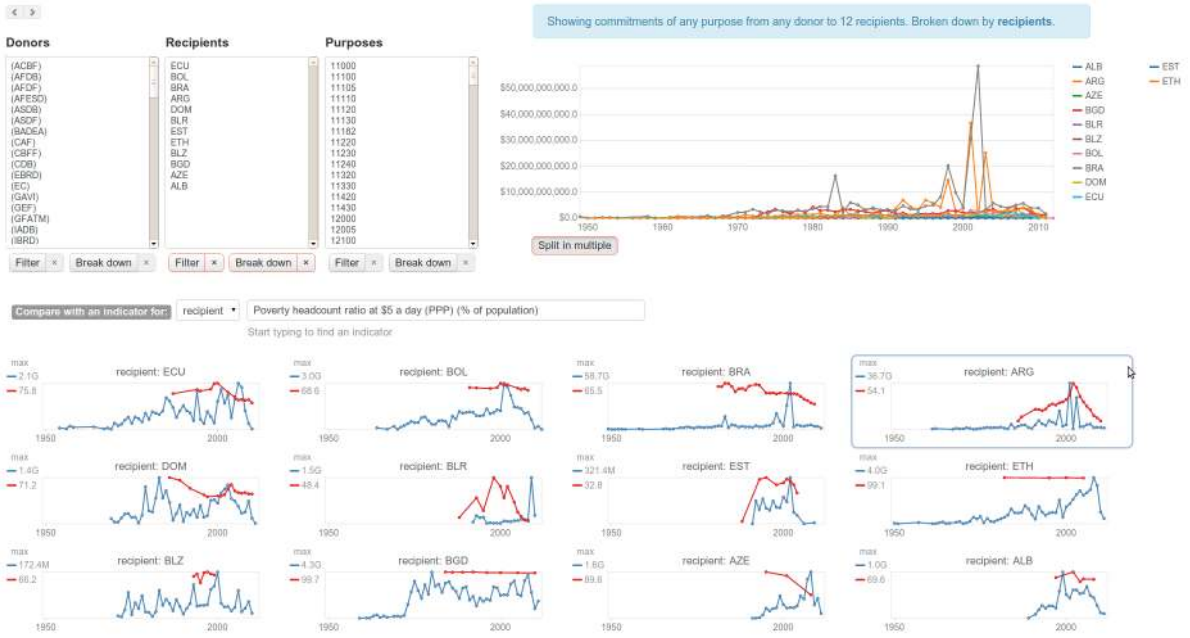


Figure 8.2: The tool we started developing for analysts which allows them to compare aid flows to time series from other datasets (in this example, “Poverty headcount ratio”).

For the small time series we used the *indexed* representation [Aigner et al., 2011a] which allows the comparison of the patterns of change in multiple time series even when they use completely different scales or when the scales differ in orders of magnitudes.

Selecting one of the small multiples allows drilling down into the selection and breaking the data down by another attribute to further explore the temporal patterns. The user has the ability to navigate back and forth in the history of the views which are displayed and the data queries which are made.

To summarize, this visual exploration tool allows representing a subset of aid flows for any donors/recipients/purposes combination and breaking the data down by one of these attributes displaying the distribution over countries and the temporal changes of the amount of aid in relation to one of the countries’ indicators. We have not yet finalized nor validated this solution, but we are optimistic about this approach as it provides support for at least some of the advanced analysis tasks which we mentioned in 8.3.

8.7 Lessons learned

In this section we summarize several important lessons we learned while working on this project. Some of them might seem obvious, but it often requires to do a field project to realize how things really work. These lessons might serve as guidelines for visualization researchers and developers.

Solving tasks vs supporting the analysts

The detailed questions which the researchers addressed in their work were quite complicated and diverse. At the beginning we were blocked by trying for too long to figure out how to address them all in a simple visualization tool. But the idea to develop a generic visualization for answering these

diverse questions turned out to be too ambitious. In fact, the AidData researchers saw the potential of these visualizations *not in solving* their tasks, but in *being supported* in solving them. As it turns out, a visualization does not always have to present a complete solution and to give full answers to complicated questions to be useful.

Two audiences

Tools for the broad public and for data analysts have very different requirements and it is difficult to develop one tool for both worlds. Most importantly, the tasks which need to be supported are not the same. We finally ended up rejecting the idea of one tool "to rule them all" and developed a rather simple solution for the broad public.

Parallel prototyping

Developing several prototypes at once might be quite time consuming. Before one specific approach is chosen, the prototypes which need to be made should be as simple as possible, ideally, just sketches on paper. Also, using tools like Tableau can be helpful for prototyping. Unlike sketches, Tableau makes it possible to see how the real data looks like.

Visual metaphor familiarity

Our guess is that one of the reasons the analysts preferred the prototype A was that the visual mappings it used were most simple and familiar to them. People might feel lost or even anxious when they see visualizations using visual metaphors they are not comfortable with. Even explaining how to read and interpret it is unlikely to change their attitude. It needs convincing arguments showing that this way of representing data is more effective to persuade people to use it. This is in line with the results of the study by Borkin et al. [2011].

8.8 Conclusion

In this chapter we discussed the design study in which we tackled the problem of the analysis of a specific dataset representing flows of financial aid between the world's countries. The goal of this work was not to develop a particular visualization technique, but to address a real-world problem working with real users, understanding their needs and preferences, proposing a solution and reflecting about the lessons learned in the process. The main contributions of this chapter are the problem characterization and the task analysis based on the interviews with the domain experts, the discussion of several design alternatives and the feedback from the experts, and the retrospective analysis which can serve as guidelines for future design studies. Most importantly, in this chapter we wanted to look at the problem of temporal OD-data visualization from the perspective of the real users. By taking a user-centered approach we tried to better understand which tasks were important for the users and which visualizations were best for the them as long as they addressed these tasks.

Although we based our task analysis on the questions which came from the experts, the deployed solution which we presented was designed for the broad public and not primarily for the experts. We also discussed the development of a specialized tool supporting the AidData researchers in finding answers to more complex analysis questions (that is, the example questions we discussed in Section 8.2). However, we have not fully solved the challenges posed by the complexity of these questions. Developing a visual exploration tool capable of addressing the diverse questions of the AidData researchers which involve analyzing relationships between data across multiple datasets remains a challenge for the future.

Chapter 9

Conclusion

A brief summary of the achievements made, concluding thoughts and suggestions for future work.

In this thesis we tackled the problem of the analysis of temporal OD-data and argued that the use of interactive exploratory visualization is a key to address the challenges created by the inherent complexity of this particular kind of data. Facilitating the development of techniques and tools enabling the visual exploration of temporal OD-data was the overall goal of this thesis. In this chapter we talk about the contributions and the impact of our work, the applications of our preliminary prototypes which show the importance of the problem and the opportunities for further development, and finally, we discuss open questions for future research.

9.1 Contributions

This work concerns different facets of temporal OD-data visualization. Therefore, in Fig. 9.1 we summarized the contributions by putting them in relation to the main entities involved in the process of interactive visualization which we discussed in Section 1.3 and to the questions we addressed in the thesis. The contributions of this thesis shown in the bottom part of Fig. 9.1 concern all of these entities in the context of visualization of temporal OD-data.

Figure 9.1: The thesis contributions put in relation to the interactive visualization process model.

The following are the individual contributions of the thesis with the references to the questions from the model of the interactive visualization process as in the figure above:

Overview of the existing techniques.
We examined various visualization techniques related to flow maps, the problems connected with their use and a number of approaches for addressing these problems. This chapter contributed to answering the questions "How can the data be visualized?" concerning non-temporal OD-data and "Which representations are effective?" (even if not very rigorously, but through a detailed consideration of the various techniques). This might be the first in-depth discussion of such a broad range of questions related to flow maps.

