UNIVERSITY OF ZURICH

MASTER THESIS – GEO 511

Evaluating visualisation types for pattern and anomaly discovery in large origin-destination data sets – a case study with global postal data

Author
Anne WEGMANN
04-350-781

Advisor
Dr. Arzu ÇÖLTEKIN

Co-Advisors
Dr. Martin TOMKO
Department of Infrastructure Engineering
The University of Melbourne, Victoria 3010 Australia
tomkom@unimelb.edu.au

Dr. José ANSÓN
Universal Postal Union
Case Postale, 3015 Bern
jose.anson@upu.int

Faculty Member
Prof. Dr. Robert WEIBEL

GIScience Center
Department of Geography

November 19, 2016

"The most beautiful thing we can experience is the mysterious. It is the source of all true art and science."

Albert Einstein

Abstract

This thesis evaluated selected visualisation types found in the network- and flow visualisation literature based on a theoretical assessment, a user requirements analysis and a usability study conducted with experts. The study focused on the visual exploration of global postal origin-destination (OD) data, with a main interest in visual pattern and anomaly discovery. There are spatial and temporal patterns in the postal OD data as well as anomalies that can be caused by disruptions related to extreme weather events, political crises, technical issues, etc. The analysis of spatio-temporal patterns in postal traffic can provide insights about global communication through the postal network and thus may reveal certain behavioural patterns as well as information about the underlying structures. The detection of anomalies may help identifying events that impact the postal network and build knowledge if those anomalies can be related to extreme weather events, political crises, etc.

Identifying visualisation methods that help to analyse large OD data sets is challenging. To address this issue, this thesis investigated how spatial properties of the postal flow network and their encoding affect the analysts' ability to visually detect spatio-temporal patterns and anomalies. It assessed the potential and limitations of different visualisation types to find out which ones should be applied or combined. It further investigated the applicability of the selected visualisation types for detecting cyclical patterns and analysed whether the tested visualisations helped the experts to explain detected anomalies.

Results show that the granularity of the spatial units, the level of connectivity as well as the distance between origins and destinations affect a) the visualisation of large OD data sets and b) the analysts' ability to visually detect patterns and anomalies. Findings show that there is a trade-off between generalisation for better visual perception and information loss. All visualisation types have their advantages and drawbacks; static visualisations can provide a quick overview, while interactive visualisations enable the analysts to explore the data in more depth. Interactive visualisations were often, but not necessarily always, preferred by the participants of the user study. Of the tested visualisations, only the time series graph enabled the participants to detect cyclical patterns in the case study at hand. Some visualisations helped the experts to reflect about the reasons for the detected anomalies, but this depended strongly on their prior knowledge and interest. Additional information from other data sources is needed for a reliable interpretation of underlying reasons.

Keywords: Origin-destination data, flow visualisation, spatio-temporal visualisation, visual data exploration, visual pattern and anomaly detection

Acknowledgements

First and foremost, I would like to thank my advisor Dr. Arzu Çöltekin for her invaluable support, her encouraging words and her patience throughout this last year. Another big thank you goes out to my advisors Dr. Martin Tomko and Prof. Dr. Robert Weibel for always supporting my ideas and sharing their knowledge during technical struggles.

My thanks also go to Dr. José Ansón for inviting me to work with the UPU in Bern and become part of the team.

I would also like to thank all the postal experts who were part of the interview sessions during the requirements analysis or who spent half a day with my usability study. Thank you, José Ansón, Anne-Claire Blet, Iacopo Iacopini, Jochen Krueck, Daniel Nieto, Lukasz Piotrowski, Stephanie Rogers, Alexander Thern-Svanberg.

I thank Stefan Winzenried for letting me work flexible when needed, Sebastian Schäufele for lending me his screen and sharing early morning coffees at work and my fellow students for their support with any cartographic or coding related question I had.

My special thanks go to Thomas Lüscher for his continuos support throughout this last year, his proofreading, his computer when mine broke down, all his patience and his encouraging words in difficult moments.

Finally, I thank my family and friends who have not only supported me during this last year, but during the last five years following my decision to go back to school and get this second degree.

To all of you, a heartfelt thank-you.
This thesis would not have possible without you.

Personal Declaration

I, Anne WEGMANN, hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Signed:

Due to reasons of data sensitivity, the scales of the graphs are not shown in this publicly available version of the thesis.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

B2C	**B**usiness to **C**ustomer
CW	**C**alendar **W**eek
EDI	**E**lectronic **D**ata **I**nterchange
EmIS	**Em**ergency **I**nformation **S**ystem
IMPC	**I**nternational **M**ail **P**rocessing **C**enter
IPN	**I**nternational **P**ostal **N**etwork
LOD	**L**evel **O**f **D**etail
OD	**O**rigin **D**estination
ODM	**O**rigin **D**estination **M**atrix
SUS	**S**ystem **U**sability **S**cale
UN	**U**nited **N**ations
UPU	**U**niversal **P**ostal **U**nion
ZCTA	**Z**IP **C**ode **T**abulation **A**rea

Chapter 1

Introduction

1.1 Motivation

Communication between people of different regions and countries is important in all societies. Hence, it is not surprising that one of the oldest physical flow networks is the International Postal Network (IPN). Already in 1874, the Universal Postal Union (UPU), a UN specialised organisation[1], was established in order to support good co-operation between actors in the postal sector. The IPN plays a vital role in personal and commercial communication around the world (Hristova et al., 2016). With recent advances in technology, it is now possible to capture and analyse data about these global flows. This opens up various fields for the investigation of this data in terms of flow dynamics and spatio-temporal patterns and anomalies that are meaningful for postal experts who are analysing the IPN (Anson and Helble, 2014).

The analysis of spatio-temporal patterns in postal traffic can provide new insights about the nature of postal connections between different places within the IPN (and thus might reveal certain information about social/behavioural patterns), and potentially also about the underlying structures, e.g. transportation infrastructure (Anson and Helble, 2014). The postal network is increasingly used for goods transportation, mainly due to e-commerce development, but also because personal communication has partly shifted to digital media (Anson and Helble, 2014). This general growth in the postal sector can also be seen in the increasing number of 'direct to consumer' shipments carried out by private parcel delivery companies (Singh et al., 2006). Along with bigger volumes in shippings, higher volatility in the data is expected by postal experts.

In the UN Hyogo Framework for Action 2005-2015 by the United Nations (2007), the protection and strengthening of public facilities and physical infrastructure is mentioned as one of the key activities for reducing the underlying risk factors of disasters related to geological events, weather, climate change, etc. The detection of anomalies in postal shipment patterns and their relationship to extreme weather events, political crises, health crises, technical issues, etc. can help to identify events that impact the postal network. Further, spatio-temporal pattern detection can help to detect periodic events happening at a specific time, e.g. Christmas, holiday seasons, etc., or showing patterns regarding seasonality, changes in flow dynamics

[1] http://www.upu.int

between weekdays and weekends, etc. Again, anomalies can indicate disruptions of different kinds. Spatial patterns might change over time and temporal patterns might be unevenly distributed over space.

Apart from this specific use case of postal data, there are various other types of origin-destination (OD) data for which visual exploration is equally beneficial. The visual representation of different types of flows is important in many applications and numerous studies have been conducted in this field. Flows include for example migration flows, traffic flows, the movement of animals or diseases or flows of various types of goods, but also information flows such as e-mails or social networks (Yang et al., 2016; Cui et al., 2008). Studying one type of flow may have implications also for other domains.

While there are a number of benefits in the visual representation of such data for analysis, there are challenges in identifying the methods that are most appropriate for the specific user group as well as the underlying data. Many research papers suggest using visual analytics for analysing large data sets, because they combine automated analysis techniques (where computers are capable) with interactive visualisations (where humans are competent). This helps the user to understand and reason about large and complex data sets (Keim et al., 2010). However, there are only few scientific studies with real-life use cases that include the intended user groups in the development and evaluation of various visualisation methods.

While other types of flows and their visual representation have been investigated in various studies (cf. section 2.3), the data about postal traffic that is collected by the UPU has not yet been adapted for visual exploration. There has only been one study focusing on the IPN, in which the postal flow network is compared to other global flow networks (Hristova et al. (2016), see also 2.4). This thesis contributes to ongoing research in the field of data visualisation and visual analytics by following an explorative approach to find out which visualisation types are promising for analysing this data. It will provide a necessary first assessment of the potential of the visualisation types tested in situ, opening new paths for further investigation of specific visualisation types.

1.2 Research Objectives

Visual analytics combines automated analysis techniques with interactive visualisations, as is discussed in Chapter 2.1 (Keim et al., 2010). As mentioned above, the contribution of this thesis lies in the visualisation part. On the basis of two case studies with the provided postal data, different visualisation types for representing flows in large postal OD data are evaluated together with postal experts. The aim is to find out which of the proposed visualisation types help the experts to analyse patterns and detect anomalies in the data they are investigating. The users are familiar with the data and content, however they are not programmers or specialised in data analytics. This means that they need tools that let them explore the data without requiring extensive customisation thereof. Hence, using existing tools for the implementation of certain types of visualisations is possibly more appropriate for this user group than programming.

Studies for evaluating such visualisations in a real-world setting are rare, but involving the actual users in the development and evaluation of such tools is important for understanding the actual requirements and needs of experts for visual data exploration and analysis.

The amount of data on global postal traffic that has become available is extensive (Anson and Helble, 2014). As it is generally the case for such large sets of OD data, it is crucial to find appropriate methods to analyse them in order to extract valuable information (Keim et al., 2010). Different visualisation types that are used in visual analytics are examined in the frame of this thesis, with a particular focus on the visual detection and analysis of meaningful spatio-temporal patterns and anomalies in the postal traffic data.

In contrast to other work which investigated rather local data sets (e.g., Brunsdon, Corcoran, and Higgs, 2007; Andrienko and Andrienko, 2010; Kelly et al., 2012), the focus of this thesis lies on a much larger spatial scale, i.e. global flows of postal shipments. Similarly to the study by Wood, Dykes, and Slingsby (2010), the aim is to visually represent a large OD data set and enable its visual exploration.

Results from the interviews and the online survey conducted during the requirements analysis suggest that postal experts are aware of certain patterns and anomalies in the postal data. They formulate various hypotheses regarding the reasons behind those patterns and anomalies, such as trade connections, seasonality, natural disasters, strikes, etc. (see also 4.2). Based on two case studies, the aim is to suggest visualisation types to help postal experts "accept" or "reject" those hypotheses and provide a set of recommendations for exploring and learning about the postal data collected at UPU.

1.3 Research Questions and Hypotheses

As stated above, the aim of this thesis is to examine data visualisation methods such as graph visualisations (e.g., Herman, Melancon, and Marshall, 2000), flow maps (e.g., Andrienko and Andrienko, 2011; Wood, Slingsby, and Dykes, 2011), or origin-destination matrices (ODM or OD matrices) (e.g., Wood, Dykes, and Slingsby, 2010) in terms of detecting and analysing meaningful spatio-temporal patterns and anomalies in the postal traffic data. The evaluation of the selected methods will be based on theoretical considerations, specific data-related constraints and expert interviews.

Based on the aforementioned study by Hristova et al. (2016), several observations can be listed:

1. There are spatial and temporal patterns in the postal OD data and they are visible at different scales.

2. Both spatial and temporal patterns change over time, for example due to shifts in e-commerce development, infrastructure, etc.

3. There are spatial and temporal anomalies in the data, caused by technical issues or disruptions related to extreme weather events, political crises etc.

After conducting a requirements analysis with UPU experts, the selected methods are applied to the postal data to find out which spatial and temporal patterns and anomalies can be detected. In order to explore the potential of various visualisation types, this study proposes the following research questions (RQ):

RQ 1: How do the spatial properties of the postal flow network and their encoding influence analysts' ability to visually interpret the spatio-temporal patterns and detect anomalies?

Three main aspects are analysed:

1. The impact of the *encoding* of origins and destinations (the granularity of spatial units) on the ability to detect spatial and temporal patterns and anomalies

2. The impact of *connectivity* on the spatial an temporal patterns and the ability of detecting spatial and temporal patterns and anomalies

3. The impact of *distance* between origins and destinations on the spatial an temporal patterns, i.e. whether the applied methods reveal spatial autocorrelation

Based on previous findings about using aggregation in the visual exploration of data (Thomas and Cook, 2005; Andrienko and Andrienko, 2010), the thesis hypothesises that granularity of the spatial units affects the visualisation and detection of patterns and anomalies (Hypothesis 1a).

In accordance with previous findings, the thesis assumes that connectivity affects the visualisation of the postal data as well as the ability to analyse and detect patterns and anomalies in the data (Andrienko and Andrienko, 2010; Wood, Dykes, and Slingsby, 2010) (Hypothesis 1b).

Further, the thesis assumes that spatial autocorrelation appears in the postal traffic data, as is found in many other geographic flows (Tobler, 1970; Wood, Dykes, and Slingsby, 2010) (Hypothesis 1c).

RQ 2: Which visualisation types should be applied or combined to detect and analyse spatio-temporal patterns and anomalies in large postal OD data, and why?

1. What are potential and limitations of the implemented visualisation types?

2. How do the selected visualisation types compare to one another or complement one another?

Visualisation approaches that have been used in the analysis of other flow networks can be applied to the postal OD data for the detection of patterns and anomalies. The thesis hypothesises that some visualisation types are well suited for showing a global overview of the flow structures (e.g., the flow map (Wood, Slingsby, and Dykes, 2011)) while others are less applicable for analysing data sets with large numbers of spatial interactions (e.g., the choropleth map (Sander et al., 2014)) (Hypothesis 2a).

Interactive visualisations can reveal more detailed information about patterns and anomalies than static ones, however the latter can provide a quick first overview and thus a combination of different visualisation types is favourable (Shneiderman, 1997; Roberts, 2007; Keim et al., 2010) (Hypothesis 2b).

The study further assumes that interactive visualisations are preferred by the participants (Hegarty et al., 2009) (Hypothesis 2c).

RQ 3: How suitable are the selected visualisation methods for detecting cyclical patterns and anomalies in the postal data?

Cyclical patterns and anomalies in the postal data can be detected on various spatial and temporal scales and indicate underlying causes such as weekends or Christmas season. Based on previous research, the study expects times series to be practical for showing cyclical patterns in the postal data, while the other tested visualisation types are less useful for this purpose (Keim et al., 2010) (Hypothesis 3).

RQ 4: How can the detected anomalies be explained by the applied visualisation methods and additional (semantic) information (e.g. seasonality, socioeconomic development, conflicts, etc.)?

The thesis hypothesises that the applied visualisation types can help to understand underlying processes (Andrienko and Andrienko, 2010), though additional variables (e.g. about socioeconomic development, seasonality, conflicts, etc.) might be needed for a reliable interpretation (Hypothesis 4).

1.4 Thesis Organization

In order to address the above mentioned research questions, the thesis is structured in the following way: Chapter 2 provides the theoretical background to this study by a literature review, discussing relevant work to develop the hypotheses. Chapter 3 discusses the reviewed visualisation types and the methodological aspects of their implementation as well as the overall study design. Chapter 4 presents the results obtained at different steps during the study and Chapter 5 discusses the results in relation to the research questions and literature. Chapter 7 concludes with a final summary of the main findings and provides an outlook for further research.

Chapter 2

Related Research

2.1 Visual Analytics

Visualising data is crucial in order to enable an analyst to reason about the data, extract relevant information and obtain knowledge (Andrienko and Andrienko, 2010). Three main attributes are needed in order to analyse data and its interactions in time and space and to extract valuable information from it: the thematic information about the phenomenon, the location of the observation and its time record (Sinton, 1977). Today, the acquisition and availability of data is not the main problem anymore. With the development of new technologies, vast amounts of complex data have become accessible (Andrienko and Andrienko, 2011). Hence, the focus shifts to the methods for exploring this data in order to extract reliable information and knowledge from it (Keim et al., 2010). Visual analytics methods combine database technologies, computerised data processing and computational analysis methods with visual exploration and analysis. They are taking into account not only the strengths of electronic data processing, but also the user's knowledge and inference capability, which helps to understand and reason about the often very large and complex data sets (Andrienko and Andrienko, 2007; Sun et al., 2013; Keim et al., 2010). Visualisations that are utilised in visual analytics include interactive features (Keim et al., 2010). This supports users in analysing the data in an exploratory way, i.e. they can test different scenarios and explore different possibilities. The analysts can vary specific parameters, select a subset of the data to display, or chose a certain spatial or temporal resolution (Roberts, 2004).