The taxonomy of the tasks which OD-data visualizations can support.
This taxonomy was built starting from the components of temporal OD-data and covers the questions which can be answered by analyzing such data. Our contribution was in applying the methodology proposed by Andrienko et al. [2011] to temporal OD-data. Using this methodology we defined the structure of the taxonomy encompassing the analysis questions. With this taxonomy whenever we discussed a technique we could position it according to the tasks defined in the taxonomy which the

technique targeted. The data model discussed in Chapter 2 and the task taxonomy together provide an answer to the question "What can be visualized?" by listing the components of the data which can be represented. The task taxonomy presents a detailed and systematic answer to "What questions can be answered?".

A systematic description of the design-space of temporal OD-data visualizations.

We systemized the alternative ways of visualizing temporal OD-data trying to identify the analysis tasks each of them is best suited for, and answering the question "How can the data be visualized?". This design space exploration helps to understand the differences between the design alternatives in the way the fundamental components of temporal OD-data are portrayed: namely, the arrangement of origins and destinations, and the way temporal changes are represented. Our classification of the design alternatives encompasses the existing approaches, but also describes solutions which have not yet been implemented. In addition, we give design recommendations based on this classification, answering the question "Which representations are effective?" depending on the tasks which need to be supported.

Flowstrates, a novel technique for the visual exploration of temporal OD-datasets.

Flowstrates focuses on the representation of temporal changes of flow magnitudes, whereas flow distances and their spatial orientation are not accurately represented. It is best suited for supporting synoptic tasks focusing on flow events and changes of flow magnitudes. But contrary to purely temporal representations Flowstrates also allows relating flow events to the geographic locations of their origins and destinations. One of the strengths of Flowstrates as a representation of temporal OD-data is that it is an easy to read and easy to navigate depiction of this complex data type. It allows the users to interactively explore these data and to look at them from very different perspectives and aggregation levels without the need to switch to a different representation. We believe that synoptic tasks focusing on flow events and changes of flow magnitudes, which are best supported by Flowstrates, play the central role in the analysis of many temporal OD-datasets, hence, Flowstrates can be effectively used for their exploration. This chapter also contributed to answering the question "How can the data be visualized?".

Analysis of insights gained with the use of animated and small multiple flow maps.

We analyzed findings made by the study participants while exploring temporal OD-data and identified the differences in the types of findings they made depending on the view they used. The study helped us to find an answer to the question "What insights can be gained?" depending on the representation used, albeit only for two specific representations which we chose to consider. In addition, we analyzed the ways in which various interaction techniques were used by the participants and found patterns of repetitive use of a once apprehended strategy. This contributed to answering the questions "How are insights gained?" and "What interactions are used to gain insights?". Based on these results we gave a number of recommendations for the use of animation and small multiples. We concluded that if a smooth mechanism for integrating these two views into one exploration tool and switching between them could be developed, it would be very promising in terms of the range of tasks it provides support for. The methodology of this user study combining the grounded theory approach with insight-based evaluation was novel and can by itself be considered a contribution.

Lessons learned from the AidData design study.

In this chapter we tried to learn more about the problem of temporal OD-data visualization involving real users in the process. We discussed how we approached the challenge of the visualization of a specific temporal OD-dataset identifying the tasks which were important for the researchers working with these data and how we implemented the visualization tools according to their needs and preferences. Thereby we first found out "Which questions are important?" and "What needs to be visualized?"

for the particular problem and users in the particular context. Then, we implemented several prototypes and, by analyzing the feedback from the users, tried to find out "Which visualization is best for users?". Hence, the main contributions of this chapter are the problem characterization and the task analysis based on the interviews with the researchers, the discussion of several design alternatives and the feedback from the researchers, and the retrospective analysis which can serve as guidelines for future design studies.

9.2 Limitations and future work

Fig. 9.1 and the previous section might give the wrong impression that we have found answers to all the questions related to the visualization of temporal OD-data. Indeed, we looked into a broad range of aspects, but we only answered *some* of the important questions or only answered them in part. In this section we briefly discuss the limitations of the contributions we made, mention research questions which are still open, and propose several opportunities for future research.

As we mentioned in the introduction, the inherent complexity of temporal OD-data presents a major challenge for their analysis. Strongly related to this problem is the question of scalability, that is, how well a visualization technique can accommodate to a growth of the amount of data which needs to be visualized and analyzed. There are two main groups of approaches which can help to address both the complexity and the scalability problems: the use of interactivity for data exploration and for querying the data and manually finding interesting details; and automatic approaches which can summarize large amounts of data or extract interesting patterns from them. In the thesis we concentrated mostly on the use of interaction and only mentioned a few automatic approaches: regionalization, segmenting flows into series of flows between adjoining regions and merging them, summarizing flows in OD-maps, flow aggregation and grouping by similarity in Flowstrates. We have not, however, proposed novel solutions in this respect. The scalability of Flowstrates could be potentially improved with the help of automatic approaches: for instance, by grouping similar flows together and showing only one representative for each group in the heatmap; or by spatially clustering nearby locations and showing only aggregated flows.

The classification of the design alternatives which we made describes the existing methods of arranging origins and destinations and of introducing the temporal dimension, but it does not describe all the theoretically possible ways to visualize temporal OD-data. Hence, it is possible that in the future other solutions appear which will not fit into our classification. Besides, our classification offers only one view of the space of design alternatives. We made a choice of properties which we found sensible to classify the alternatives by, but other classifications would be possible. For instance, we did not include the ways the flow magnitudes are represented in the taxonomy, or the exact way of positioning the locations. Such classifications are always generalizations presenting an overall view of the alternatives and every entry in the taxonomy can be implemented in different ways.

It is worth noting that several of the design alternatives we described have not yet been implemented. Some of them may not make much sense, as, for instance, D4 ("nesting" and "3rd dimension as time") in Fig. 5.18. But anyway, it could be an interesting opportunity to implement them or even to develop a flexible visualization system which would allow the user to easily change the representation within our taxonomy while maintaining a traceable logical connection between the same objects in different representations. As we learned from the user study, different representations provide better support for different tasks and switching between them could help users to look at the same data from different perspectives and gain additional insight. This also relates to one of the conclusions of our user study in which we suggested to integrate animation and small multiples into one representation with the possibility to easily switch between the two allowing the users to take advantage of the complementary task support of the views.

Flowstrates, the technique we proposed, does not address all of the tasks in the taxonomy of temporal

OD-data analysis tasks, which we introduced. Only a subset of the tasks which we found to be important for specific datasets are supported. The same concerns the AidData visualizations, as what we learned from the users about the tasks which are important is only fully valid for the particular use cases and cannot be automatically extrapolated to other datasets. Of course, it would be hardly possible to produce a single visualization which supports all of the tasks in the taxonomy. But the choice of tasks to support in visualizations when there is not just one single "most important question" to answer often implies self-willed decisions. It would be interesting to look at this from the perspective of visualization development as a social process and study how these decisions are made and what their implications are.

One question which we only partly addressed in the thesis is the use of interactions. In the introduction we argued that the visual analysis of temporal OD-data can be facilitated by the use of interactivity because the complexity of the data makes it extremely difficult to fully represent them in a comprehensible way in a static visualization. We could find some support for that analyzing the results of our study and the extensive use of the interactions by the participants. However, we did not systematize the interactions for temporal OD-data visualization in general and did not consider them in our design space exploration. This could be an opportunity for future research.

In the user study we tried to answer the question "What kinds of insights can be gained depending on the representation used?". However, we only compared two particular representations of temporal OD-data: animated and small multiple flow maps. We had good reasons to choose these two representations, but it would be interesting to carry out a more thorough study comparing the types of findings which can be made with other visualizations as well. Besides, our study was conducted in a lab setting and the participants were not domain experts carrying out data analysis as part of their job. A long-term study with real data analysts could help to better answer the question about the real outcomes of visualization and understand how they depend on the choice of a particular data representation.