The development of an environment for the visual exploration of OD data, i.e. the investigation of movement of objects or people in time and space, can be useful for pattern detection, information extraction and knowledge building (Wood, Dykes, and Slingsby, 2010). As discussed above, the human visual-cognitive abilities play a crucial part in visual analytics. Humans learn more efficiently and effectively within a visual environment in comparison to using text or working in a numerical setting (e.g., Lewalter, 2003). Spatial connections between objects enhance the analysts' awareness of similar objects and may thus help to better memorise associated information (Lloyd, 1997, and Bertin, as cited in Hernandez (2007)). In the case of spatial and spatio-temporal data, geovisualisation methods aim at taking advantage of this visual and cognitive capability to enable the user to detect patterns or interpret visual cues (Hernandez, 2007).

To support the analysts in the visual investigation of data, it is important to create representations that display the task-relevant information in a salient way, which is traditionally also one of the core concepts of cartography. Especially with large data sets, it is equally important to minimise visual clutter of information (Hegarty et al., 2009).

2.2 Visualisation of Spatio-Temporal Data

The visual representation of flows between different geographic locations is important in order to enable their analysis (Yang et al., 2016). Flows between geographic locations include for example movement of people or animals, diseases, goods or knowledge, commuting behaviour or urban traffic (Tobler, 1987; Zhao, Forer, and Harvey, 2008; Slingsby, Kelly, and Dykes, 2014; Guo, 2007; Gilbert et al., 2005; Paci and Usai, 2009; Wood, Slingsby, and Dykes, 2011; Ferreira et al., 2013). As mentioned before, networks can consist of physical flows as well as information flows. The particular characteristic of postal data is that apart from flows of goods (e.g. from e-commerce), it also contains communication flows which are distributed via physical entities such as letters and parcels.

As mentioned above, Sinton (1977) defined three data attributes that are needed for the investigation of data and its interactions in time and space (thematic information, location, time record). Andrienko et al. (2010) represent this structure more formally as $S \times T \rightarrow$ A, where S = space/set of places, T = time/set of moments or intervals, A = thematic component (attribute). Hence, S and T are independent variables, whereas the data attributes are space- and time dependent. Based on this structure, questions about the spatial situation and the temporal variation can be asked, e.g. how spatial patterns of an attribute change over a certain time period or how attributes change over time at a certain place. Andrienko, Andrienko, and Gatalsky (2003) propose three classes of spatio-temporal data in relation to the changes happening over time: 1) existential changes (appearance and disappearance), 2) changes of the spatial properties (location, shape, orientation, etc.) and 3) changes of the thematic properties (qualitative attribute changes and increase or decrease of ordinal or numeric characteristics).

Especially for large and complex data sets, it is generally difficult to adequately represent the data without losing relevant information because some of the data elements overlap others, thus it is important to evaluate different methods for specific tasks and data sets (Andrienko and Andrienko, 2010; Wood, Dykes, and Slingsby, 2010). Andrienko and Andrienko (2010) further argue that existing approaches for visual analysis of large (movement) data should be systemised in a framework and Andrienko and Andrienko (2013) point out that ideally, visualisation techniques should not only allow users to analyse patterns, but to also document and share those findings.

2.2.1 Aggregation

A key strength of data visualisation lies in generalisation and abstraction of the underlying data (Andrienko and Andrienko, 2010). One way of generalising data and encouraging abstraction is the combination of several data elements into one unit, i.e. the aggregation of data elements. This reduces the amount of displayed information and thus facilitates visualisation and perception, but it also leads to a loss of information. Depending on the interest of an analyst, highlighting the main characteristics at the expense of losing fine-detailed information can be either an advantage or a disadvantage (Thomas and Cook, 2005; Andrienko and Andrienko, 2010). Especially in the case of large data sets, the visual representation of spatio-temporal structures requires a reduction of the data, e.g. through the selection of subsets or data aggregation (Rae, 2009).

Following a similar categorisation to other GIScience scholars, Fredrikson et al. (1999) suggested in a study about spatio-temporal data three types of aggregation: 1) spatial aggregation, 2) temporal aggregation and 3) categorical aggregation. This is similar to others talking about space-time attributes. In the case of spatial aggregation, data elements are aggregated to suitable spatial units, e.g. administrative units (Andrienko and Andrienko, 2010). For temporal aggregation, time is divided into time intervals and individual data elements are grouped according to those time intervals. Categorical aggregation groups data items together according to their attributes. Depending on the data that is being analysed, such categorical attributes can, for example, be the vehicle type as in Fredrikson et al. (1999) or, in the case of the postal data, the postal item class (Hristova et al., 2016).

If data is aggregated to time intervals, those can be represented as so-called sequenced snapshots. This way of storing spatio-temporal data basically represents a series of images, which, if shown in quick succession, enable the analyst to investigate the changes of objects over time at a certain place (Thomas and Cook, 2005).

2.2.2 Interactivity

Visual analytics methods combine interactive visualisations with automated analysis techniques (Keim et al., 2010). This allows the user to decide e.g. which part of the data he or she wants to explore in more detail. A basic principle for visual data exploration was introduced by Shneiderman (1997) by what he called the "The Visual Information Seeking Mantra: Overview first, zoom and filter, then details-on-demand". This lets the data analyst define to a certain level what he or she wants to see and visualise. Similar to this, Bertin (1983) specified three "levels of reading,": The elementary level (allowing the analyst to look at the information about a single data record), the intermediate level (showing summarised information about a group of data records), and the global level (providing an overview of all data elements).

According to Hegarty et al. (2009), researchers in the field of human-computer interaction state that users generally assume that dynamic computer displays are superior to static ones. In their study about people's intuitions regarding display effectiveness and the actual effectiveness of various displays, they find that their participants consistently prefer dynamic representations over static ones. However, they find that this does not necessarily match their task performance. Thus, they suggest that is important to assess the actual efficiency of displays and not just rely on user preference for implementing new systems, as the latter may not be the best indication of display effectiveness. Other findings similarly suggest that interaction does not necessarily improve a user's comprehension of a visualisation and task performance (Dou et al., 2010). A recent study by Severtson and Roth (2015) indicates that interactivity enhanced interpretation with maps showing uncertain information. However, the authors state that overall, it remains unclear in which contexts interactivity improves reasoning with uncertain information.

2.3 Previous Studies about Flow Visualisations

Postal OD data, which is the focus of this thesis, can essentially be seen as a network of postal flows. Thus, OD data analysis may benefit from network analysis techniques. In terms of network analysis, Sun et al. (2013) recognise matrix visualisations as well as graph layouts and clutter reduction methods as important categories in the visualisation of network data. Other frequently used visualisation types include flow maps (Yang et al., 2016) and OD maps (Slingsby et al., 2012). The following sections discuss the aforementioned visualisation types, along with other methods found in research about visualising spatio-temporal data.

2.3.1 Graph visualisation

As Herman, Melancon, and Marshall (2000) state, the basic data structure needed for graphs are a given set of nodes and a given set of edges, as they are often drawn as node-link diagrams. In the postal data set, nodes could represent the origins and destinations of postal flows, e.g. countries, mail processing centres or post offices, depending on the spatial scale. Edges represent the flows between the respective origins and destinations. Wong et al. (2006) define a semantic graph as a "network of heterogeneous nodes and links annotated with a domain ontology. In intelligence analysis, investigators use semantic graphs to organise concepts and relationships as graph nodes and links in hopes of discovering key trends, patterns, and insights." (Wong et al., 2006, p. 67).

A key issue in graph visualisation is the size of the graph, i.e. the size of the data to visualise. Herman, Melancon, and Marshall (2000) mention different problems in this regard, such as the limitations given by viewing platforms if the number of elements is too large, or what they call viewability or usability issues. They relate the problems in usability to the difficulties in differentiating between nodes and edges if there are too many of them. With a growing amount of data to display, graphs can become too complex and overburdening for the analyst's cognitive capacity. It thus becomes difficult for the user to conduct significant analysis (Wong et al., 2006).

Because of the issues described above, research often focuses on ways to solve the problems of visual clutter, e.g. by aggregation or clustering techniques, which is also one of the main topics of cartographic generalisation (e.g., Hegarty et al., 2009).

For example, Cui et al. (2008) investigate node-link diagrams for network visualisation. Their study focuses on the problem of visual cluttering in graph visualisations, as this is one of the main issues in the representation of relationships among large data (Herman, Melancon, and Marshall, 2000). They introduce a framework for geometry-based edge clustering to group edges into bundles and hence reduce visual cluttering caused by the crossing of the high number of edges (see figure 2.1).

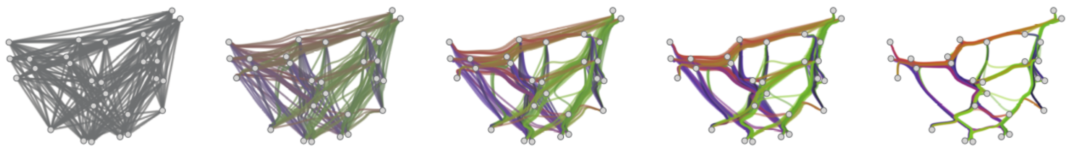


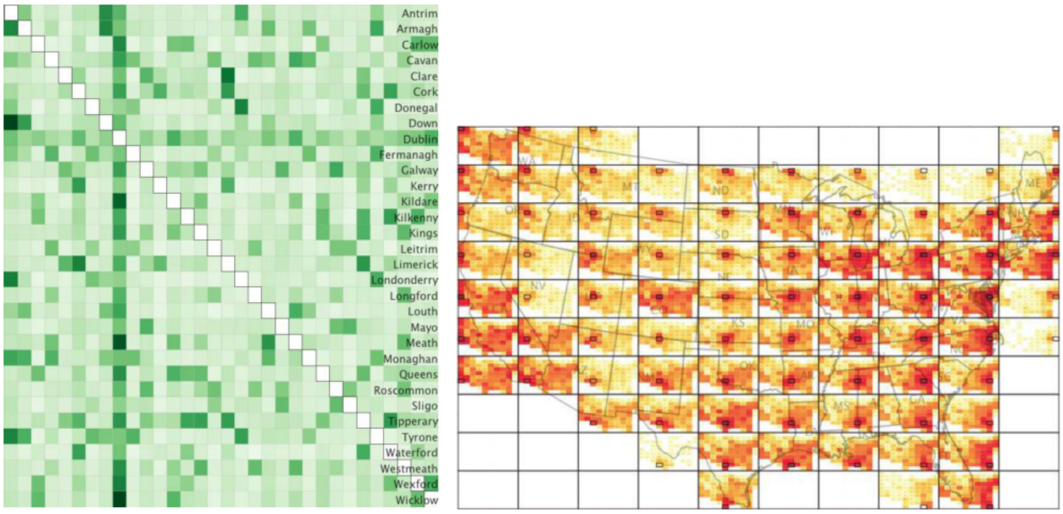
FIGURE 2.1: An animation sequence for edge clustering. *Source: Cui et al. (2008)*

Henry and Fekete (2006) use node-link diagrams and matrix visualisations to visualise social networks. They developed a network visualisation system called MatrixExplorer that consists of two interlinked and synchronised visualisation types, matrices and node-link diagrams. Filtering and clustering is possible, as well as reordering the matrix. The tool further allows to annotate and examine findings across different layouts.

2.3.2 OD matrix and OD map

As mentioned above, another frequently used way for representing OD data are so-called origin-destination matrices (ODM or OD matrices). The OD matrix is the most basic matrix approach, representing each origin as a row (R) and each destination as a column (C) in the matrix. Each cell (R,C) depicts the flow from origin to destination (see figure 2.2 (a)). As for graph visualisations, the abstract positions of the nodes and links in ODMs (rows and columns) do not represent the actual geographic locations (Yang et al., 2016). There are several attempts to enhance OD matrices, e.g. in developing reorderable matrices, in which rows and columns can be (re)ordered by location (e.g. east to west, or north to south), hierarchy, time, etc. (Guo et al., 2006).

Another approach is found in so-called OD *maps*, with the aim of preserving the actual spatial properties of the data (Yang et al., 2016; Slingsby et al., 2012). Slingsby



(a) OD matrix showing county to county flows in Ireland; rows = origin counties, columns = destination counties. *Source:* *Kelly et al. (2012)*

(b) US county to county migration vectors shown as an OD map. Each large grid cell represents origin location, within which the map of destination densities using the same grid is shown. *Source: Wood, Dykes, and Slingsby (2010)*

FIGURE 2.2: OD matrix (left) and OD map (right) of migration in Ireland and the USA, respectively.

et al. (2012) use OD maps to visualise historical migration within Ireland, Wood, Dykes, and Slingsby (2010) use this representation to visualise county to county migration flows for the USA (see figure 2.2 (b)). Slingsby et al. (2012) order the cells of the OD matrix according to their original two-dimensional geographic location, avoiding overlapping cells. Wood, Dykes, and Slingsby (2010) map OD vectors as cells of a regular grid instead of lines, and construct a gridded two-level spatial treemap in order to preserve the spatial layout of the origin and destination locations. Slingsby et al. (2012) find that OD maps are effective for studying spatial patterns of Irish internal migration, they also mention that the geography of Ireland is well suited for their approach. Wood, Dykes, and Slingsby (2010) see their approach of OD maps as a complementary way rather than as replacement for other forms of representations of large OD data, offering a new spatial view of such data.

2.3.3 Flow map

Flow maps have been known for a long time for representing flows of various kinds. One of the first flow maps and possibly the most famous one was created by Charles Joseph Minard in 1861, depicting the losses of Napoleon's army in the Russian campaign of 1812 by combining a visualisation of movement data and a time series graph (see figure 2.3) (Tufte, 1983; Andrienko, Andrienko, and Gatalsky, 2003).

Until today, flow maps remain popular for visualising various kinds of flows. As Yang et al. (2016) assert, flow maps are very intuitive, which may well be one of the reasons for their popularity. However, Yang et al. (2016) also note that flow maps are well suited for showing flows from one or very few sources, but less so

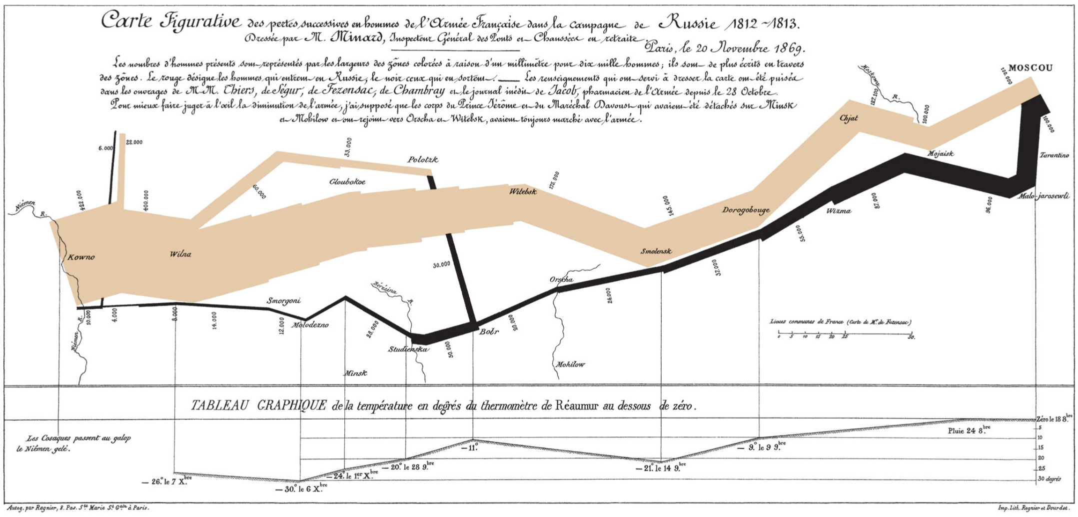


FIGURE 2.3: Charles Minard's map of the Russian campaign of 1812.
Source: http://www.en.wikipedia.org (Accessed: 12-09-2016)

for depicting a large number of sources (or targets). Again, as has been discussed for other visualisation types, cluttering and visual overload becomes a problem with large amounts of data.

More recent examples of flow maps include movement related visualisations by Wood, Slingsby, and Dykes (2011) as well as Boyandin et al. (2010). Wood, Slingsby, and Dykes (2011) propose three methods for visualising bicycle hire use and travel patterns in London, among them flow maps for providing an overview of flow structures. They state that flow mapping helped to provide structural overviews which reveal patterns otherwise not easily observable. Boyandin et al. (2010) present a tool called *JFlowMap* which provides several visualisation methods for creating and analysing flow maps. They present the tool by applying it to migration data (see figure 2.4).

FIGURE 2.4: Refugee Flows between the World's Countries in 1996, 2000, and 2008 *Source: Boyandin et al. (2010)*

2.3.4 Choropleth map

Choropleth maps are frequently used to visualise data that is aggregated to administrative areas such as countries, states, districts, etc. One application of this visualisation type is to show change in certain areas between different moments in time, e.g. by using shades of two colours to show the magnitude of increase and a decrease of an attribute value, respectively (Andrienko, Andrienko, and Gatalsky, 2003). According to Heer, Bostock, and Ogievetsky (2010), an often seen error is that raw data values are depicted instead of normalised values to create a density map. Further, they note that the size of the unit areas can affect the analyst's perception of the shaded values. This issue was also mentioned by Kelly et al. (2012), who showed historical internal migration in Ireland with choropleth maps, among other visualisation types (see figure 2.5). Kelly et al. (2012) replaced the choropleth map in a next step with a grid map to create an OD map (cf. 2.3.2) to ensures a more equal visual salience of each county. Sander et al. (2014) mention that choropleth maps are not ideal for analysing data sets with large numbers of spatial interactions, as large collections of such data are difficult to visualise with such basic visualisation methods.

Carr, White, and MacEachren (2005) propose a new approach for working with choropleth maps for a multivariate data analysis (conditioned choropleth maps or CCmaps). In this approach, the data is partitioned into subsets which allows to control the changes in the dependent variable that is related to two potential explanatory variables.

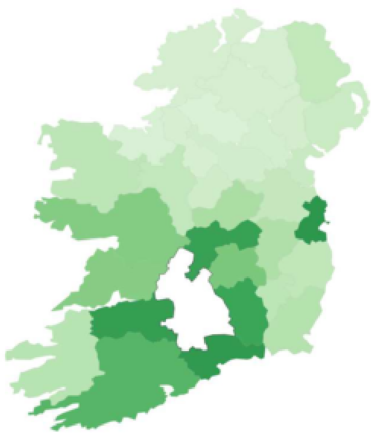


FIGURE 2.5: Choropleth map showing migration out of Tipperary county in Ireland (white) *Source: Kelly et al. (2012)*

2.3.5 Graduated symbol map

For showing absolute values, Heer, Bostock, and Ogievetsky (2010) suggest to use graduated symbol maps instead of choropleth maps, which avoids mistaking the underlying geographic area with the data values. Further, graduated symbol maps allow to depict more than one dimension by using several visual variables (shape, size, colour of the symbol). In their study about aggregation methods used for the visual exploration of movement data, Andrienko and Andrienko (2010) recommend graduated symbol maps for a situation-oriented view of movement (e.g. analysis of space use and accessibility). As relevant aggregation methods they note S (spatial aggregation) and $S \times T$ (position of elements at different discrete events or snapshots). Rae (2009) used graduated symbol maps to represent migration flows into the city of Manchester (see figure 2.6). He converted the inflow lines to dots and showed the changes in amounts by using the visual variables 'size' and 'colour'.

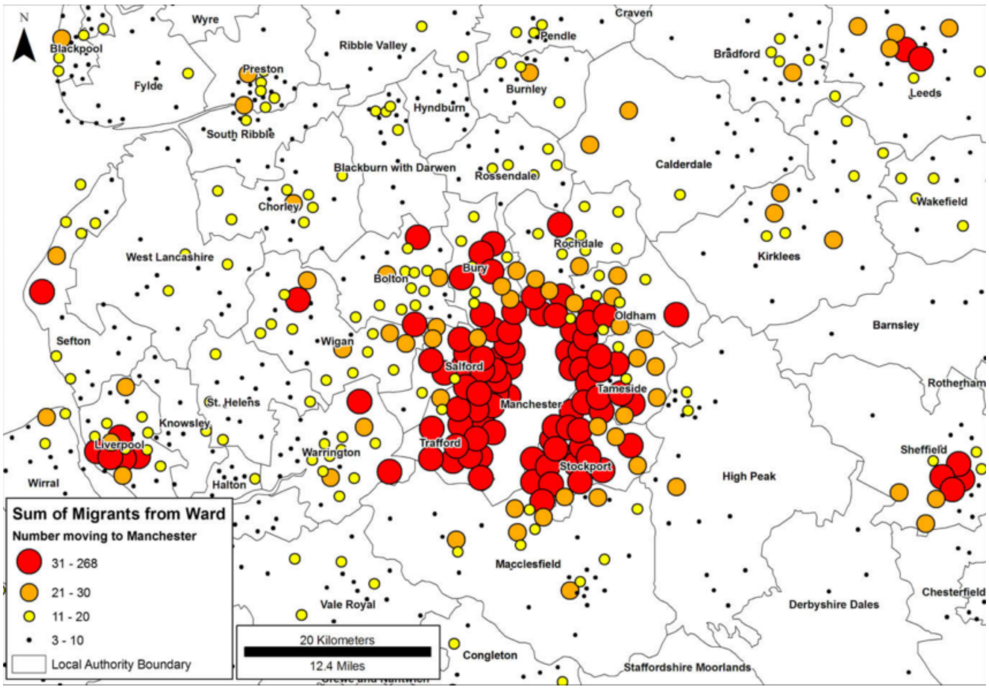


FIGURE 2.6: Graduated symbol map showing migration inflows to Manchester *Source: Rae (2009)*

2.3.6 Time series

The analysis of the temporal dimension in data is important in many research areas, but bears challenges for data visualisation and analysis (Keim et al., 2010). Visual representations of temporal data can be divided into categories by the time characteristics they are showing: 1) linear vs. cyclic time, 2) time points vs. time intervals, 3) order time vs. branching time vs. time with multiple perspectives. For example, line graphs are practical for showing general trends or outliers, whereas spiral visualisations or ringmaps can show cyclical patterns in the data (Keim et al., 2010; Zhao, Forer, and Harvey, 2008).

Andrienko and Andrienko (2005) mention several aspects regarding the visualisation of time series data to improve the effectiveness of its analysis, for example the interactive access to the actual values by hovering over a part of the time line or the possibility to show multiple lines simultaneously for comparing them. In their study, they evaluate various methods for exploring time series data that is spatially distributed. They propose three different methods for a) providing an overview of the value changes across the whole region, b) detecting spatial patterns of identical temporal behaviour and c) detecting spatial patterns with similar changes over time. For b) and c) they used dynamically linked displays for the visualisation.

Hochheiser and Shneiderman (2004) suggest a different approach for enhancing the usability of time series visualisations. They introduce so-called 'timeboxes' for querying the data and selecting a specific part of it that is of interest (see figure 2.7). The box helps to select data by specifying its extent on the time axis (time period of interest) as well as the y axis (value range of interest in the given time period). The timebox thus acts as a filter for choosing values in a given range during the time interval specified by the width of the box.

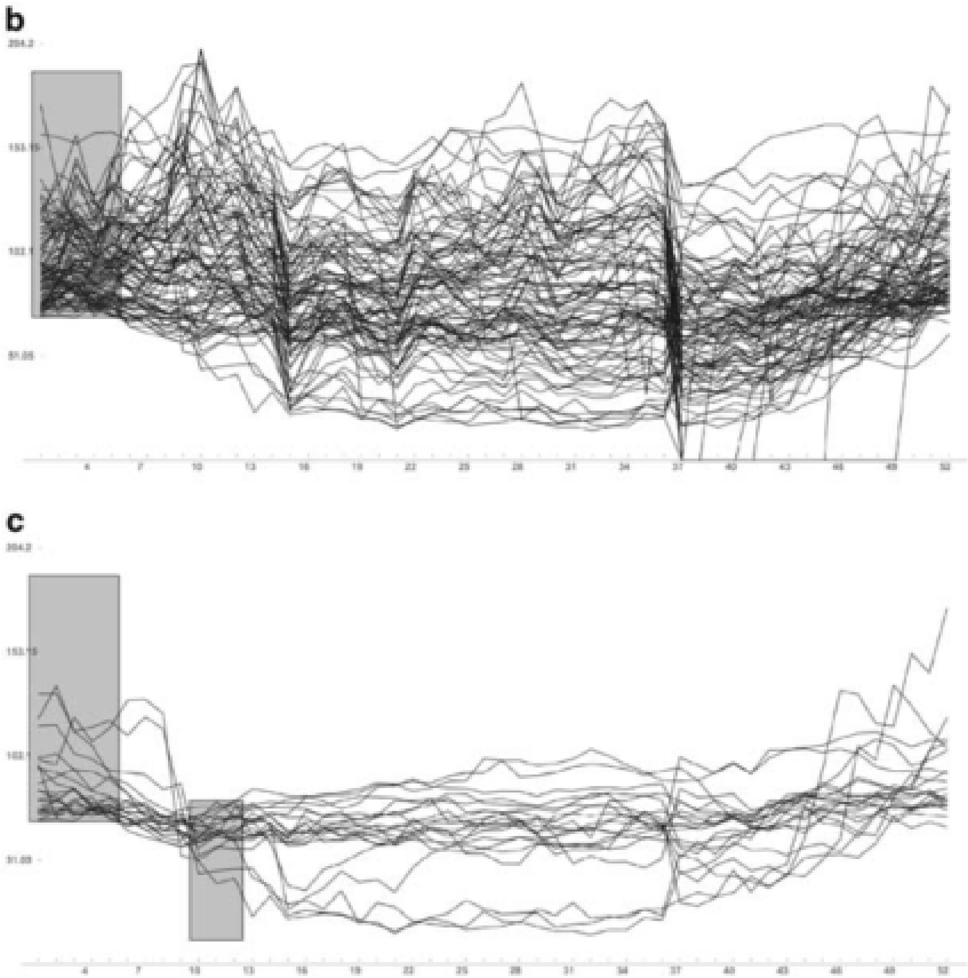


FIGURE 2.7: Time series graph showing 52 weekly stock prices for 1430 stocks (x-axis = weeks 1–52, y-axis = stock prices). Top: single timebox query for items that traded between 70 and 190 during weeks 1-5. Bottom: query containing two timeboxes, refining the query in (b). *Source: Hochheiser and Shneiderman (2004)*

2.3.7 Chord diagram

Chord diagrams have been used in recent studies to visualise flows, for instance of people[1] (Sander, Abel, and Bauer, 2014), traffic (Li et al., 2016), or social insurance (Heil and Walch, 2015). Both Sander, Abel, and Bauer (2014) and Heil and Walch (2015) state that their goal was mostly to visualise complex data flows for a wider audience, including non-technical or non-scientific users. While Sander, Abel, and Bauer (2014) find that circular migration plots (see figure 2.8)improve not only the ability to communicate, but also to analyse and understand the patterns in the data, Heil and Walch (2015) state that a large amount of time was needed for data transformation in order to produce chord diagrams.

FIGURE 2.8: Interactive circular migration plot of migration flows between regions in 2005-10, created with D3. *Source: Sander et al. (2014)*

2.3.8 Treemap

A treemap is a space filling enclosure diagram (see figure 2.9). As the name suggests, it uses containment instead of adjacency to visualise the hierarchy in the data (Heer, Bostock, and Ogievetsky, 2010). As the authors state, a treemap subdivides the total area into rectangles in a recursive way. This allows the analyst to quickly retrieve information about the data that is visualised.

2.3.9 Further visualisations

Flow visualisations can also reach public space in the form of installations[2], as shown by Nagel and Pietsch (2015). They want to engage the public with urban mobility as they show visualisations of bike-sharing journeys in three cities (Berlin, London, New York City) side by side on large high-resolution screens.

[1] http://www.global-migration.info
[2] https://uclab.fh-potsdam.de

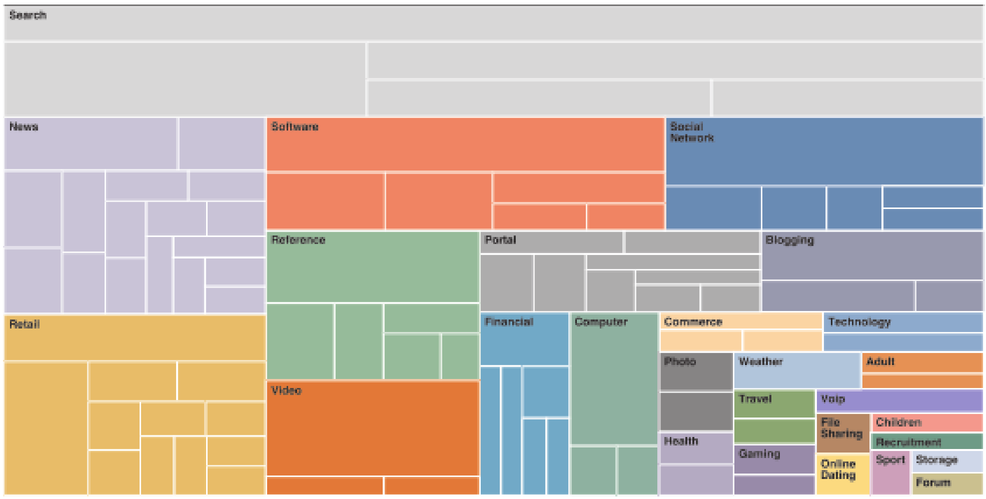


FIGURE 2.9: Treemap showing the number of unique users of the top 100 websites within January, 2010. *Source: Lee et al. (2016)*

Brunsdon, Corcoran, and Higgs (2007) evaluated three different visualisation techniques (map animation, isosurface, and comap[3]) on a theoretical basis regarding their suitability for the exploration of spatio-temporal patterns of crime data. They conclude that, in the case of crime patterns as well as in other areas of pattern analysis, more studies are needed regarding the suitability and effectiveness of visualisation techniques for knowledge discovery. In particular, they state that it is important to conduct user studies with experts in the respective field to test the effectiveness of the applied techniques in the daily operational work.

Further, Brunsdon, Corcoran, and Higgs (2007) as well as Keim et al. (2010) stress the importance of taking in account not only the spatial, but also the temporal dimension of the data.

2.3.10 Combining and linking visualisations

Combining and linking different visualisations of the data being analysed can help analysts to explore the data, as they view it through different representations, each of which can bring across only partial information on its own (see e.g. figure 2.10) (Roberts, 2007; Keim et al., 2010; Andrienko and Andrienko, 2005; Brunsdon, Corcoran, and Higgs, 2007; Guo et al., 2006). Combining not only different visualisations of the same data, but representing different types of data on the same screen is a further possibility. As Keim et al. (2010) state, new developments, especially in web based technologies, offer new possibilities of producing and using data visualisations. Visualisations can for example be linked to various sensors, therefore enabling the integration of data from several sources, e.g. weather reports or news feeds. Most often, such web based visualisation platforms are interactive and allow

[3]A geographical version of the conditional plot (coplot) for investigating the relationship between two variables conditioned on a third variable, where maps are used in the panels instead of scatter plots (Brunsdon, 2001).

the user to explore the data. Gapminder[4] is an example of such an approach, presenting world census data in a combined view.