9.3 Impact

In the course of this thesis we experimented with various temporal OD-data visualizations and developed an open-source tool JFlowMap (Fig. 9.2) which has been publicly available online[1]. The tool includes a flow map representation which visualizes changes over time with the use of animation and small multiples. It also provides support for edge bundling in flow maps and includes a change map view and our implementation of Flowstrates. The users can load their own datasets along with maps to visualize and interactively explore them in one of these views.

Since the publication of the tool we have received dozens of emails from actual and prospective users. We learned from their feedback that JFlowMap has been utilized to visualize many different datasets, including the following:

- Scientific co-publications (flow map, Flowstrates) [Degelsegger and Gruber, 2012]
- NYC taxi flows (flow map, Flowstrates) [Zhang, 2011]
- Global resource flows (flow map, Flowstrates) [Boon et al., 2012]
- Supply chain distribution of a logistics company (flow map)
- Domestic animal trade (flow map, Flowstrates)
- Working labor migration in Chile (Flowstrates) [Rowe and Bell, 2012]
- Working migrants in Slovenia (flow map, Flowstrates) [Konjar et al., 2010]

[1] http://code.google.com/p/jflowmap/

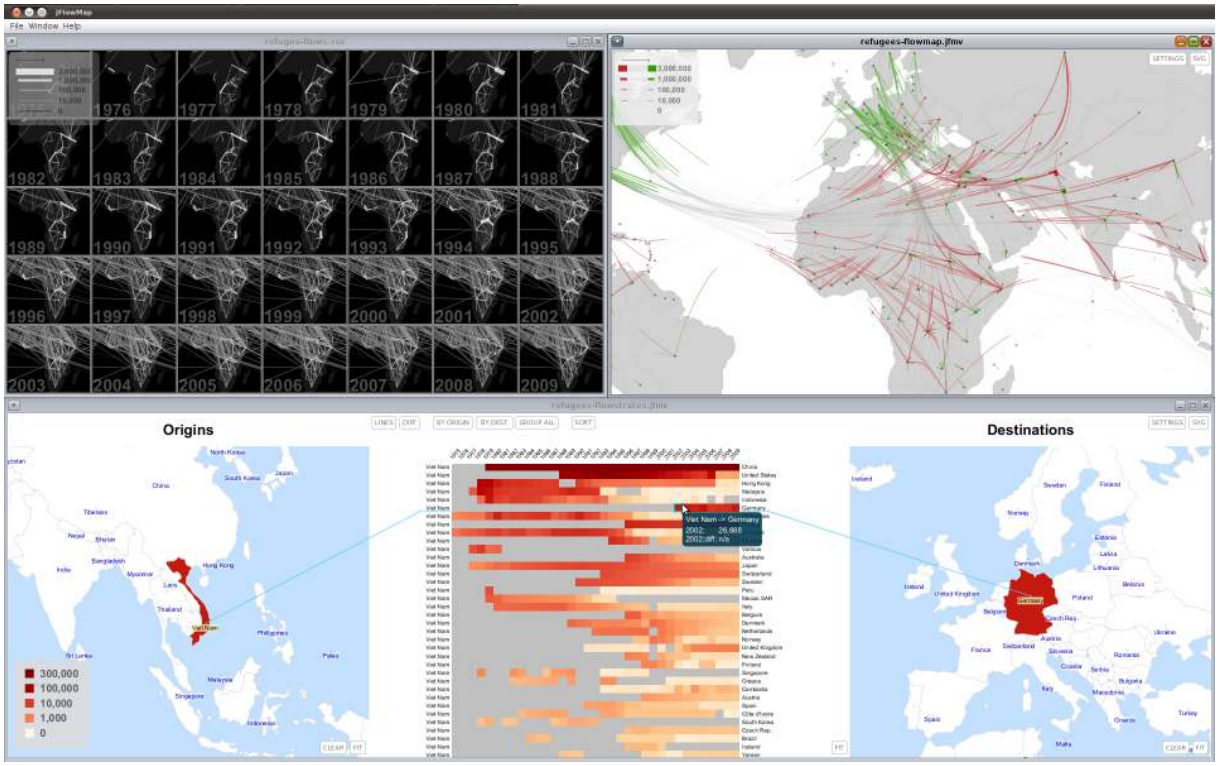


Figure 9.2: JFlowMap: the open-source tool we developed for temporal OD-data visualization.

– Employers and workers on freelancer.com [Mill, 2011] (flow map)

– Export flows from the state of Louisiana to other parts of the world

– Young people coming to study in the US from different countries (flow map)

– Mobility of university graduates in Australia (Flowstrates)

– Connections within the human brain (flow map, bundling) [Böttger et al., 2012]

Despite the fact that our tool is a rough research prototype it has been provoking interest of many people working in a broad range of domains. This shows that the analysis of temporal OD-data is important for a large number of problems and that tools for visual exploration of such datasets have a great potential in supporting data analysts in addressing these problems.

* * *

We hope that our work on this thesis will eventually lead to the development of more comprehensive and universal solutions for the visualization of temporal OD-data which can help analysts working in various domains to gain insight into their data and can support decision makers in finding and implementing the right policies.

Bibliography

Abel, G. J. [2010]. *Estimation of international migration flow tables in Europe*. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(4), pages 797–825. ISSN 1467-985X. doi:10.1111/j.1467-985X.2009.00636.x. http://dx.doi.org/10.1111/j.1467-985X.2009.00636.x. (Cited on page 12.)

Adali, S., T. Eren, A. Turk, and S. Balcisoy [2012]. *HeatCube: Spatio-Temporal Data Visualization with GPU-Based Ray Tracing Volume Rendering*. *International Workshop on Visual Analytics*. (Cited on pages 60 and 63.)

Aigner, W., C. Kainz, R. Ma, and S. Miksch [2011a]. *Bertin was Right: An Empirical Evaluation of Indexing to Compare Multivariate Time-Series Data Using Line Plots*. In *Computer Graphics Forum*, volume 30, pages 215–228. Wiley Online Library. (Cited on page 105.)

Aigner, W., S. Miksch, H. Schumann, and C. Tominski [2011b]. *Visualization of time-oriented data*. Springer. (Cited on page 11.)

Amar, R., J. Eagan, and J. Stasko [2005]. *Low-Level Components of Analytic Activity in Information Visualization*. In *Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS'05)*, pages 15–15. Minneapolis, MN, USA. doi:10.1109/INFOVIS.2005.24. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1575773. (Cited on page 39.)

Andrienko, G., N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel [2011]. *A conceptual framework and taxonomy of techniques for analyzing movement*. *Journal of Visual Languages & Computing*. ISSN 1045926X. doi:10.1016/j.jvlc.2011.02.003. http://linkinghub.elsevier.com/retrieve/pii/S1045926X11000139. (Cited on pages 10, 42, 43, 44, 46, 59 and 108.)

Andrienko, G., N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M. Kraak, H. Schumann, and C. Tominski [2010]. *Space, time and visual analytics*. *International Journal of Geographical Information Science*, 24(10), pages 1577–1600. ISSN 1365-8816. doi:10.1080/13658816.2010.508043. http://www.informaworld.com/openurl?genre=article&doi=10.1080/13658816.2010.508043&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3. (Cited on page 2.)

Andrienko, G. L. and N. V. Andrienko [1999]. *Interactive maps for visual data exploration*. *International Journal of Geographical Information Science*, 13(4), pages 355–374. (Cited on page 72.)

Andrienko, N. and G. Andrienko [2006]. *Exploratory analysis of spatial and temporal data : a systematic approach*. Springer. ISBN 9783540259947. (Cited on pages 3, 38, 42, 43, 45 and 46.)

Andrienko, N. and G. Andrienko [2011]. *Spatial generalization and aggregation of massive movement data*. *Visualization and Computer Graphics, IEEE Transactions on*, 17(2), pages 205–219. (Cited on pages 11 and 24.)

Andrienko, N. and G. Andrienko [2012]. *Visual analytics of movement: An overview of methods, tools and procedures*. *Information Visualization*. ISSN 1473-8716, 1473-8724. doi:10.1177/1473871612457601. http://ivi.sagepub.com/lookup/doi/10.1177/1473871612457601. (Cited on page 2.)