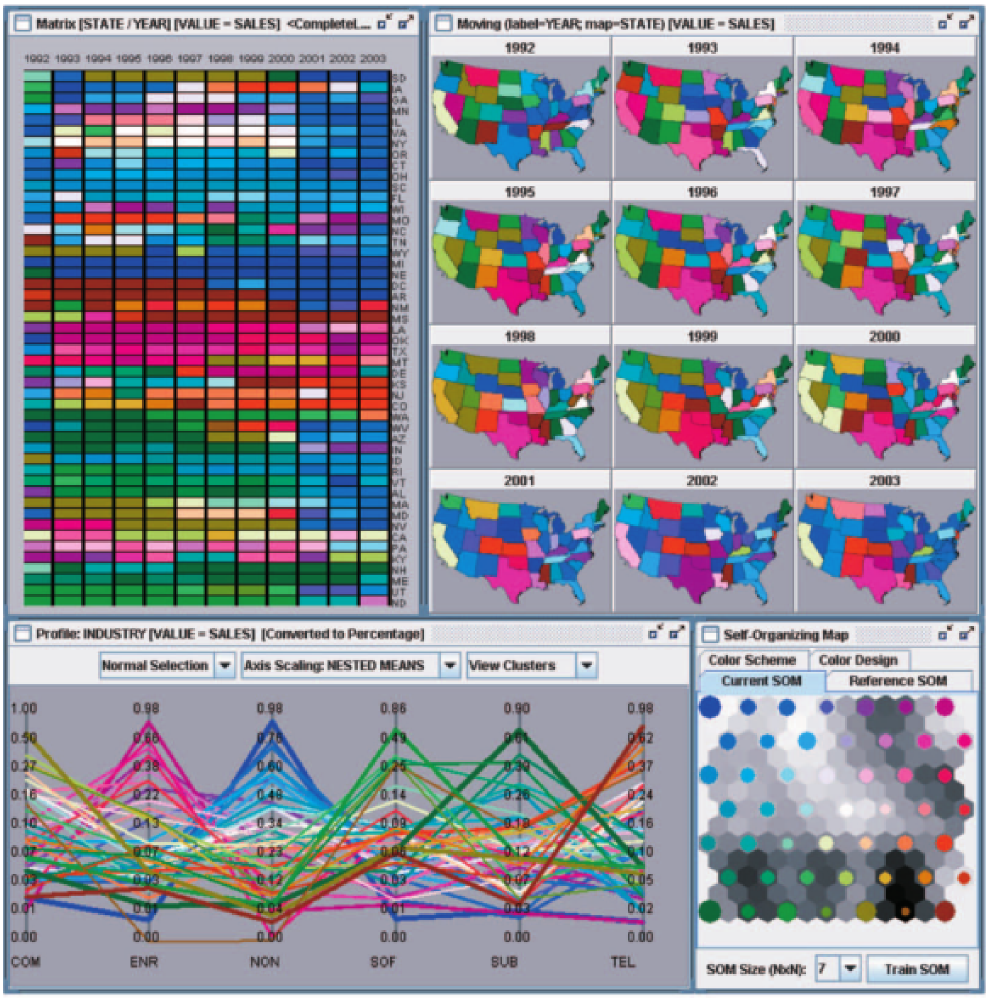


FIGURE 2.10: Linked plots showing changing characteristics of US industries; Top-left: reorderable matrix, top-right: map matrix, bottom-left: multivariate "legend"-PCP (Parallel Coordinate Plot), bottom-right: SOM (Self-Organising Map). *Source: Guo et al. (2006)*

Most of the aforementioned studies followed a 'top-down', i.e. a theoretical approach for assessing the efficacy of the visualisations they were analysing, thus the visualisations were not tested with actual users, for example experts in the respective fields. However, it is necessary and important to test the applied visualisation methods with user studies in order to obtain better knowledge about the requirements of the analysts and about the visualisations' effectiveness in the day-to-day operational work (cf. Brunsdon, Corcoran, and Higgs, 2007).

The visualisation methods that were investigated in the previously mentioned studies were further mostly applied to a specific data set or a particular location,

[4] http://www.gapminder.org

hence their asserted advantages or limitations might at least partly be context dependent. It is thus important to test those visualisation methods for other types of flow data, as results from a study focusing on, for example, migration might not necessarily hold true for other types of OD data such as postal flows. The evaluation of visualisation methods with another type of OD data might thus provide additional insights for flow data with a similar structure.

2.4 Previous Studies with Postal Data

Only few studies have been conducted with postal data, especially in the field of data visualisation and visual exploration of the data. In their analysis, which was conducted with data from Serbia, Marković, Grgurović, and Štrbac (2011) focused on the postal service market in urban areas. They highlight the importance of spatial data and the benefits of applying geographic information systems (GIS) for obtaining valuable information for the segmentation of the postal service market with its largely heterogenous client base. Especially for those parts of the postal service market that are not reserved for public postal operators, it is crucial to optimise services, e.g. delivery time, in order to persist in the highly competitive market. Marković, Grgurović, and Štrbac (2011) state that spatial data and its analysis is very beneficial for such optimisations.

Tabak and Kljak (2009) developed a traffic matrix model for postal flows in the public postal network of Croatia. Their aim is to enable the identification of the main geographical distribution routes according to the location of acceptance points (input) and delivery points (consumer demands for outputs). This will help to distribute the total costs of the postal network through all sections of the value chain and thus recalculate tariffs that are put in place for accessing the public postal network. In contrast to Marković, Grgurović, and Štrbac (2011), Tabak and Kljak (2009) do not apply any spatial analysis supported by a GIS, as their matrix model is created with standard spreadsheet software.

Ben-Ayed and Hamzaoui (2012) followed a mathematical approach to a 'multi-objective multiproduct parcel distribution timetabling problem', i.e. the creation of effective timetables for parcel distribution. They focused on time and cost reduction based on a real-world case study, but did not explicitly include the factor space in their analysis. However, they stressed that the collaboration with the practitioners is crucial for defining the relevant requirements and objectives in order to develop a successful model.

As mentioned in the introduction, the postal data collected by the UPU has only been the focus of one study so far. Hristova et al. (2016) compared the international postal network to other global networks such as the trade network, the migration network, the flight network, the digital network and social media networks. They conclude that it is possible to predict several socioeconomic indicators by measuring the position of each country in those networks. By analysing multiplexity (the multi-channel exchange of goods or information) across the above mentioned networks in order to construct a degree measure, Hristova et al. (2016) were able to show that the strength of relationships between countries is a strong indicator of wellbeing on an international level. Further, they find that countries that belong to the same community over multiple networks are more alike regarding their socioeconomic profiles. Hence, according to the authors, the postal data has the potential to provide interesting insights on economic trends, migration, etc., not least because it is one of the oldest global networks which is present even in the most remote areas of the world.

By analysing the postal data at country level, Hristova et al. (2016) could extract interesting information about the intensity of relationship between countries. It needs to be said that the data they used is not exactly the same as the data that has been provided for this thesis. Hristova et al. (2016) used data on dispatch level (aggregated data about total postal shipments), whereas this study uses data about individual tracked items. The data about tracked items provides a higher resolution than country level (cf. section 3.1.1) and thus allows studying patterns of postal flows on different spatial and temporal scales, which is one of the aims of this study. As Nieto Corredera (2015) states in the report on postal e-services development, the shift towards online shopping behaviour creates a growing demand for optimised logistics and personalised delivery of parcels. He suggests that the big data approach that is followed by collecting tracking information has several advantages: The rich data sets that are produced by using tracking technologies help service providers, marketers and merchants to understand consumer trends. They can thus improve and optimise their business models and lower operational costs across the value chain.

This thesis focuses not only on the analysis of patterns, but also on the detection of anomalies in the postal data. The Memorandum by the Chairman of the Ad Hoc Group on Disaster Risk Management of the UPU states that, according to international observers, the scale of, and the vulnerability and exposure to hazards is expected to increase in future. According to the Memorandum, the postal sector (among many others) has been affected severely by natural disasters, and postal service has been interrupted. However, it is also stated that the Post can serve as a key player in disaster response, for example as distribution point for emergency supplies or coordinator of emergency aid operations (Widloecher, 2013; Ad hoc group on Disaster Risk Management, 2015). Analysing patterns and especially disruptions in postal shipment data could help the postal experts understand and, if combining the postal data with other data sources, potentially explain how the postal network is affected by certain types of disrupting events.

Chapter 3

Data and Methods

The first section of this chapter describes the postal data as well as additional data sources used for implementing the visualisations. The second section describes the data processing, mainly of the postal data. The third section presents the case studies. The fourth section illustrates the setup of the requirements analysis, followed by the fifth section that describes the Emergency Information System (EmIS) data. The sixth section presents the reviewed and implemented visualisation methods and the seventh section provides an outline of the usability study.

3.1 Postal Data

The postal data used in this thesis is provided by the UPU. It contains information about tracked items from the postal traffic within the UPU International Postal Network (IPN). Data about domestic shipments as well as non-tracked items are not included. The data which is captured on item level does not include any information about transportation mode. It is further important to mention that information about shipments by private postal and express operators such as DHL, DPD, UPS, FedEx, etc. are mostly not included in this data. Market shares of private logistics providers differ between countries. While some of them work independently through their own integrated network, others collaborate with historical postal operators and exchange data with them. In addition, the public postal operators of a few industrialised countries may sometimes avoid sharing data through UPU systems, given the sensitivity of traffic information with a number of their key partners abroad. This results in issues regarding data completeness, as can be observed particularly well for the US inbound data that was obtained for this study. For example, it is known that there is a substantial postal exchange between Germany and the US, however this flow appears as almost non-existent in the data set that was provided.

The item level data is mostly focused on Business to Customer (B2C) shipments, which means that the majority of the data that is captured is related to e-commerce or private shipments.

TABLE 3.1: Structure of postal data (EDI message chain from sender to recipient)

Event A (Sender PC) →	Event A (Orig. Office) →	Event B (Disp. Office) →	Event C (Disp. Office) →
Sender Post Code: often missing values or code unclear	Sender Office: often missing values or code unclear	Dispatch Office (arriving): IMPC codes often missing values	Dispatch Office (leaving): IMPC codes relatively good quality
Event D (Inb. Office) →	**Event F (Inb. Office) →**	**Event G (Deliv. Office) →**	**Event I (Deliv. Office)**
Inbound Office (arriving): IMPC codes relatively good quality	Inbound Office (leaving): often missing values	Delivery Office: often missing values	Receiver Post Code: often missing values or code unclear

3.1.1 EDI message chain

In order to monitor and analyse the postal traffic within the IPN, EDI (Electronic Data Interchange) messages are collected by postal operators worldwide and passed on to UPU. EDI messages are generated by barcode scanning and other computerised entries at various key points in the shipping process from sender to receiver, e.g. postal sorting centres and post offices. This data is collected since 2010 until today, amounting to nearly 14 million records related to dispatch level information, and more than 5 billion records available for individually tracked items sent between countries (Hristova et al., 2016). Each item is represented as one entry in the data set. The collected data contains information about the origin and destination country, possible transit stops, the item class (U = letters and parcels < 2kg / C = parcels 2-30kg / E = express mail), as well as eight *Events* (date/time and location) as shown in table 3.1. In contrast to other OD data, in some cases this data thus contains 'multihops' because of transit stops. This structural characteristic places the data somewhere in between the traditional OD data and trajectories. In this study, however, the focus lies on the flows between origins and destinations and the in-between stops are not considered at this point.

The data further contains information about planned and actual departure and arrival time of the items at specific *Events*. The most complete information is available for the items of class C, namely parcels, as these are the items that are scanned individually. Letters and lighter parcels do not necessarily have to be tracked, which results in incomplete data. As this study focuses on overall patterns and anomalies, the data is not split into the different item classes. This could, however, be an interesting approach for further investigation.

Generally, the data quality is best for *Event C* and *Event D* (the dispatch and inbound offices of the sending and receiving country, respectively). Data entries for *Event C* and *Event D* refer to IMPCs (International Mail Processing Centres, see also section 3.1.2). For other *Events*, it is not always clear what the data entries in the respective columns refer to. According to UPU experts, they may refer to post codes, but often it is not clear what they represent. The data entries could also be a reference to local post offices or something else, given that the data input depends on the choices operated by each country.

As mentioned above, there is a time stamp for each *Event* of the tracked items. These time stamps are recorded in local time, thus for analysis with a focus on the temporal dimension this needs to be kept in mind. It is suggested to convert the time stamps into UTC time prior to analysis with a focus on the duration of shipments.

3.1.2 IMPCs

Each IMPC is has a unique code which consists of 6 letters, for example USMIAA: the first two stand for the respective country (US), the third through fifth specify the location, e.g. an airport (MIA = Miami), and the sixth letter provides additional information, e.g. which company is responsible for the handling of the items at the given location. The sixth letter was deleted, as this information is not of any value for the analysis in this study.

3.2 Data Processing

This section mostly focuses on the postal data, as the data generated from interviews and EmIS reports is rather qualitative than quantitative and does not require the same amount of processing.

3.2.1 Data selection

The thesis entails two case studies, US and Chile (CL). For each case study, a specific set of data was selected to conduct the analysis (for further details on the selection criteria, see section 3.3). For the US, the case study includes calendar weeks (CW) three to eight of the years 2012, 2013, 2014 and 2015 and for Chile, CWs 31 to 36 of the years 2012, 2013 and 2014, as the data for 2015 was not yet available for these CWs.

For the purpose of this thesis, a CW spans from Sunday to Saturday. This is usual for the US, where the largest data set analysed in this study comes from.

For the visualisation of outbound data, *Event A* (when parcels are sent) was the relevant date, for visualising inbound data, *Event D* (when parcels enter the country) was relevant.

3.2.2 Data cleaning

The data contains missing, as well as incorrect values. Further, some data fields do not contain homogeneous values. For example, the US ZIP codes are mixed between 5- and 9-digit codes, some of which do not correspond to any of the official ZIP Code Tabulation Areas (ZCTA) from the US Census.

In a first step, the data was cleaned and processed in a way that it can be used for applying the visualisations. The following data were excluded: shipments where the origin and destination countries are the same, shipments where the date of *Event C* was before *Event B*, shipments where the code for *Event C* (outbound data) or *Event D* (inbound data) did not contain the respective country code (USXXX, CLXXX). The US data used in this study only considers direct shipments (data without transit stops), whereas the CL data includes all shipments because of the small total number of shipments.

3.2.3 Adding spatial information

A shapefile with spatial information about the IMPC locations was obtained from UPU. The file contained a total of 86'062 lines, each consisting of the columns FID, Shape, code, country, name, lat, lon, function, status, and zone.

28'224 entries were missing spatial information, i.e. their lat/lon columns consisted of a value of 0. Among those entries with missing coordinates were a considerable amount of important hubs such as Montreal in Canada, or Rio de Janeiro in Brazil.

To enable an automated import of coordinates for the important IMPCs (i.e. hubs close to airports), the list was merged with data obtained from OpenFlights.org[1]. Among other information, this data contains the 2-letter ISO country code as well as the 3-letter IATA/FAA code specifying the respective airports. Airports are assumed to represent the main hubs for international postal shipment. The two above mentioned codes were used to merge the airport data with the original IMPC data, which resulted in a reduction of about 2'500 missing coordinates. The remaining entries with missing coordinates were disregarded for this analysis, as the vast majority of them never appear in the EDI messages. However, this data incompleteness should be taken in account for future studies.

3.2.4 Spatial aggregation

In order to reduce the data size and facilitate visual exploration, the data was aggregated. This can be done spatially, e.g. by administrative units (countries, sortation areas, even postal code areas) or by IMPCs, if the corresponding data is available. For the visualisations of this study, the data was aggregated to two different levels of detail (LOD), country level (LOD 1) and IMPC level (LOD 2).

The following two LODs were considered in the review, but not included in the usability study:

- Regional level: For Chile, the data could be aggregated to a regional level, which could potentially provide further insights. However, spatial data for these exact postal regions is not available, therefore, only visualisations that do not include any geographic information would be possible at this stage.

[1] http://openflights.org/data.html

- ZIP Code level: For the US data, an aggregation on ZIP code level is possible. The postal data includes information about 5- and 7-digit codes. Spatial data however was only available for 5-digit ZIP codes. Some tests were run with this limited data (see figure 3.6), but it was decided not to include it in the final analysis, as the incomplete data did not yield acceptable results.

As each of those two LODs would only be available for either CL or the US, no comparability would be possible. Additionally, the usability studies lasted about 3.5 hours for each participant. Asking them to evaluate more visualisations would only result in limited additional benefit.

3.2.5 Temporal aggregation

Temporal aggregation is another possibility, e.g. aggregating the the data to different temporal segments (snapshots). In this thesis, following advice from postal experts, the data for most of the visualisations was aggregated to calendar weeks. For the time series visualisation, the data was aggregated to years, thus visualising the whole period of six weeks at once. Other possibilities include the aggregation to daily snapshots, for example if one is interested in looking more closely at patterns within a week, or compare certain weekdays across the year, or weekends vs. weekdays. Even hourly segments could provide interesting insights for analysis with a strong focus on time.

Information about the planned arrival/departing time is available in the data. As anomalies may also be detected there, this information was taken into account as well. Time difference was calculated in days or hours for inbound or outbound data, respectively. The distribution was displayed in the form of histograms, however, this analysis was not included in the usability study.

Time zone conversion

As was mentioned before, the time stamps of the individual *Events* are stored in the respective local time. For analyses with a strong focus on time, especially regarding the duration of the parcel journeys, it is suggested that the time stamps are converted to UTC time. In order to do so, information about the respective time zone of post offices and IMPCs needs to be added to the data. If spatial information for the post offices and IMPCs is available, this can be done relatively easy in a GIS by performing a 'spatial join' between the layer with the postal locations and a layer containing the information about UTC time zones, e.g. from Natural Earth[2]. In a next step, the time stamps of the *Events* need to be converted according to their respective time zone. In R, this was found to be a rather complicated process, however, in Python this could be done in a more straightforward way. For this thesis, it was decided to use local time stamps because the focus did not lie on the duration of parcel shipment and because of technical reasons.

[2] http://www.naturalearthdata.com

3.2.6 Normalisation

Instead of looking at absolute values, the data can be normalised. Normalisation reduces emphasis on flows with high absolute values, e.g. shipments from China to the US. For LOD 1, the values were normalised to the population size of either the sending or receiving countries. Population data was obtained from the World Bank[3]. In the case of the IMPCs, normalisation to population size is not possible, as the IMPCs do not correspond to any administrative units. Therefore, absolute values were used for LOD 2. Further, the data could be normalised with baseline values, e.g. weekly averages. It is recommended to include this in further investigation.

3.3 Case Studies and Tasks

As mentioned above, this thesis focuses on two different countries (USA, Chile) as case studies. The two countries are defined by different characteristics relevant to the analysis focusing on the postal network. The USA are an industrialised country, whereas Chile is less industrialised. The USA are characterised by one of the highest global connectivity levels regarding postal operations, whereas Chile is also connected to many countries in the world, but its connectivity level is still much lower than the one of the USA.[4] Further, the overall volumes sent to and from the US are a lot higher than the ones sent to and from Chile (see table in Annex **??**). This has consequences for analysing the data: Large flows of data allow a better reasoning about patterns and anomalies, as the flows in itself are usually more stable, whereas with small overall flows, sudden changes in volumes may also be related to the natural volatility in the data rather than to actual disruptions. The data quality was also a factor to take in account for the selection of the case studies. For both countries, the data quality of the EDI messages is rather good, according to postal experts. Also, spatial data is available for both countries and there is some information about specific disruptions in the EmIS database, which helped to select the time period.

3.3.1 Patterns and anomalies

Patterns

The patterns of interest in this study are patterns that are meaningful in the context of the postal network and thus to postal experts. For example, strong and stable connections are expected to be seen between major postal partners and between countries with historical and/or cultural ties. Trade is a major driver of postal exchange, with an acceleration of traffic with the rise of e-commerce (Hristova et al., 2016).

[3] http://data.worldbank.org

[4] For the investigated data sets (six CWs of four and three years, respectively), the US has outbound connections with 185 countries (weekly average: 131, including the very low year of 2013) and inbound connections with 57 countries (weekly average: 15), whereas CL has outbound connections with 57 (weekly average: 34) and inbound connections with 78 countries (weekly average: 43).

Thus, major connections between producing and consuming countries of products being distributed through e-commerce channels should be visible. Temporal patterns such as weekly patterns should also be visible, in the present case studies mainly in the outbound data. For detecting yearly patterns, a larger time spans than the period of six weeks used in this study needs to be analysed.

Anomalies

Anomalies are expected to be seen in data where regular shipments are either affected by external factors such as natural disasters, weather, etc. or social/political issues such as strikes. Further, anomalies are possible because of technical issues or reporting problems in the postal network.

3.4 Requirements Analysis

In order to learn about the main needs of professionals working with postal data, a requirements analysis was conducted. The analysis consisted of two parts: 1) an online survey, which was distributed to people working at UPU and 2) expert interviews that were conducted with the employees of UPU who are working with the postal traffic data on a regular basis. The online questionnaire serves as a way to get insights from various perspectives and can thus provide inputs from a broader range of people within UPU, whereas the expert interviews are conducted to obtain more detailed information from the people most familiar with the data itself. The next two sections explain the two parts of the requirements analysis in more detail.

3.4.1 Online questionnaire

The online survey was created with the tool SurveyMonkey[5]. It allows to combine and edit several question types as well as the integration of figures in an interactive way. The questionnaire was built based on recommendations made by Courage and Baxter (2005) (the complete set of questions can be found in Appendix A.2). It consisted of a total of 32 questions, split up into two parts:

1. General information: The first eight questions aimed at collecting general information about the participant such as age, gender, professional background, and experience in the field of information visualisation.

2. Visualisation types: The remaining 24 questions focused on the actual topic of visual analytics (visualisation types for visualizing large OD data) in order to obtain information about the current usage of such methods, the participants' familiarity with them, etc. As all of the participants came from within UPU, questions about their use of postal data, the data's purpose for their specific group/division, as well

[5] https://www.surveymonkey.com/

TABLE 3.2: Overview of interviews for the requirements analysis

Interviewee	Status	Format	Length	Recording	Transcript
Lead Economist	Conducted in person	Semi-structured	1 hr 07 mins	Audio recording & concurrent notes	Available
Data Scientist	Conducted in person	Semi-structured	1 hr 15 mins	Audio recording & concurrent notes	Available
Expert for Environment and Sustainable Development	Conducted in person	Semi-structured	45 mins	Audio recording & concurrent notes	Available
GIS Analyst	Conducted in person	Semi-structured	55 mins	Audio recording & concurrent notes	Available

as their interest regarding time and space were included. Another section contained questions about patterns and anomalies in the data and the participant's awareness of either of them.

The online survey was distributed by e-mail to a total of 20 people working at UPU. The e-mail consisted of a very brief introduction to the survey and its purpose and a direct link to the questionnaire. Of the 20 addressees, 13 completed the survey.

3.4.2 Expert interviews

As outlined above, semi-strucured expert interviews were conducted in addition to the online questionnaire. Four people were interviewed; three of them work closely with the postal data, one works in the Sustainable Development Group within UPU but takes a high interest in the outcomes of the postal data analysis.

All of the interviews were audio recorded and notes were taken during the interviews, as is recommended by Höld (2009). In one of the interviews, the recording device stopped after 40 minutes, but since notes were taken, the key points of the answers could still be collected in a satisfactory quality for the requirements analysis. Table 3.2 provides an overview of the interviews according to suggestions made by Bleich and Pekkanen (2013) regarding the reporting of interviews. The interview guide can be found in Appendix A.1.

3.5 EmIS Reports

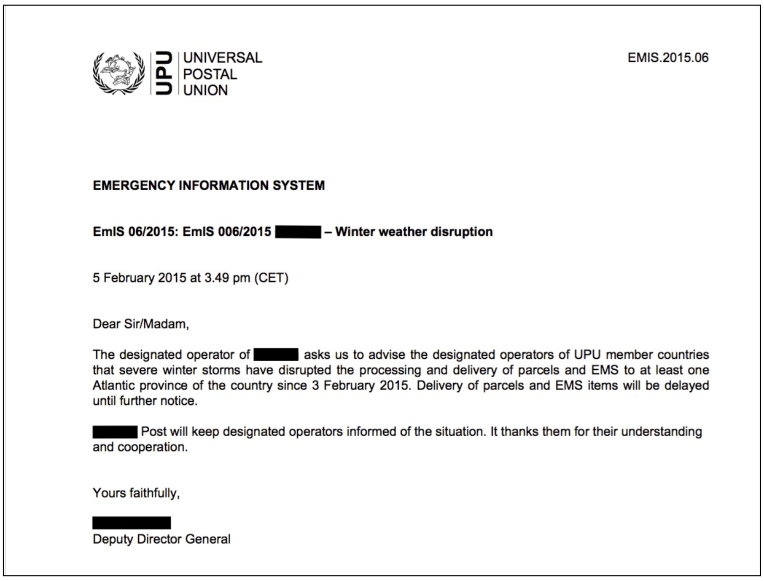
The UPU collects information about different kinds of disruptions in its member countries via its Emergency Information System (EmIS) since 2003. Messages that are contributed by affected countries are communicated through this system, which is accessible to all of the UPU member countries. According to UPU employees, the participation in this messaging system is not compulsory for the members. This means that the collection of messages is most likely incomplete. However, it reflects what the countries regard as important and/or severe enough for reporting to the

TABLE 3.3: Classification of EmIS reports

Category	Abbreviation	Subcategory	Category	Abbreviation	Subcategory
Political	POL	Political protest	Social	SOC	Strike
		Political conflict			National Holiday
		Elections			
		Terrorism	Health	HEA	Virus
Natural	NAT	Flooding	Technical	TEC	Server problems
		Hurricane			EDI messaging
		Cyclone			Internet problems
		Snow storm			Flight cancellation
		Fire			Power cut
		Volcanic eruption			
		Earthquake			

UPU and the other member states, as some of the disruptions may have strong impacts on the postal network and therefore on other countries. There were 101 entries in 2013, 99 in 2014, and 121 in 2015.

The messages are stored in the format of a formal letter (see figure 3.1). To get a better overview of all the messages, an excel sheet was created with the contents including the date of the EmIS message, the start and end date of the disruption, the country, the post codes affected (if mentioned), and key comments of the message (see figure in Annex B.1). The list also includes the information whether it was the first message of a specific disruption, or a status update or closing message. All of the items in the table were then categorised based on a two-level classification scheme (see table 3.3).


UPU | UNIVERSAL POSTAL UNION

EMIS.2015.06

EMERGENCY INFORMATION SYSTEM

EmIS 06/2015: EmIS 006/2015 [redacted] – Winter weather disruption

5 February 2015 at 3.49 pm (CET)

Dear Sir/Madam,

The designated operator of [redacted] asks us to advise the designated operators of UPU member countries that severe winter storms have disrupted the processing and delivery of parcels and EMS to at least one Atlantic province of the country since 3 February 2015. Delivery of parcels and EMS items will be delayed until further notice.

[redacted] Post will keep designated operators informed of the situation. It thanks them for their understanding and cooperation.

Yours faithfully,

[redacted]

Deputy Director General

FIGURE 3.1: An example of an EmIS message showing the original format of the collection.

3.6 Data Visualisation Tools

Three tools were used for creating the visualisations that were included in the usability study: Tableau[6] a (business) data visualisation application, R Studio[7] (R), and ArcGIS Online[8]. The visualisations created with R are static, the ones implemented with Tableau and ArcGIS Online provide interactive features.

Tableau was tested for several other visualisation types, but eventually it was only used for implementing the time series visualisation. R and Gephi[9] were tested for implementing graph visualisations, however those visualisations were eventually not included in the usability study. R, more specifically the R package 'circlize' was used for a chord diagram implementation, however this was also not included in the usability study for evaluation.

3.7 Data Visualisation and Analysis Methods

The first part of this section provides an overview of the implemented visualisation types and the tools used for the implementation. The second part briefly explains all of the reviewed visualisation types and provides reasons for not implementing all of them in the study. An overview of all reviewed visualisations is presented in table 3.4. The ones marked with x in the column 'Selected' were implemented and evaluated in the usability study. An overview of the implemented methods for both data sets (US and CL) is given in tables 3.5 and 3.6.

All of the visual representations use a range of blue values in order to homogenise the visualisations for the study as much as possible. However, since the data was visualised with different tools and methods, the blue values vary between the visualisation types.

As stated above, a selection of several visualisation methods was made based on literature research and the results from the requirements analysis, as well as technical possibilities.

The data sets were aggregated to CWs and all the visualisations (except for the time series) were created with these weekly data sets (cf. Andrienko and Andrienko, 2010). This allowed the users to compare the data not only between the years, but also between the weeks within the the whole period of one particular year. The visualisations that were tested in the usability study were informationally equivalent in the sense that all representations of the same LOD and either in- or outbound direction used the same underlying data. However, by the definition given by Larkin and Simon (1987), 'informationally equivalent' means that all of the information in one representation is also inferable from the other representation, and vice versa. In

[6]http://www.tableau.com
[7]https://www.rstudio.com
[8]http://www.arcgis.com
[9]https://gephi.org

TABLE 3.4: List of reviewed visualisation types

Type	Static	Interactive	Applicable (spatial) scale	Selected	Tool
OD Matrix	x	–	Country, Region, IMPC	x	R
Flow Map	x	–	Country, IMPC	x	R
Choropleth Map	–	x	Country, ZIP Code	x	ArcGIS Online
Graduated Symbol Map	–	x	Country, IMPC	x	ArcGIS Online
Time Series	–	x	Country, IMPC	x	Tableau
Graph	x	x	Country, Region, IMPC	–	R, Gephi
Chord Diagram	x	–	Country, Region, IMPC	–	R
Treemap	–	x	Country	–	Tableau
Bubble Chart	–	x	Country	–	Tableau
Histogram	x	–	Time difference	–	R

Static solutions were implemented with R, interactive solutions with Tableau and ArcGIS Online.

TABLE 3.5: Overview of implemented visusalisations for Chile

Vis. Type	IMPC (IN)	IMPC (OUT)	Country (IN)	Country (OUT)	Tool
Flow Map	x	x	x	x	R
Choropleth Map	–	–	x	x	ArcGIS Online
Graduated Symbol Map	x	x	–	–	ArcGIS Online
Time Series	–	–	x	x	Tableau

the case of the here implemented visualisations, the additional interactive features in some of the visualises are partly interfering with this concept.

3.7.1 Reviewed visualisation types

All reviewed visualisation types are described in this section. The ones that were implemented are described in more detail, whereas only a brief overview is given of the ones that were reviewed but not implemented.

Origin Destination matrix

The Origin Destination Matrices (ODM) were created in R by using the package 'plotrix'. The visualisation was implemented in a static form for LOD 2 for the US data set (inbound and outbound flows), thus displaying US IMPCs at origin and

TABLE 3.6: Overview of implemented visualisations for the USA

Vis. Type	IMPC (IN)	IMPC (OUT)	Country (IN)	Country (OUT)	Tool
OD Matrix	x	x	–	–	R
Flow Map	x	x	x	x	R
Choropleth Map	–	–	x	x	ArcGIS Online
Graduated Symbol Map	x	x	–	–	ArcGIS Online
Time Series	x	x	x	x	Tableau

destination countries at destination and vice versa (see figure 3.2). The colour scale is set in blue values and is applying a logarithmic scale, resulting in smaller perceived differences for value changes in larger numbers, relative to value changes in small numbers. The absolute numbers are plotted inside the cells (see Appendix C.1 for example code for the ODM).