Archambault, D., H. Purchase, and B. Pinaud [2011]. *Animation, Small Multiples, and the Effect of Mental Map Preservation in Dynamic Graphs*. *IEEE Transactions on Visualization and Computer Graphics*, 17, pages 539–552. ISSN 1077-2626. doi:10.1109/TVCG.2010.78. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5473226. (Cited on pages 80, 82 and 92.)

Aris, A. and B. Shneiderman [2007]. *Designing semantic substrates for visual network exploration*. *Information Visualization*, 6, pages 281–300. ISSN 1473-8716. doi:10.1145/1375935.1375938. http://portal.acm.org/citation.cfm?id=1375935.1375938. (Cited on page 71.)

Bahoken, F. [2011]. *Représentation graphique des matrices. Graphe et/ou carte des flux?* (Cited on page 22.)

Becker, R., S. Eick, and A. Wilks [1995]. *Visualizing network data*. *IEEE Transactions on Visualization and Computer Graphics*, 1(1), pages 16–28. ISSN 10772626. doi:10.1109/2945.468391. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=468391. (Cited on pages 25, 51 and 81.)

Beddoe, D. [1978]. *An alternative cartographic method to portray Origin-Destination data*. PhD Thesis, University of Washington. (Cited on page 17.)

Bederson, B. B., J. Grosjean, and J. Meyer [2004]. *Toolkit design for interactive structured graphics*. *Software Engineering, IEEE Transactions on*, 30(8), pages 535–546. (Cited on page 77.)

Berry, B. [1966]. *Essays on commodity flows and the spatial structure of the Indian economy.* Technical Report, DTIC Document. (Cited on page 70.)

Bertin, J. [1967]. *Sémiologie graphique les diagrammes, les réseaux, les cartes*. Editions Gauthier-Villars, Paris, Paris. (Cited on pages 40, 46 and 50.)

Black, A. [1990]. *The Chicago area transportation study: A case study of rational planning*. *Journal of Planning Education and Research*, 10(1), pages 27–37. (Cited on page 17.)

Black, W. [1973]. *Toward a factorial ecology of flows*. *Economic Geography*, 49(1), pages 59–67. (Cited on page 70.)

Blok, C. [2000]. *Monitoring Change: Characteristics of Dynamic Geo-spatial Phenomena for Visual Exploration*. In *Spatial Cognition II, Integrating Abstract Theories, Empirical Studies, Formal Methods, and Practical Applications*, pages 16–30. Springer-Verlag, London, UK, UK. ISBN 3-540-67584-1. http://dl.acm.org/citation.cfm?id=645973.675227. (Cited on pages 41 and 46.)

Boon, B., E. Johnson, M. Studer, G. Tsalidis, J. van Houten, and C. Vercauteren [2012]. *Resource Atlas: Global Resource Flow Data Visualization*. Technical Report, Universiteit Leiden and TU Delft. (Cited on page 111.)

Borkin, M., K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister [2011]. *Evaluation of artery visualizations for heart disease diagnosis*. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12), pages 2479–2488. (Cited on page 106.)

Bostock, M. [2012]. *Uber Rides by Neighborhood*. http://bost.ocks.org/mike/uberdata/. [Online; accessed 20-Dec-2012]. (Cited on page 49.)

Bostock, M., V. Ogievetsky, and J. Heer [2011]. *D^3 Data-Driven Documents*. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12), pages 2301–2309. (Cited on page 103.)

Böttger, J., A. Schäfer, G. Lohmann, A. Villringer, and D. Margulies [2012]. *Force-directed edge-bundling for the visualization of functional connectivity*. *Poster at OHBM 2012*. (Cited on page 112.)

Boyandin, I. [2010]. *JFlowMap: Flow map visualization tool*. `https://code.google.com/p/jflowmap/`. [Online; accessed 1-May-2013]. (Cited on page 77.)

Boyandin, I., E. Bertini, P. Bak, and D. Lalanne [2011]. *Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data*. *Computer Graphics Forum*, 30, pages 971–980. ISSN 01677055. doi:10.1111/j.1467-8659.2011.01946.x. `http://doi.wiley.com/10.1111/j.1467-8659.2011.01946.x`. (Cited on page 67.)

Boyandin, I., E. Bertini, and D. Lalanne [2010]. *Using Flow Maps to Explore Migrations Over Time*. *Geospatial Visual Analytics Workshop in conjunction with The 13th AGILE International Conference on Geographic Information Science*. (Cited on pages 25, 27 and 68.)

Boyandin, I., E. Bertini, and D. Lalanne [2012]. *A Qualitative Study on the Exploration of Temporal Changes in Flow Maps with Animation and Small-Multiples*. In *Computer Graphics Forum*, volume 31, pages 1005–1014. Wiley Online Library. (Cited on pages 79 and 86.)

Brewer, C. and M. Harrower [2009]. *Colorbrewer: Color Advice for Maps*. `http://colorbrewer2.org`. [Online; accessed 20-Dec-2012]. (Cited on page 72.)

Brodbeck, D. and L. Girardin [2012]. *TreeMap visualization tool*. `http://www.treemap.com/`. [Online; accessed 1-May-2013]. (Cited on page 51.)

Brunet, R. [1986]. *La carte-modèle et les chorèmes*. *Mappemonde*, 86(4), pages 2–6. (Cited on page 31.)

Buchin, K., B. Speckmann, and K. Verbeek [2011]. *Flow map layout via spiral trees*. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12), pages 2536–2544. (Cited on pages 21, 22 and 25.)

Burch, M., C. Vehlow, N. Konevtsova, and D. Weiskopf [2012]. *Evaluating partially drawn links for directed graph edges*. In *Proceedings of the 19th international conference on Graph Drawing*, pages 226–237. GD'11, Springer-Verlag, Berlin, Heidelberg. ISBN 978-3-642-25877-0. doi:10.1007/978-3-642-25878-7_22. `http://dx.doi.org/10.1007/978-3-642-25878-7_22`. (Cited on page 25.)

Card, S., J. Mackinlay, and B. Shneiderman [1999]. *Readings in information visualization: using vision to think*. Morgan Kaufmann. (Cited on pages 2, 4 and 5.)

Charmaz, K. [2006]. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage Publications Ltd. (Cited on page 81.)

Chen, C.-H., H.-G. Hwu, W.-J. Jang, C.-H. Kao, Y.-J. Tien, S. Tzeng, and H.-M. Wu [2004]. *Matrix visualization and information mining*. In *COMPSTAT 2004—Proceedings in Computational Statistics*, pages 85–100. Springer. (Cited on page 51.)

Chicago Area Transportation Study [1957]. *Chicago Area Transportation Study: brief description*. Chicago Area Transportation Study. (Cited on page 17.)

Chicago Area Transportation Study [1959]. *Chicago Area Transportation Study : final report in three parts (1959)*. Chicago Area Transportation Study. (Cited on page 18.)

Cleveland, W. and R. McGill [1987]. *Graphical perception: The visual decoding of quantitative information on graphical displays of data*. *Journal of the Royal Statistical Society. Series A (General)*, pages 192–229. (Cited on page 20.)

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein [2001]. *Introduction to algorithms*. MIT press. (Cited on page 8.)

Cui, W., H. Zhou, H. Qu, P. C. Wong, and X. Li [2008]. *Geometry-based edge clustering for graph visualization*. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6), pages 1277–1284. (Cited on page 29.)

De Chiara, D., V. Del Fatto, R. Laurini, M. Sebillo, and G. Vitiello [2011]. *A chorem-based approach for visually analyzing spatial data*. *Journal of Visual Languages & Computing*, 22(3), pages 173–193. (Cited on page 31.)

Degelsegger, A. and F. Gruber [2012]. *Bibliometric studies on ASEAN research output and ASEAN-EU cooperation*. Presentation held at 11 December 2012 in Brussels on the occasion of the closing event of the ASEAN-EU Year of Science, Technology and Innovation 2012. (Cited on page 111.)