For all 24 matrices that were generated, the colour scale is identical and consistent over all CWs and years. This allows a comparison between the individual CWs.

This visualisation type was perceived very positively in the requirements analysis. An implementation is only reasonable for data sets and LODs with multiple origins and destinations. Thus, in this study it was only applied for the US data set for LOD 2 because for CL, all international postal traffic runs through the same IMPC (Santiago de Chile). If there are too many origins or destinations (rows or columns), the visualisation might be difficult to read because of information overload. In the present study, a minimum threshold value of 100 items was defined for displaying the number of items. Other possibilities to avoid information overload is selecting a number of specific postal partners to display.

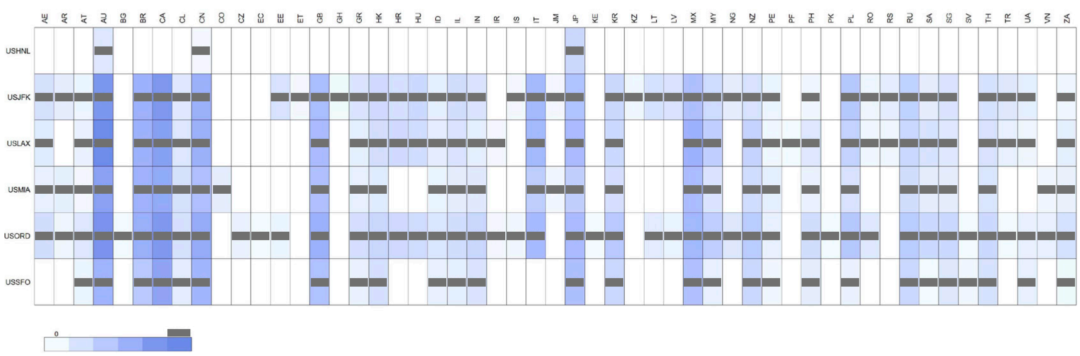


FIGURE 3.2: ODM for US outbound, LOD 2, CW 4 2015. Rows = origins (US IMPCs), columns = destinations (countries).

Flow map

The flow map visualisation was also created in R in a static display by using the R packages 'maps', 'geosphere' and 'scales' and altering open source code provided by Flowingdata[10]. The flow lines were drawn using great circles rather than straight lines. According to the author of Flowingdata, aircrafts and ships actually do use great circle routes. However, the routes displayed in the flow maps do not represent the actual route of the postal shipments, since the code was altered such that all of the lines are completely displayed on the map, which is crucial because the focus lies on the connections between origins and destinations. Thus, a line that runs out of the map to re-enter on the other side would hinder the analyst in the analysis of the flows. The code used for the flow map in figure 3.3 is provided in Appendix C.2.

[10]http://flowingdata.com

The visualisation was implemented for both the US and the Chile data set, and for both inbound and outbound postal traffic (figures 3.3 and 3.4, respectively). It was implemented for LOD 1 and LOD 2. For the LOD 1, normalised data was used for computing the line thickness (items per capita of the sending or receiving country, respectively). For LOD 2, absolute values were used. In both LODs, the maximum width of the lines was consistent across all years and CWs to allow a relative comparison between and within the six weeks periods of one year. A legend showing maximum and minimum flow widths with according values was shown in the usability study (not depicted in the graphs shown in the thesis for legibility reasons). Again, a threshold value of 100 items was set for the US data for LOD 2 in order to prevent a visual overload and to show the same data as in the ODM.

For Chile, the IMPCs of the origin (inbound) and destination (outbound) countries were mapped, as all international postal traffic enters or leaves Chile through the same IMPC. For the US data set, the US IMPCs were mapped, but for the sending (inbound) and receiving (outbound) countries, the respective country centroids were mapped. This resulted in more clarity in the map, as there would have been even more lines crossing each other if flows were mapped from US IMPCs to other countries' IMPCs. For each CW one flow map was produced, both for inbound and outbound postal flows.

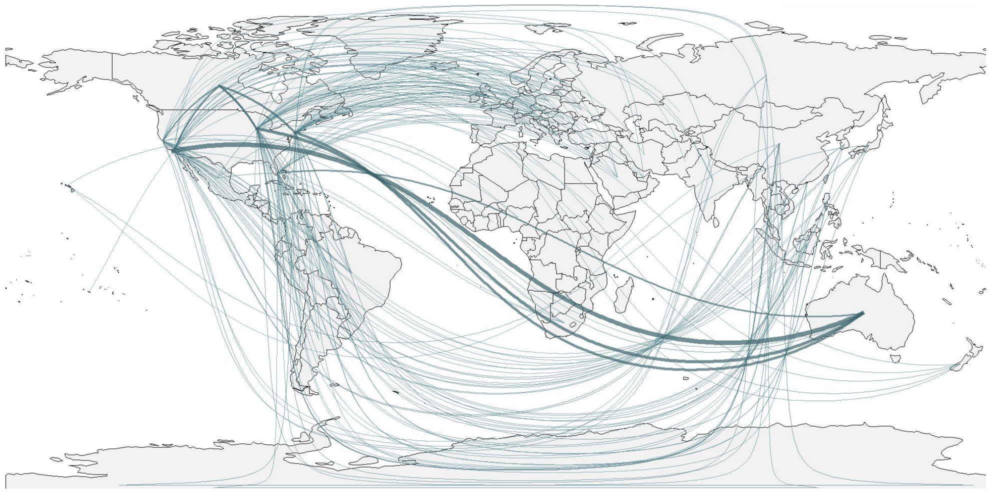


FIGURE 3.3: Flow map for US outbound, LOD 2 (traffic from US IMPCs to country centroids), CW 4 2015

Choropleth map

Choropleth maps were implemented for LOD 1 for both data sets (US and CL). For each CW, a map was created with normalised values showing items per capita of the sending or receiving country, respectively. Five classes were defined according to a quantile classification of the complete data set (data of the six weeks for all three years). In order to enable comparability across the CWs, the defined classes were

FIGURE 3.4: Flow map for Chile inbound, LOD 2 (traffic from world IMPCs to Santiago de Chile) in CW 31 2014

kept consistent across all CWs and years. The applied classification may not be perfect for the single CW, as the data is characterised by a high range in values and an uneven distribution with a few very high and many low values. If one would just look at a single week, natural breaks classification would lead to better results, but since this study aims at comparing different snapshots, identical classes are needed. In order to minimise the effect of this skewed distribution, the classes were assigned according to the logarithmic scale of the data. Another option would have been to choose a continuous scale with the identical maximum value across all weeks, however since the data is defined by clear administrative units (country borders), classification is found to be more appropriate in this case to make differences better distinguishable. The maps were implemented with ArcGIS Online, in order to allow the participants to use the visualisation in an interactive way.

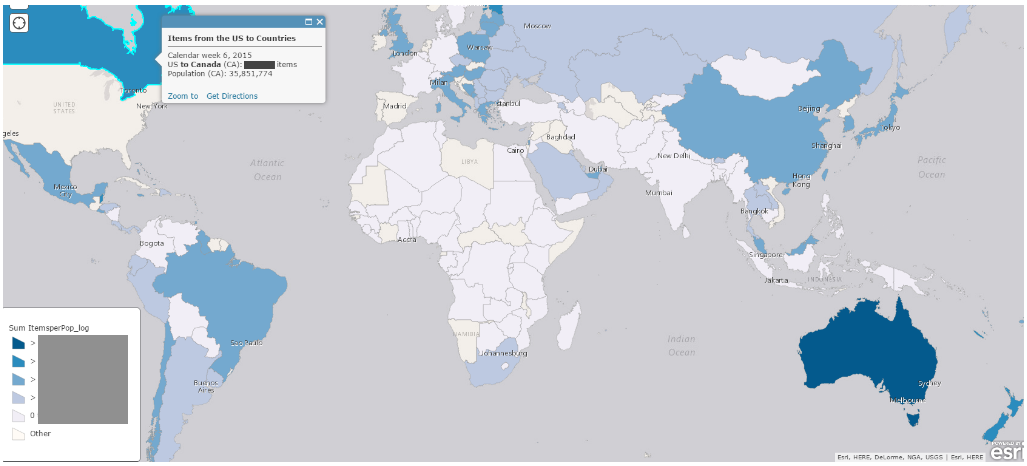
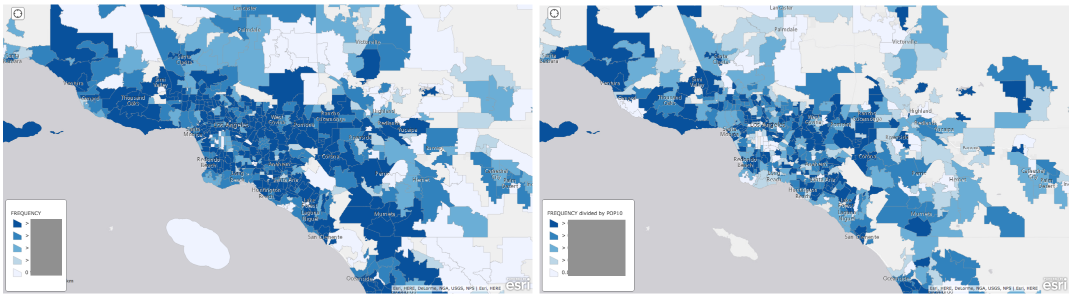


FIGURE 3.5: Choropleth map for the US, outbound, CW 6 2015



(a) Los Angeles ZCTA counts (b) Los Angeles ZCTA normalised

FIGURE 3.6: Los Angeles ZCTA (counts and items per capita)

The following additional options for data analysis with this visualisation type have been considered but were not further pursued (see Fig. 3.6 and 3.7):

- Implementation at different LODs such as ZIP code level (see figure 3.6)
- The display of ratios for showing the variation within one country over time. The ratio of items per population and week (RATIOweek) and the average of items per population per week (RATIOmean) could be computed. RATIOweek/RATIOmean could be computed and displayed for each country and week. Thus, values that lie above or below the average can easily be distinguished. This method might be very useful to detect anomalies and potentially also patterns (see figure 3.7).

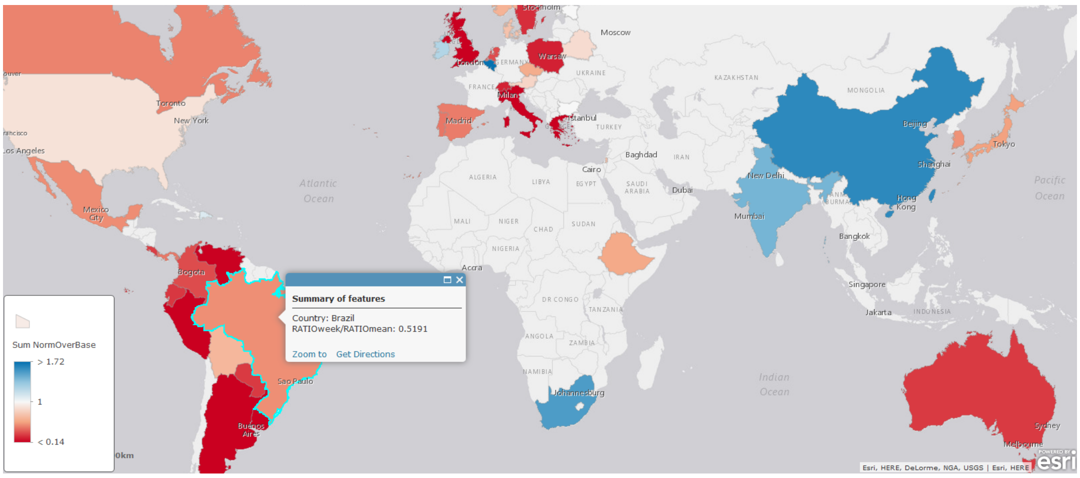


FIGURE 3.7: Ratio of items per population of CW 31 over the average items per population per week (CL inbound)

Graduated symbol map

To represent count values of the number of items sent between countries, a graduated symbol map was chosen. It was implemented in ArcGIS Online to provide the users with some interactivity. Six subsequent calendar weeks were shown on the same screen (for one CW, see figures 3.8 and 3.9). The background map was chosen in a light grey colour in order not to distract the user from the data. The size of the circles represent the number of items either sent to or from the country (US or Chile). By clicking on a dot, the user can access detailed information about the place of origin and destination and the exact number of items sent.

For the Chile data, actual values were used for the dot sizes. In the case of the US, the data was logarithmized to make differences in smaller values more visible, as the distribution of the values included few very large values and many small values.

FIGURE 3.8: Graduated symbol map for Chile inbound, CW 34 2014

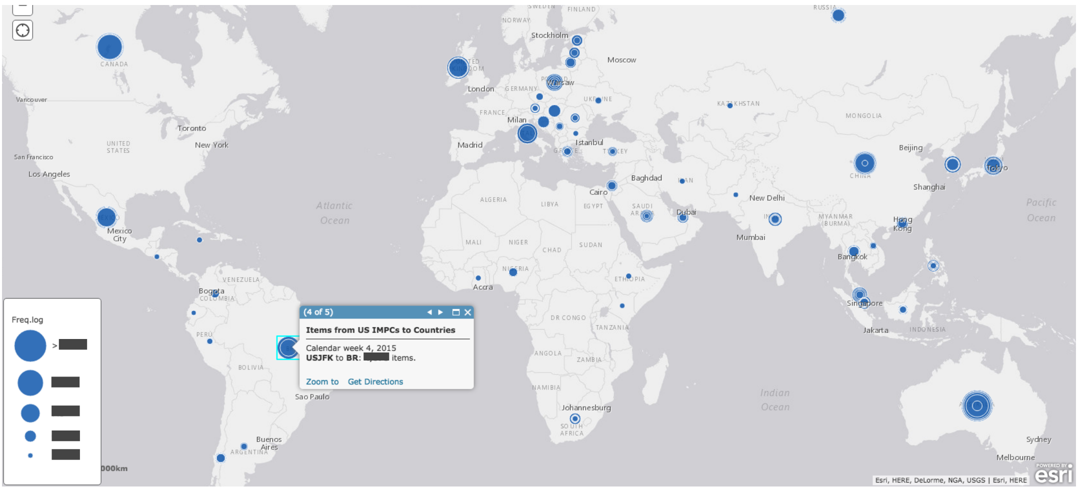


FIGURE 3.9: Graduated symbol map for the US outbound, CW 4 2015

Time series

The time series visualisation was implemented with Tableau. It enables the user to interact with the data, e.g. he or she can apply filters or read exact numbers by hovering over the lines.

As opposed to the other visualisation types, the data was not split into single weeks here, but the whole time span of six weeks was was displayed. For Chile, all three years were displayed on the same screen, for the US, the years 2012 and 2013 were displayed on one 'page' of the screen, the years 2014 and 2015 on the next page and users could switch between them.

For Chile, LOD 1 was visualised, for the US, LOD 1 and LOD 2 were visualised (see figure 3.10 for US outbound, LOD 2). For LOD 2, a colour scale adjusted to colour blindness provided by Tableau was used for enabling visual differentiation of the IMPCs.

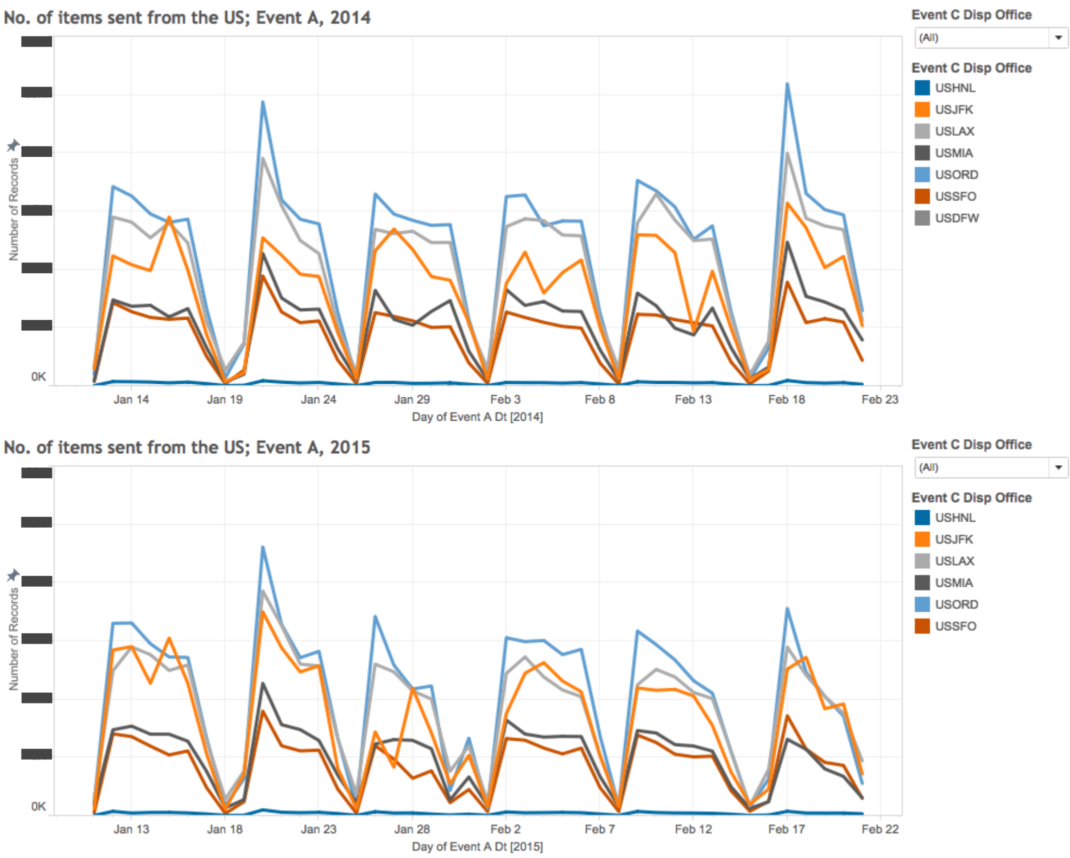


FIGURE 3.10: Time series for the US outbound, LOD 2, 2014-2015

Histogram of time difference *(not included in evaluation)*

The EDI message chain also contains information about the scheduled and actual time of departure or arrival of items. The time difference between the planned and actual arrival time (*Event D*) was computed for the CL inbound data and displayed in the form of histograms (see figure 3.11 for the histograms of CW 33 of 2012–2014). This provides insights about the delay and the amount of items that were reported during the respective time interval. This analysis/visualisation was not included in the usability study, however it might provide an additional way for confirming discovered anomalies in the data.

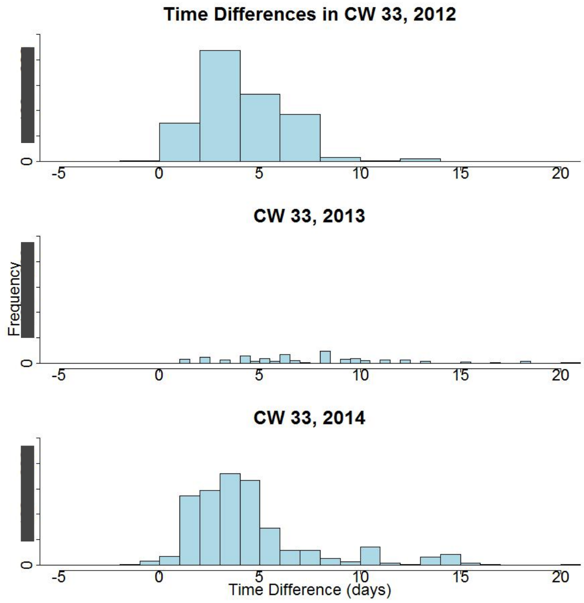


FIGURE 3.11: Histograms of time difference (in days) between scheduled and actual arrival in CW33 of 2012 (top), 2013 (middle), 2014 (bottom). In 2013, there were clearly fewer items reported than in the other two years.

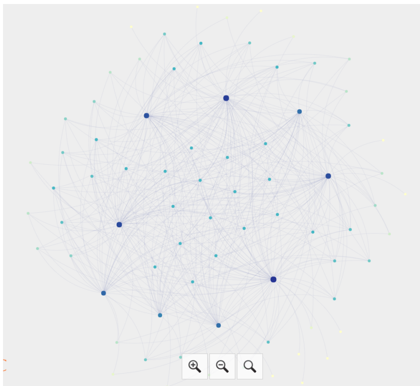
Graph *(not included in evaluation)*

The graph visualisation type received medium ranking in the requirements analysis. It is a rather abstract visualisation type and not easy to understand for lay people (cf. interview results in table 4.3), however it may prove useful during analysis conducted by data experts or programmers. Graphs can be drawn in R or Gephi (see figure 3.12), a tool that is specifically designed for this type of data visualisation.

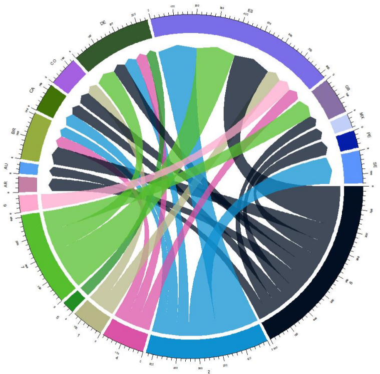
Adding interactivity, which is possible with a plugin available for Gephi, improves the situation, but it was decided not to include it in the final implementation, as the target group consists mostly of postal experts, not data scientists or programmers.

Chord Diagram *(not included in evaluation)*

Chord diagrams were implemented with R (R package 'circlize') for different LODs for both Chile and the US (see for example figure 3.12). However, it was eventually decided to not include it in the user evaluation, as it was only implemented in a static version. In order to provide valuable insights, it should be implemented in an interactive way.

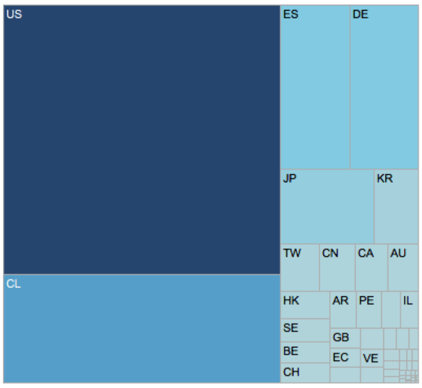


(a) Graph

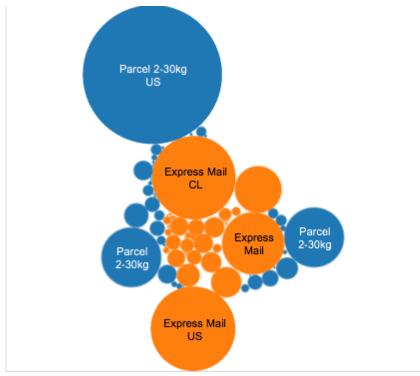


(b) Chord Diagram

FIGURE 3.12: Graph and Chord Diagram



(a) Treemap



(b) Bubble Chart

FIGURE 3.13: Treemap and Bubble Diagram

Treemap ***(not included in evaluation)***

First tests were run with treemap visualisations in Tableau (see figure 3.13), however it was decided not not include it in the final user evaluation in order to be able to conduct the usability study within a reasonable time frame.

Bubble chart ***(not included in evaluation)***

Like with the treemap visualisation, first tests were run with bubble chart visualisations in Tableau (see figure 3.13), but it was eventually decided not to include it in the final user evaluation due to time restrictions of the usability study.

3.8 Usability Study

For the evaluation of the implemented visualisation types, a usability study was conducted with the actual user group, i.e. postal experts. In contrast to Yang et al. (2016), who conducted a quantitative user study, the focus of this study lies more on the qualitative feedback. For this purpose, usability is defined as a quality metric as described by Nielsen (2001). Three main points are addressed in order to assess the usability of the selected methods:

- Users' subjective satisfaction and opinion about usability
- Success rate (can users perform the tasks)
- The time a task requires

The time needed to perform the tasks was recorded, but is not the main focus of this study, as it can be assumed that the users would have enough time to perform similar tasks in reality. The participants were asked to perform certain tasks with the visualisation methods and data sets described in the previous section (cf. tables 3.5 and 3.6).

This section is structured in the following way: First, an overview about the participants is given, followed by the second section describing the tools used for the study. In the third section, the usability study procedure is explained in more detail.

3.8.1 Participants

The six participants of the usability study are all employees of the UPU. They are experts in the postal sector but not necessarily very familiar with data analysis or data visualisation (see table E.2). The group of participants consisted of five men and one woman. All of them work in a different sector within the UPU and have a different professional background. The following positions are held by the participants: Economist, Statistician, Legal Expert, Manager Markets Development and Market Research, E-Services and E-Commerce Expert, Environmental and Sustainable Development Expert.

3.8.2 Apparatus and materials

In the tasks given to the participants, they were asked to focus on detecting patterns and anomalies by exploring the visualisations of the selected data sets. Both case studies (US and CL) consisted of six consecutive weeks, as described in section 3.2.1. For the US, the case study includes the events of snow storms during three short periods in January and February 2014. 2013 shows a major lack of data for the US data set, presumably because of a data reporting issue. The CL data set includes a strike that occurred in August 2013 and lasted for about three weeks.

The participants were exploring the data with the visualisations on a computer screen. They were also asked to complete an online survey on the same computer screen during the session. For the time series visualisation, they had to switch to a laptop, as this visualisation was implemented with an application that requires a license (see photography of study setup in Appendix E.1).

3.8.3 Study design

For the usability study, the data sets of the case studies were split into three different sections: LOD 1 (Country), LOD 2 (IMPC), Time series. Between the participants, these sections were rotated, i.e. participant I and IV started with LOD 1, participant II and V with LOD 2 and participant III and VI with Time Series. Within the sections, the data sets and visualisation types were systematically rotated, according to the Latin square (Zhang, 2010). For an example of the design for participants I and II, see table E.1.

The selected visualisation types (flow map, choropleth map, ODM, graduated symbol map and time series) were tested at the respective LODs by collecting reactions of the participants through observation, questionnaires and interviews. The following reactions were collected and studied: System Usability Scale (SUS), time, personal opinion regarding usefulness of the visualisations, insight (detected patterns and anomalies and additional information such as ideas about underlying reasons for those patterns and anomalies).

Test Procedure

At the beginning of the session, the participants were introduced to the purpose and procedure of the study. They were reminded that it is not them who are being tested, but that they are helping to assess the visualisation methods regarding the detection of patterns and anomalies in large origin-destination flows in postal data. As the study needed around 3.5 hours of full concentration per participant, they were told to take breaks whenever needed.

In a first step, they were asked to provide some information about their previous experience and familiarity with different visualisation methods, data types and tools (see tables E.2).

Next, they were asked to perform the tasks with the first set of data and visualisations. They were given a short introduction about the data set they were going to look at (inbound or outbound flows, aggregation level (CWs), normalisation, LOD, etc.) and the tools (interactive, static, etc.). They were given some time to familiarise themselves with the respective tools and ask some general or technical questions before starting the tasks. Only the time they needed to actually perform the task was recorded.

After completing the tasks for each section and before continuing to the next, the participants were asked to state their opinion about the usefulness of the visualisation types they just worked with by completing a very brief online survey. They were asked to rate the visualisation types regarding their usefulness between 1 (not useful at all) and 5 (very useful) and briefly note the main advantages and disadvantages with their own words.

After completing all task sessions, the participants were asked to fill in a System Usability Scale (SUS) questionnaire consisting of ten questions (see Appendix E.4).

Finally, a semi-structured interview was conducted to gain more in-depth knowledge about the participants' opinion and suggestions regarding the visualisation of the data. They were asked to state their preference regarding visualisation methods and explain difficulties that they encountered while exploring the data. They were also asked to comment on the advantages and disadvantages and potential improvements of the different visualisation types.

Through the entire session, the experimenter took observational notes (specifically noting the moments of insights and moments of difficulty with the visualisations). The session was audio-recorded for future reference.

Chapter 4

Results

This chapter presents the results obtained during this study. It is structured as follows: the first section briefly repeats the research questions. The second part presents the results from the requirements analysis, as these findings were crucial for selecting data sets and especially for selecting the visualisation types that were implemented. The third section provides an overview of the results from the analysis of the EmIS reports, as these findings were taken into account for selecting the data sets for the case studies. The fourth part presents the results obtained from the usability study that was conducted with postal experts at UPU.

4.1 Research Questions and Hypotheses

RQ 1: How do the spatial properties of the postal flow network and their encoding influence analysts' ability to visually interpret the spatio-temporal patterns and detect anomalies?

Three main properties are analysed:

1. The impact of the *encoding* of origins and destinations (the granularity of spatial units) on the ability to detect spatial and temporal patterns and anomalies
2. The impact of *connectivity* on the spatial an temporal patterns and the ability of detecting spatial and temporal patterns and anomalies
3. The impact of *distance* between origins and destinations on the spatial an temporal patterns, i.e. whether the applied methods reveal spatial autocorrelation

Based on previous findings about using aggregation in the visual exploration of data (Thomas and Cook, 2005; Andrienko and Andrienko, 2010), the thesis hypothesises that granularity of the spatial units affects the visualisation and detection of patterns and anomalies (Hypothesis 1a).

In accordance with previous findings, the thesis assumes that connectivity affects the visualisation of the postal data as well as the ability to analyse and detect patterns and anomalies in the data (Andrienko and Andrienko, 2010; Wood, Dykes, and Slingsby, 2010) (Hypothesis 1b).

Further, the thesis assumes that spatial autocorrelation appears in the postal traffic data, as is found in many other geographic flows (Tobler, 1970; Wood, Dykes, and Slingsby, 2010) (Hypothesis 1c).

RQ 2: Which visualisation types should be applied or combined to detect and analyse spatio-temporal patterns and anomalies in large postal OD data, and why?

1. What are potential and limitations of the implemented visualisation types?
2. How do the selected visualisation types compare to one another or complement one another?

Visualisation approaches that have been used in the analysis of other flow networks can be applied to the postal OD data for the detection of patterns and anomalies. The thesis hypothesises that some visualisation types are well suited for showing a global overview of the flow structures (e.g., the flow map (Wood, Slingsby, and Dykes, 2011)) while others are less applicable for analysing data sets with large numbers of spatial interactions (e.g., the choropleth map (Sander et al., 2014)) (Hypothesis 2a).

Interactive visualisations can reveal more detailed information about patterns and anomalies than static ones, however the latter can provide a quick first overview and thus a combination of different visualisation types is favourable (Shneiderman, 1997; Roberts, 2007; Keim et al., 2010) (Hypothesis 2b).

The study further assumes that interactive visualisations are preferred by the participants (Hegarty et al., 2009) (Hypothesis 2c).

RQ 3: How suitable are the selected visualisation methods for detecting cyclical patterns and anomalies in the postal data?

Cyclical patterns and anomalies in the postal data can be detected on various spatial and temporal scales and indicate underlying causes such as weekends or Christmas season. Based on previous research, the study expects times series to be practical for showing cyclical patterns in the postal data, while the other tested visualisation types are less useful for this purpose (Keim et al., 2010) (Hypothesis 3).

RQ 4: How can the detected anomalies be explained by the applied visualisation methods and additional (semantic) information (e.g. seasonality, socioeconomic development, conflicts, etc.)?

The thesis hypothesises that the applied visualisation types can help to understand underlying processes (Andrienko and Andrienko, 2010), though additional variables (e.g. about socioeconomic development, seasonality, conflicts, etc.) might be needed for a reliable interpretation (Hypothesis 4).