Ding, H., G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh [2008]. *Querying and mining of time series data: experimental comparison of representations and distance measures*. *Proceedings of the VLDB Endowment*, 1(2), pages 1542–1552. (Cited on page 74.)

Dorling, D. [1996]. *Area cartograms, their use and creation*. *Concepts and techniques in modern geography*. (Cited on page 31.)

Ellis, G. and A. Dix [2006]. *An explorative analysis of user evaluation studies in information visualisation*. In *Proc. of the BELIV'06 workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7. BELIV'06. (Cited on page 80.)

Ellis, G. and A. Dix [2007]. *A taxonomy of clutter reduction for information visualisation*. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6), pages 1216–1223. (Cited on page 24.)

Ericsson, K. and H. Simon [1980]. *Verbal reports as data. Psychological review*, 87(3), page 215. (Cited on page 38.)

Ersoy, O., C. Hurter, F. Paulovich, G. Cantareiro, and A. Telea [2011]. *Skeleton-based edge bundling for graph visualization*. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12), pages 2364–2373. (Cited on page 29.)

Ersoy, O. and A. Telea [2011]. *Graph Edge Bundling by Medial Axes*. In *Proc. of the Sixteenth Annual Conference of the Advanced School for Computing and Imaging (ASCI/IPA/SIKS, Nov. 14–15, 2011, Veldhoven, The Netherlands)*. (Cited on page 29.)

Fabrikant, S. I., S. Rebich-Hespanha, N. Andrienko, G. Andrienko, and D. R. Montello [2008]. *Novel Method to Measure Inference Affordance in Static Small-Multiple Map Displays Representing Dynamic Processes*. *Cartographic Journal, The*, 45, pages 201–215. ISSN 00087041, 17432774. doi:10.1179/000870408X311396. http://openurl.ingenta.com/content/xref?genre=article&issn=0008-7041&volume=45&issue=3&spage=201. (Cited on pages 80 and 81.)

Farrugia, M. and A. Quigley [2011]. *Effective Temporal Graph Layout: A Comparative Study of Animation versus Static Display Methods*. *Information Visualization*, 10(1), pages 47–64. (Cited on pages 80 and 82.)

Fekete, J. and N. Henry [2009]. *Matrix Reordering Survey*. *Visualisation Summer School, Peking University*. (Cited on page 51.)

Francis, A. and J. Schneider [1984]. *Using computer graphics to map origin-destination data describing health care delivery systems*. *Social Science & Medicine*, 18(5), pages 405–420. (Cited on page 17.)

Gesler, W. [1986]. *The uses of spatial analysis in medical geography: a review*. *Social Science & Medicine*, 23(10), pages 963–973. (Cited on page 17.)

Gleicher, M., D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts [2011]. *Visual comparison for information visualization*. *Information Visualization*, 10(4), pages 289–309. (Cited on pages 59 and 81.)

Gray, J., S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh [1997]. *Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals*. *Data Mining and Knowledge Discovery*, 1(1), pages 29–53. (Cited on page 4.)

Griffin, A. L., A. M. MacEachren, F. Hardisty, E. Steiner, and B. Li [2006]. *A Comparison of Animated Maps with Static Small-Multiple Maps for Visually Identifying Space-Time Clusters*. *Annals of the Association of American Geographers*, 96, pages 740–753. ISSN 0004-5608, 1467-8306. doi:10.1111/j.1467-8306.2006.00514.x. http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.2006.00514.x. (Cited on pages 80, 81 and 92.)

Guo, D. [2009]. *Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data*. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pages 1041–1048. (Cited on pages 27 and 28.)

Guo, D., J. Chen, A. MacEachren, and K. Liao [2006]. *A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP)*. *Visualization and Computer Graphics, IEEE Transactions on*, 12(6), pages 1461 –1474. ISSN 1077-2626. doi:10.1109/TVCG.2006.84. (Cited on page 52.)

Hahsler, M., K. Hornik, and C. Buchta [2007]. *Getting things in order: an introduction to the R package seriation*. (Cited on page 51.)

Harris, R. [1999]. *Information graphics : a comprehensive illustrated reference*. Oxford University Press, New York. ISBN 9780964692503. http://www.oup.com/us/catalog/general/subject/Communication/VisualCommunication/?view=usa&ci=9780195135329. (Cited on page 68.)

Heer, J. [2012]. *Datavore, a small in-browser database engine written in JavaScript*. http://vis.stanford.edu/projects/datavore/. [Online; accessed 20-Dec-2012]. (Cited on page 104.)

Heer, J., S. K. Card, and J. A. Landay [2005]. *Prefuse: a toolkit for interactive information visualization*. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM. (Cited on page 77.)

Heer, J., N. Kong, and M. Agrawala [2009]. *Sizing the horizon*. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09*, page 1303. Boston, MA, USA. doi:10.1145/1518701.1518897. http://portal.acm.org/citation.cfm?doid=1518701.1518897. (Cited on pages 72 and 103.)

Holten, D. [2006]. *Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data*. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5), pages 741–748. (Cited on page 29.)

Holten, D., P. Isenberg, J.-D. Fekete, and J. Van Wijk, Jarke [2010]. *Performance Evaluation of Tapered, Curved, and Animated Directed-Edge Representations in Node-Link Graphs*. Research report. http://hal.inria.fr/hal-00696823. (Cited on page 19.)

Holten, D., P. Isenberg, J. Van Wijk, J., and J. Fekete [2011]. *An Extended Evaluation of the Readability of Tapered, Animated, and Textured Directed-Edge Representations in Node-Link Graphs*. In Press, I. (Editor), *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, page 8p. Hong Kong, China. http://hal.inria.fr/inria-00548180/en/. (Cited on pages 19 and 20.)

Holten, D. and J. J. van Wijk [2009a]. *Force-Directed Edge Bundling for Graph Visualization*. *11th Eurographics/IEEE-VGTC Symposium on Visualization (Computer Graphics Forum; Proceedings of EuroVis 2009)*, 28(3), pages 983–990. ISSN 01677055. doi:10.1111/j.1467-8659.2009.01450.x. http://blackwell-synergy.com/doi/abs/10.1111/j.1467-8659.2009.01450.x. (Cited on pages 20, 29 and 30.)

Holten, D. and J. J. van Wijk [2009b]. *A user study on visualizing directed edges in graphs*. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09*, page 2299. Boston, MA, USA. doi:10.1145/1518701.1519054. http://portal.acm.org/citation.cfm?doid=1518701.1519054. (Cited on page 19.)

Hurter, C., O. Ersoy, and A. Telea [2012]. *Graph Bundling by Kernel Density Estimation*. *Computer Graphics Forum*, 31. (Cited on page 29.)

Kaufmann, M. and D. Wagner [2001]. *Drawing graphs: methods and models*. Springer, Berlin, [etc.]. ISBN 3540420622 9783540420620. http://www.springerlink.com/openurl.asp?genre=issue&issn=0302-9743&volume=2025&issue=preprint. (Cited on pages 11 and 24.)

Keefe, D. F., M. Ewert, W. Ribarsky, and R. Chang [2009]. *Interactive Coordinated Multiple-View Visualization of Biomechanical Motion Data*. *IEEE Transactions on Visualization and Computer Graphics (IEEE Visualization 2009)*, 15(6), pages 1383–1390. (Cited on pages 92 and 95.)

Keim, D. [2001]. *An introduction to information visualization techniques for exploring very large databases.* Tutorial notes, Information Visualization'00 (Salt Lake City, UT, Oct. 9–13, 2000). (Cited on page 3.)

Keogh, E. J., H. Hochheiser, and B. Shneiderman [2002]. *An Augmented Visual Query Mechanism for Finding Patterns in Time Series Data*. In *Proceedings of the 5th International Conference on Flexible Query Answering Systems*, pages 240–250. FQAS '02, Springer-Verlag, London, UK, UK. ISBN 3-540-00074-7. http://portal.acm.org/citation.cfm?id=645424.652296. (Cited on page 72.)

Kern, R. and G. Rushton [1969]. *MAPIT: A Computer Program for Production of Flow Maps, Dot Maps and Graduated Symbol Maps*. *Cartographic Journal, The*, 6(2), pages 131–137. (Cited on page 17.)

Konjar, M., I. Boyandin, D. Lalanne, A. Lisec, and S. Drobne [2010]. *Using flow maps to explore functional regions in Slovenia*. *2nd International Conference on Information Society and Information Technologies - ISIT 2010, Dolen*. (Cited on page 111.)

Kraak, M. [2003]. *The Space-Time Cube Revisited from a Geovisualization Perspective*. In *Proceedings of the 21st International Cartographic Conference (ICC)*. (Cited on pages 60 and 63.)

Kraak, M., R. Edsall, and A. MacEachren [1997]. *Cartographic animation and legends for temporal maps: Exploration and or interaction*. *Proceedings of the 18th International Cartographic Conference*, 1. (Cited on page 103.)

Krempel, L. and T. Plümper [1999]. *International division of labor and global economic processes: an analysis of the international trade in automobiles*. *Journal of World-Systems Research*, V(3), pages 487–498. http://jwsr.ucr.edu/archive/vol5/number3/krempel/. (Cited on page 24.)

Krzywinski, M., I. Birol, S. J. Jones, and M. A. Marra [2011]. *Hive plots: Rational approach to visualizing networks*. *Briefings in Bioinformatics*. doi:10.1093/bib/bbr069. http://bib.oxfordjournals.org/content/early/2011/12/09/bib.bbr069.abstract. (Cited on page 54.)

Lam, H., T. Munzner, and R. Kincaid [2007]. *Overview Use in Multiple Visual Information Resolution Interfaces*. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), pages 1278–1285. ISSN 1077-2626. doi:10.1109/TVCG.2007.70583. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4376151. (Cited on page 72.)

Lambert, A., R. Bourqui, and D. Auber [2010]. *Winding roads: Routing edges into bundles*. In *Computer Graphics Forum*, volume 29, pages 853–862. Wiley Online Library. (Cited on page 29.)

Lavin, S. and R. Cerveny [1987]. *Unit-vector density mapping*. *Cartographic Journal, The*, 24(2), pages 131–141. (Cited on page 33.)

Lazar, J. [2010]. *Research methods in human-computer interaction*. Wiley, Chichester West Sussex U.K. ISBN 9780470723371. (Cited on page 86.)

Liu, L. [1995]. *PPFLOW: An interactive visualization system for the exploratory analysis of migration flows*. *Geographic Information Sciences*, 1(2), pages 118–123. (Cited on page 17.)

MacEachren, A., F. Boscoe, D. Haug, and L. Pickle [1998]. *Geographic visualization: designing manipulable maps for exploring temporally varying georeferenced statistics*. pages 87–94,. IEEE Comput. Soc. ISBN 0-8186-9093-3. doi:10.1109/INFVIS.1998.729563. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=729563. (Cited on page 81.)

MacEachren, A. M. [2004]. *How Maps Work: Representation, Visualization, and Design*. The Guilford Press. ISBN 1-57230-040-X. (Cited on pages 41 and 60.)

MacEachren, A. M., J. Thacher, and C. Reeves [1994]. *Some truth with maps: A primer on symbolization and design*. Association of American Geographers Washington, DC. (Cited on page 72.)

Mansmann, F., D. Keim, S. North, B. Rexroad, and D. Sheleheda [2007]. *Visual analysis of network traffic for resource planning, interactive monitoring, and interpretation of security threats*. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6), pages 1105–1112. (Cited on pages 57 and 103.)

Marble, D. F., Z. Gou, L. Liu, and J. Saunders [1997]. *Recent advances in the exploratory analysis of interregional flows*. *Innovations in GIS 4*, pages 75–88. (Cited on pages 2 and 70.)

Mill, R. [2011]. *Freelancer.com visualization*. http://www.stanford.edu/~roymill/freelancer.com/flow_map_city.html. [Online; accessed 20-Dec-2012]. (Cited on page 112.)

Mitchell, A. and ESRI [1999]. *The ESRI guide to GIS analysis*. 1st ed. Edition. Environmental Systems Research Institute. ISBN 9781879102064. (Cited on page 40.)

Morse, M. D. and J. M. Patel [2007]. *An efficient and accurate method for evaluating time series similarity*. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 569–580. ACM. (Cited on page 74.)

Nagel, T., E. Duval, A. Vande Moere, K. Kloeckl, and C. Ratti [2012]. *Sankey Arcs – Visualizing edge weights in path graphs*. *Computer Graphics Forum (Proceedings of EuroVis 2012)*. (Cited on page 50.)

Noguchi, T. and J. Schneider [1977]. *Data display techniques for transportation analysis and planning: an investigation of three computer-produced graphics*. *Transportation Planning and Technology*, 4(1), pages 23–36. (Cited on page 17.)

North, C. [2006]. *Toward measuring visualization insight*. *IEEE Computer Graphics and Applications*, 26, pages 6–9. ISSN 0272-1716. doi:10.1109/MCG.2006.70. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1626178. (Cited on page 80.)

Peuquet, D. J. [1994]. *It's About Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems*. *Annals of the Association of American Geographers*, 84(3), pages 441–461. ISSN 0004-5608, 1467-8306. doi:10.1111/j.1467-8306.1994.tb01869.x. http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1994.tb01869.x. (Cited on pages 41, 42, 46, 59 and 60.)

Phan, D., L. Xiao, R. Yeh, P. Hanrahan, and T. Winograd [2005]. *Flow Map Layout*. In *Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS'05)*, pages 29–29. (Cited on page 21.)

Proulx, P., A. Khamisa, and R. Harper [2010]. *Integrated Visual Analytics Workflow with GeoTime and nSpace*. *VAST 2010 Mini Challenge*. (Cited on page 63.)

Pupyrev, S., L. Nachmanson, S. Bereg, and A. Holroyd [2012]. *Edge routing with ordered bundles*. In *Graph Drawing*, pages 136–147. Springer. (Cited on pages 30 and 31.)

Pupyrev, S., L. Nachmanson, and M. Kaufmann [2011]. *Improving layered graph layouts with edge bundling*. In *Graph Drawing*, pages 329–340. Springer. (Cited on page 29.)

Purchase, H. and A. Samra [2008]. *Extremes are better: Investigating mental map preservation in dynamic graphs*. In *Diagrams 2008. Fifth International Conference on the Theory and Application of Diagrams*. LNAI, Springer Verlag. http://eprints.gla.ac.uk/35837/. Session 2.: Diagram Aesthetics and Layout (joint with VL/HCC). (Cited on page 82.)

Rae, A. [2009]. *From spatial interaction data to spatial interaction information: Geovisualisation and spatial structures of migration from the 2001 UK census*. *Computers, Environment and Urban Systems*, 33(3), pages 161–178. http://dx.doi.org/10.1016/j.compenvurbsys.2009.01.007. (Cited on pages 2 and 34.)

Ravenstein, E. [1885]. *The laws of migration*. *Journal of the Statistical Society of London*, 48(2), pages 167–235. (Cited on page 28.)

Richards J. Heuer, J. [1999]. *Psychology of Intelligence Analysis*. Central Intelligence Agency. ISBN 1929667-00-0. (Cited on page 92.)

Robertson, G., R. Fernandez, D. Fisher, B. Lee, and J. Stasko [2008]. *Effectiveness of Animation in Trend Visualization*. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), pages 1325–1332. ISSN 1077-2626. doi:10.1109/TVCG.2008.125. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4658146. (Cited on pages 80 and 81.)

Robinson, A. [1982]. *Early thematic mapping in the history of cartography*. University of Chicago Press Chicago. (Cited on pages 16 and 17.)

Robinson, A. H. [1955]. *The 1837 Maps of Henry Drury Harness*. *The Geographical Journal*, 121(4), pages pp. 440–450. ISSN 00167398. http://www.jstor.org/stable/1791753. (Cited on page 16.)

Rowe, F. and M. Bell [2012]. *Creating an integrated database for the analysis of spatial mobility in Chile*. Working Papers 2012/02, Queensland Centre for Population Research, School of Geography, Planning and Environmental Management, The University of Queensland, Brisbane. (Cited on page 111.)

Saffrey, P. and H. Purchase [2008]. *The "mental map" versus "static aesthetic" compromise in dynamic graphs: a user study*. In *Proceedings of the ninth conference on Australasian user interface - Volume 76*, page 85–93. AUIC '08, Australian Computer Society, Inc., Darlinghurst, Australia, Australia. ISBN 978-1-920682-57-6. http://dl.acm.org/citation.cfm?id=1378337.1378354. (Cited on page 82.)

Saraiya, P., C. North, and K. Duca [2005]. *An Insight-Based Methodology for Evaluating Bioinformatics Visualizations*. *IEEE Transactions on Visualization and Computer Graphics*, 11, pages 443–456. ISSN 1077-2626. doi:10.1109/TVCG.2005.53. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1432690. (Cited on page 80.)

Schmidt, M. [2006]. *Der Einsatz von Sankey-Diagrammen im Stoffstrommanagement*. Hochsch. Pforzheim. (Cited on page 21.)

Schneider, J. [1983]. *Mapping Origin/Destination data: Now we can "see" what's going on out there*. *ITE Journal*, 53(12). (Cited on page 17.)

Sedlmair, M., M. Meyer, and T. Munzner [2012]. *Design Study Methodology: Reflections from the Trenches and the Stacks*. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 18(12), pages 2431–2440. (Cited on page 38.)

Sharp, H., Y. Rogers, and J. Preece [2007]. *Interaction design : beyond human-computer interaction*. Wiley, Chichester; Hoboken, NJ. ISBN 9780470018668 0470018666. (Cited on page 38.)

Shneiderman, B. [1996]. *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations*. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–. VL '96, IEEE Computer Society, Washington, DC, USA. ISBN 0-8186-7508-X. http://dl.acm.org/citation.cfm?id=832277.834354. (Cited on pages 26 and 39.)

Shneiderman, B. and A. Aris [2006]. *Network Visualization by Semantic Substrates*. *IEEE Transactions on Visualization and Computer Graphics*, 12, pages 733–740. ISSN 1077-2626. doi:http://dx.doi.org/10.1109/TVCG.2006.166. http://dx.doi.org/10.1109/TVCG.2006.166. (Cited on page 71.)

Shneiderman, B. and M. Wattenberg [2001]. *Ordered treemap layouts*. In *Proceedings of the IEEE Symposium on Information Visualization 2001*, volume 73078. (Cited on page 51.)

Slocum, T., R. Sluter, F. Kessler, and S. Yoder [2004]. *A Qualitative Evaluation of MapTime, A Program For Exploring Spatiotemporal Point Data*. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39, pages 43–68. ISSN 0317-7173, 1911-9925. doi:10.3138/92T3-T928-8105-88X7. http://utpjournals.metapress.com/openurl.asp?genre=article&id=doi:10.3138/92T3-T928-8105-88X7. (Cited on pages 80 and 81.)

Speckmann, B. and K. Verbeek [2010]. *Necklace maps*. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6), pages 881–889. (Cited on page 49.)

Spencer, D. and T. Warfel [2004]. *Card sorting: a definitive guide*. *Document available online on 03.11.2011*. http://www.boxesandarrows.com/view/card_sorting_a_definitive_guide. (Cited on page 87.)

Stefaner, M. [2010]. *Map your moves - A visual exploration of where New Yorkers moved in the last decade*. http://moritz.stefaner.eu/projects/map%20your%20moves/. [Online; accessed 1-May-2013]. (Cited on pages 31 and 32.)

Stolte, C., D. Tang, and P. Hanrahan [2003]. *Multiscale visualization using data cubes*. *Visualization and Computer Graphics, IEEE Transactions on*, 9(2), pages 176–187. (Cited on page 4.)

Telea, A. and O. Ersoy [2010]. *Image-Based Edge Bundles: Simplified Visualization of LargeGraphs*. *Computer Graphics Forum*, 29(3), pages 843–852. doi:http://dx.doi.org/10.1111/j.1467-8659.2009.01680.x. http://www.cs.rug.nl/~alext/PAPERS/EuroVis10/paper.pdf. (Cited on page 29.)

Thompson, W. and S. Lavin [1996]. *Automatic Generation of Animated Migration Maps*. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 33(2), pages 17–28. ISSN 0317-7173. doi:10.3138/P4Q4-1220-3774-M430. http://utpjournals.metapress.com/openurl.asp?genre=article&id=doi:10.3138/P4Q4-1220-3774-M430. (Cited on pages 17, 33 and 81.)

Thornthwaite, C. and H. Slentz [1934]. *Internal migration in the United States*. University of Pennsylvania Press. (Cited on pages 21 and 22.)

Tobler, W. [1970]. *A computer movie simulating urban growth in the Detroit region*. *Economic geography*, 46, pages 234–240. (Cited on page 23.)

Tobler, W. [1981]. *A model of geographical movement*. *Geographical Analysis*, 13(1), pages 1–20. (Cited on pages 17, 18, 28, 30 and 33.)

Tobler, W. [1987]. *Experiments in migration mapping by computer*. *The American Cartographer*, 14(2), pages 155–163. (Cited on pages 20, 22, 26 and 27.)

Tobler, W. [1995]. *Migration: Ravenstein, Thornthwaite, and beyond*. *Urban Geography*, 16(4), pages 327–343. (Cited on page 29.)

Tominski, C., P. Schulze-Wollgast, and H. Schumann [2005]. *3d information visualization for time dependent data on maps*. In *Information Visualisation, 2005. Proceedings. Ninth International Conference on*, pages 175–181. IEEE. (Cited on page 63.)

Tufte, E. R. [1986]. *The visual display of quantitative information*. Graphics Press, Cheshire, CT, USA. ISBN 0-9613921-0-X. (Cited on pages 17, 21 and 32.)

Tukey, J. [1977]. *Exploratory Data Analysis*. Addison-Wesley, Reading MA. (Cited on page 3.)

UNHCR [2010]. *2009 Global Trends: Refugees, Asylum-seekers, Returnees, Internally Displaced and Stateless Persons*. http://www.unhcr.org/4c11f0be9.pdf. (Cited on pages 12 and 68.)

van de Ven, B. [2007]. *Algorithms for flow maps*. (Cited on pages 22, 23 and 25.)

Van Liere, R. and W. De Leeuw [2003]. *Graphsplatting: Visualizing graphs as continuous fields*. *Visualization and Computer Graphics, IEEE Transactions on*, 9(2), pages 206–212. (Cited on page 34.)

Voorhees, A. [1956]. *A general theory of traffic movement*. (Cited on page 12.)

Ware, C. [2012]. *Information visualization: perception for design*. Morgan Kaufmann, San Francisco CA. (Cited on page 2.)

Ware, C. and R. Bobrow [2004]. *Motion to support rapid interactive queries on node–link diagrams*. *ACM Transactions on Applied Perception*, 1, pages 3–18. ISSN 15443558. doi:10.1145/1008722.1008724. http://portal.acm.org/citation.cfm?doid=1008722.1008724. (Cited on pages 82 and 92.)

Weiner, E. [1986]. *Urban transportation planning in the United States: an historical overview (revised edition)*. Technical Report, Department of Transportation, Washington, DC (USA). Office of the Assistant Secretary for Policy and International Affairs. (Cited on page 12.)

Wilkinson, L. [2005]. *The grammar of graphics*. Springer. (Cited on page 4.)

Wittick, R. [1976]. *A computer system for mapping and analyzing transportation networks*. *Southeastern Geographer*, 16(1), pages 74–81. (Cited on page 17.)

Wood, J. and J. Dykes [2008]. *Spatially Ordered Treemaps*. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), pages 1348–1355. (Cited on pages 31, 32, 57 and 103.)

Wood, J., J. Dykes, and A. Slingsby [2010]. *Visualisation of Origins, Destinations and Flows with OD Maps*. *Cartographic Journal, The*, 47(2), pages 117–129. ISSN 00087041. doi:10.1179/000870410X12658023467367. http://openurl.ingenta.com/content/xref?genre=article&issn=0008-7041&volume=47&issue=2&spage=117. (Cited on pages 52 and 53.)

Wood, J., A. Slingsby, and J. Dykes [2011]. *Visualizing the Dynamics of London's Bicycle-Hire Scheme*. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(4), pages 239–251. (Cited on page 34.)

Wu, H.-M., S. Tzeng, and C.-h. Chen [2008]. *Matrix Visualization*. In *Handbook of Data Visualization*, pages 681–708. Springer Handbooks Comp.Statistics, Springer Berlin Heidelberg. ISBN 978-3-540-33036-3. doi:10.1007/978-3-540-33037-0_26. http://dx.doi.org/10.1007/978-3-540-33037-0_26. (Cited on page 51.)

Yi, J. S., Y.-a. Kang, J. T. Stasko, and J. A. Jacko [2008]. *Understanding and characterizing insights*. In *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV '08*, page 1. Florence, Italy. doi:10.1145/1377966.1377971. http://portal.acm.org/citation.cfm?doid=1377966.1377971. (Cited on page 80.)

Zhang, J. [2011]. *NYC taxi flows*. http://www-cs.ccny.cuny.edu/~jzhang/NYCTaxiFlow.htm. [Online; accessed 20-Dec-2012]. (Cited on page 111.)

Curriculum Vitae

Personal Details

Lastname	Boyandin
Firstname	Ilya
Birthday	November 1st, 1978
Address	Rue de la Neuveville 30, 1700 Fribourg
Hometown	Saint-Petersburg, Russia
Nationality	Russia

Education

MSc Computer Science, St. Petersburg State University, Russia, 2003
Thesis title: Statistical Query Transformations for Question Answering in the Web
Summary: Developed an improvement for a state-of-the-art algorithm transforming natural language questions into search engine queries achieving a better quality of question answering.

Publications

A Qualitative Study on the Exploration of Temporal Changes in Flow Maps with Animation and Small-Multiples, Ilya Boyandin, Enrico Bertini, Denis Lalanne.
Computer Graphics Forum, International Journal of the Eurographics Association, Eurographics/IEEE-VGTC Symposium on Visualization, Vienna, Austria, June 2012.

Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data., Ilya Boyandin, Enrico Bertini, Peter Bak, Denis Lalanne.
Computer Graphics Forum, International Journal of the Eurographics Association, Eurographics/IEEE-VGTC Symposium on Visualization, Bergen, Norway, June 2011.

Visualizing migration flows and their development in time: flow maps and beyond., Ilya Boyandin, Enrico Bertini, Denis Lalanne.
IEEE VisWeek Doctoral Colloquium, Salt Lake City, United States, October 2010.

Visualizing the World's Refugee Data with JFlowMap, Ilya Boyandin, Enrico Bertini, Denis Lalanne.
Poster Abstract at Eurographics/IEEE-VGTC Symposium on Visualization, Bordeaux, France, June 2010.

Using Flow Maps to Explore Migrations Over Time, Ilya Boyandin, Enrico Bertini, Denis Lalanne.
Workshop in Geospatial Visual Analytics: Focus on Time, GeoVA(t), Guimarães, Portugal, May 2010.

Statistical Query Transformations for Question Answering in the Web (in Russian), slides (in English), Ilya Boyandin, Igor Nekrestyanov.
Fifth Russian Conference on Digital Libraries (RCDL'2003), St. Petersburg, Russia, October 2003.

Work Experience

PhD Student, Assistant, University of Fribourg, Switzerland – *since April 2009*

- Working on a thesis on visualization of temporal changes in origin-destination data (e.g. migrations).
- Developing web-based visualizations of AidData.
- Developed Flowstrates, a novel approach for visualizing and exploring temporal origin-destination data.
- Carried out a qualitative user study comparing animated and small-multiple representations of changes in flow maps.
- Developed JFlowMap, an experimental tool for the visualization of spatial interactions.
- Contributed to the BirdEye visualization library developed at the UN Centre for Advanced Visual Analytics.
- Assisting in courses on Web technologies and Functional programming. Tutoring, giving occasional lectures, preparing materials, building supporting websites and utilities.
- Technical maintenance of the DIVA research group website.
- Helping to organize and tutoring in workshops on programming and computer graphics for students and school children.

Senior Software Engineer, Technical Team Leader, IT Department, University of Applied Sciences FH Joanneum, Graz, Austria – *March 2007 - April 2009*

- Maintained and developed web applications used by the students, lecturers and employees of the university for the online administration.
- Maintained the web and database server infrastructure for the online administration.
- Designed and developed a web application for collaborative data collection and consolidation which provided a statistical overview of study- and research-relevant indicators.

Software Engineer/Research Assistant, Dept of Information Design, University of Applied Sciences FH Joanneum, Graz, Austria – *September 2005 - March 2007*

- Designed and developed CGVis, a visualization tool facilitating hierarchical clustering, zooming and animation for the exploration of multidimensional datasets.
- Designed and developed a standalone and a web version of a proteomic data classification tool implementing a cancer diagnosis method based on mass-spectrometry data facilitating multi-step feature reduction and SVM classification. The tool achieves 99% classification accuracy on the NCI Cancer SELDI-TOF study dataset.
- Participated in the development of a presentation management tool for the information screens installed at the university. Developed the visual layout editor for arranging multimedia objects on the screen and the schedule editor similar to calendar in Outlook.
- Improved the implementation of an algorithm detecting the behavior type of a user looking at a web page based on the real-time eye-tracking data.

Software Engineer, Ecofinance Finanzsoftware & Consulting GmbH, Graz, Austria – *June 2004 - August 2005*

- Participated in the development of a web based treasury system for Deutsche Bahn and Commerzbank.
- Implemented the Java infrastructure and XSLT stylesheets for the runtime generation of the front end UI code from XML sources.
- Evaluated and optimized the performance of the XSL transformations.
- Implemented support for long running jobs on the server-side of the system.
- Developed the context-sensitive help for the system and the help authoring infrastructure based on DocBook.

Software Engineer, Elbrus MCST (by contract with Sun Microsystems), St. Petersburg, Russia – *August 2003 - May 2004*

- Worked in the Sun's Java Swing UI library maintenance team. Was responsible for fixing bugs and implementing requests for enhancements in the button classes (JButton, JRadioButton, JCheckBox, etc). Fixed a total of about 50 bugs in the Sun JDK.

Software Engineer, Aloha, St. Petersburg, Russia – *January 2001 - July 2003 (during studies)*

- Designed and developed an e-commerce system with order tracking, credit card processing, back-office, inventory, statistical reports, etc. Developed the whole system from scratch, supported and customized it adapting it to changing requirements. The system is still in use on several e-commerce websites.

Web Developer, ALife, St. Petersburg, Russia – *March 2000 - January 2001 (during studies)*

- Participated in the development of a system of intelligent agents capable of chatting to visitors of a website in a natural language and promoting its products. Implemented a highly dynamic web-interface for the subsystem that controlled the chats and let operators intervene in a chat if a bot was in trouble.

Teacher, Anichkov Lyceum, St. Petersburg, Russia – *September 1998 - April 2000 (during studies)*

- Taught school children programming.

Spoken Languages

- Russian (mother tongue)
- English (fluent)
- German (fluent)
- French (intermediate)