TABLE 4.1: Demographics of online survey participants

Age (yrs)	Average 36.9	Male 38.6	Female 34	
Gender	Male 10	Female 3		
First language	EN 30.8%	ES 38.5%	FR 15.4%	Other 15.4%
Education	Bachelor's degree 15.4%	Master's degree 69.2%	Doctoral degree 15.4%	

4.2 Requirements Analysis

The requirements analysis provided insights about which visualisations the postal experts consider useful, which helped to select the visualisation types to implement for the analysis.

4.2.1 Online questionnaire

The online questionnaire was answered by 13 postal experts working in various departments of the UPU (see table 4.1). Only one of them had a background in Geography. The main purposes that the postal data has for them are listed in figure 4.1. For a majority of them, statistical analysis, online reports and publication are the most important uses of postal data, followed by written reports and process optimisation. Most participants (85%) stated that their main target groups are external to the UPU. For 77% of the participants, governments and policy makers are among their main target groups (see figure 4.2). If results of analyses conducted by the participants are published, it is mostly in reports, either internal (85%) or public (54%) (see figure 4.3).

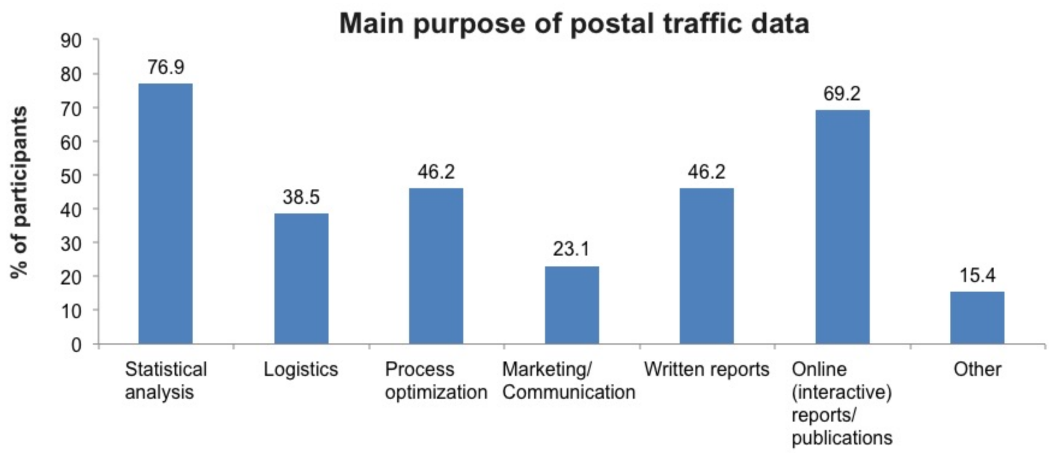


FIGURE 4.1: Main purposes of postal data as reported by participants

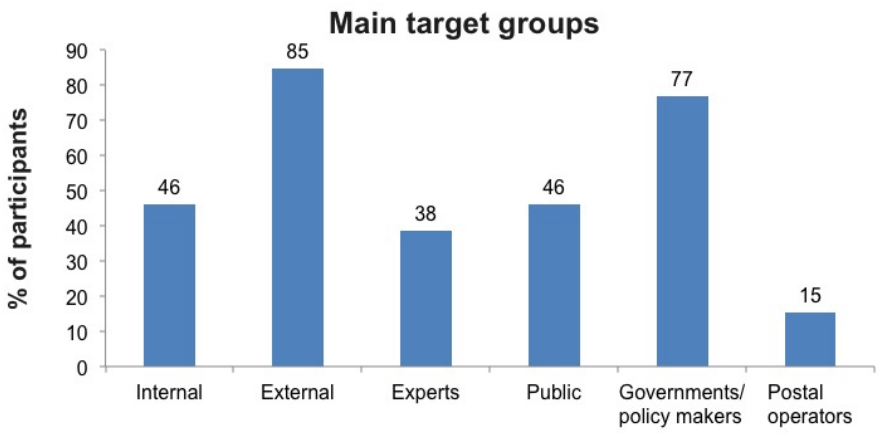


FIGURE 4.2: Main target groups of postal experts

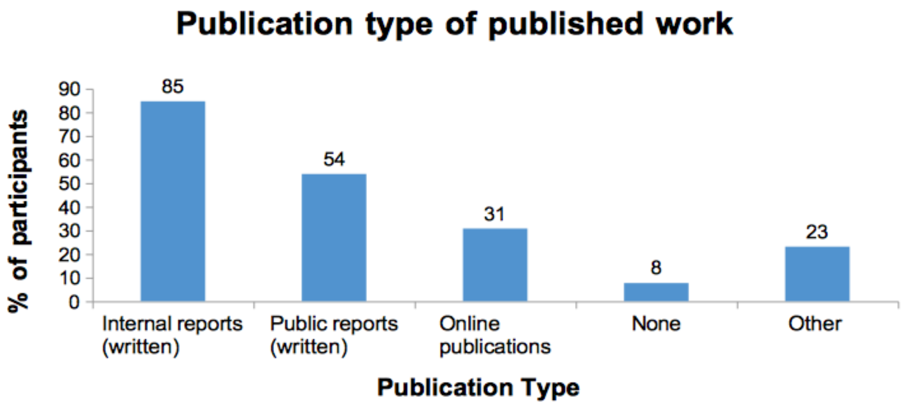


FIGURE 4.3: Publication type of documents published by postal experts

Regarding the importance of the spatial scale, for many of the participants the global scale is most important, followed by the national and the regional scale. The importance of the temporal scale varies strongly between the participants. Overall, the annual temporal scale seems to be the most important, followed by monthly, weekly and daily scale. For more information, see figures (see D.1 and D.2 in the Annex).

Many of the participants are aware of certain spatial and temporal patterns, as can be seen in figure 4.4. The anomaly awareness is lower. However, some participants are aware of certain types of anomalies (see figure 4.5). When asked whether they think that visualisation could help detect patterns and anomalies in the data, a majority (92%) answered with 'yes' for the patterns, and 85% responded with 'yes' for the anomalies.

Regarding the expected usefulness of different visualisation types for detecting patterns and anomalies, the answers varied widely between the participants (see figure D.4 in the Annex).

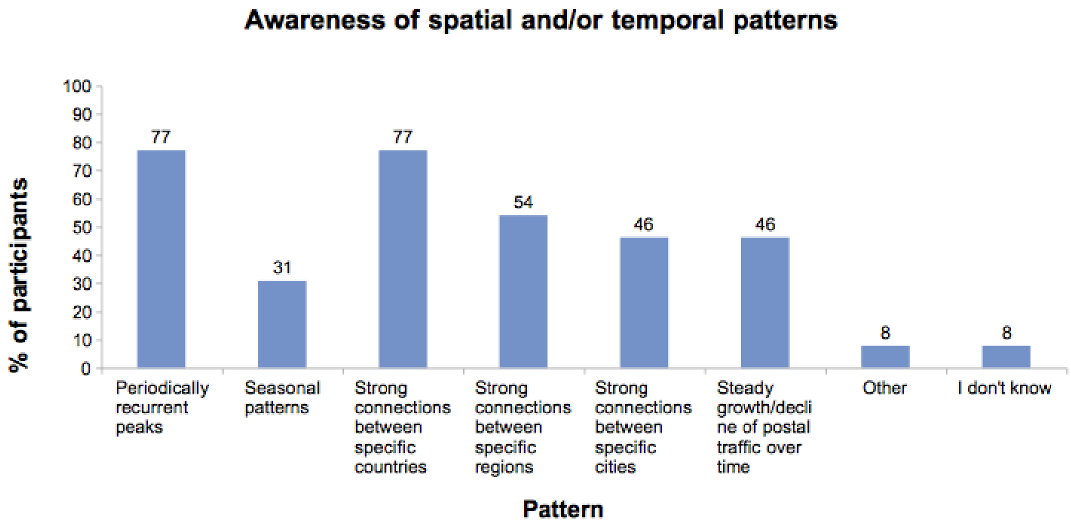


FIGURE 4.4: Awareness of spatial/temporal patterns

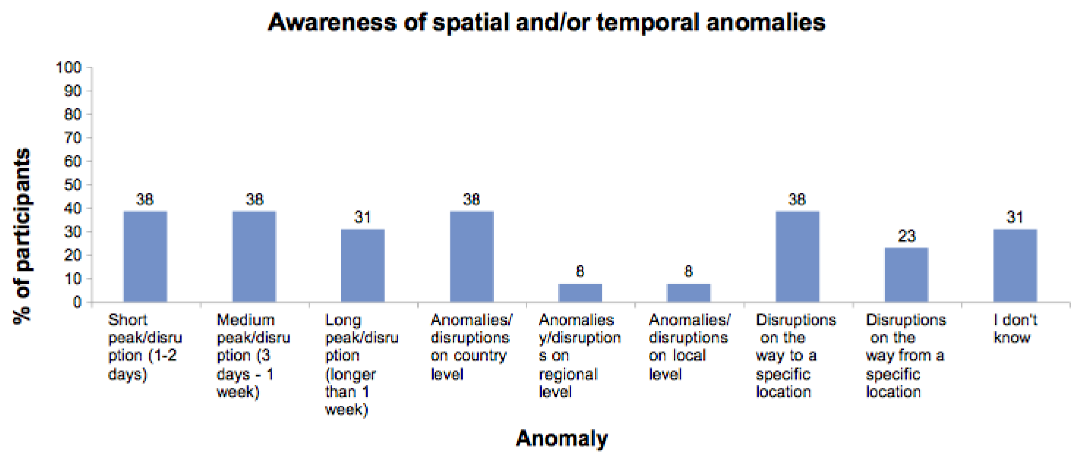


FIGURE 4.5: Awareness of spatial/temporal anomalies

Only 8% of the participants stated that they were completely happy with the current solutions they have at hand for their work regarding data analysis. 54% said they would appreciate having other tools to work or combine with the currently used solutions (see table 4.2.)

4.2.2 Expert interviews

The expert interviews were conducted with four employees of UPU in order to get a more thorough understanding of the situation and needs of postal experts with regard to data analysis and visualisation.

Regarding the target groups, it confirmed the results from the online questionnaire: the first main target group consists of the affiliated governments and policy

TABLE 4.2: Opinion about currently used solutions

Statement	% of participants
I am completely happy with the current solutions, I don't need anything else for my work.	8%
I can work well with the current solution, but I have some ideas about how they could be improved.	23%
I would be happy to have other tools at hand to work with, or to combine with currently used solutions.	54%
Other: mapping real time data and workflow optimisation	8%

makers, the second consists of the posts and postal operators of the member countries.

Key interests

One key interest of the interviewees lies in the analysis of the big postal data and developing an index for postal network development. The index ranks countries according to four critical dimensions: network reliability, network reach (global connectivity), relevance (for people/customers) and network resilience (is the network ready to be confronted with disruptions like technological, economical, environmental shocks). Another focus of the interviewees lies in analysing the EDI messaging system in order to make predictions about the future postal network behaviour for all people that are involved (delivering companies, airports, customers, etc.). The analysis of the EDI messaging system can also provide information regarding network efficiency, trends in terms of flows of goods, comparison of the postal network to other networks, main bottlenecks, improvement of postal infrastructure, etc.

Spatial and temporal scale

The interviewees stated that the main focus so far has been on a global level, but there is a high interest in better understanding what is happening on a local level, always from a global connectivity aspect (how well are specific locations connected to the rest of the world). There is also an interest in seeing how things might change if the scale of analysis is changed. Most analysis has been conducted for industrialized countries, as the quality of the available data is better than for less developed countries.

Similarly, for the temporal scale, most of the indicators focus on a yearly scale, but there is a high interest to analyse different temporal scales. The interest for future analysis lies in a monthly, weekly, or daily scale. For the prediction model, even hourly scales are of high interest.

Types of patterns and anomalies (awareness and interest in detection)

The interviewees mentioned various patterns they are aware of in the data. These include periodically recurrent peaks (e.g. Christmas), seasonal patterns, weekly patterns depending on working days of a country (Sunday slowdown and peaks on Mondays and Tuesdays in western countries, Friday slowdown in Muslim countries), strong connections between specific countries/regions/cities, or a steady growth of postal traffic over time. They also stated that for a long time, there was a clear seasonality towards the end of the year (holiday season) but now with growing e-commerce, this has shifted also towards other events with discount rates for online purchasing such as Black Friday, Halloween, Singles Day in China, etc. This results in higher volatility in the data, as the global marketing ecosystem is "sort of starting to replace the traditional holiday season pattern", as one of the interviewees put it.

Regarding anomalies, the interviewees mentioned that sometimes things get held up in customs for a long time in less well organised countries, which causes delays in the delivery. They also mentioned weather related events that could help to explain why there was a bottleneck in the mail delivery from an affected country (short time period) or political conflicts that are expected to have an impact, however on a longer period of time.

The interviewees referred to events like weather disruptions, natural disasters, etc. as important things they would want to detect in the postal data. They stated that weather is extremely important in logistics and e-commerce, in both purchasing behaviour of people and logistics operations. Strikes were also mentioned as an important factor to look at, as the postal industry is still more labor- than capital intensive. Further, political conflicts within and between countries are a key factor in the postal network as well as health issues that might affect the network strongly (for example Ebola).

It was stated that the time needed for the network to get back to a normal situation depended on both the type of the disruption and the resilience of the postal infrastructure of the specific place. The latter in turn strongly depends on a country's resources and whether it is more or less developed.

Expected usefulness of visualisation

All of the interviewees state that they regard visualisation as being very important in their work. As main advantages they view the fact that visualisations can help to see immediately where and when there could be a problem in the data. Further, they feel that people tend to forget what happened in the past and think that visualisations could be helpful to show that it is often the same areas that are affected by disasters. In their opinion, visualisations could be helpful for better explaining anomalies, however they think it would be difficult to explain *why* there was a disruption.

They further express concerns regarding data quality, as it is only possible to visualise what is in the data. This means that if the data quality is poor, the visualisations will not be able to show correct information.

TABLE 4.3: Expected usefulness of visualisation types

visualisation type	Key observations
Chord diagram	Good for marketing/communication, not so much for analysis. Needs to be very interactive.
Graph	Interesting and useful, but not easy to understand for everyone.
OD matrix	Very useful, gets the point across, but might be confusing if there are too many cells. The differences in the colours might not always be easy to differentiate. Helpful when exact numbers are printed in cells.
Flow map	Very informative because one can see start and end of a flow and the patterns. Absence of exact values and overlapping lines can be a problem.
Choropleth map	Very simple and easy to understand. It brings across the point if there is only one attribute to map.
OD map	Not easy to understand because it needs a lot of explanation and the user needs to know what he/she is looking at.
Statistical plots	Get the point across, but are very boring. Used often in analysis.

From the visualisation types that were discussed in the interviews, the OD matrix and the flow map were seen as most useful. For a collection of the key statements made in this regard, see table 4.3.

4.3 EmIS Reports

The EmIS reports were collected and classified as described in chapter 3.5. They were used to identify the countries for the case studies and select a reasonable time span, as the data first had to be extracted from the EDI messages database by the UPU.

The overall number of events that are reported vary highly between the different categories, as can be seen in figure 4.6. Between the countries, there is a variation in the number of total reported events as well as the number of events per category, as can be seen in figure 4.7 or in the Appendix in figure B.2.

4.4 Visualisation Tools

As mentioned in the previous chapter, several tools were tested for creating the visualisations. This section provides a brief summary about the advantages and drawbacks that were found during this testing.

Tableau offers several possibilities for visualising data, some of which were tested for the postal data (see also 3.4). The tool includes options for geographic visualisations such as choropleth maps and graduated symbol maps. For those two visualisation types, however, ArcGIS was preferred after testing both tools, as it offers a higher degree of customisation for creating, presenting and sharing maps. Especially for comparing the CWs, ArcGIS Online proved to be very useful, because it allows, for example, to choose a classification type for the choropleth map or to set the class limits manually. Similarly, ArcGIS Online allows to make adjustments to

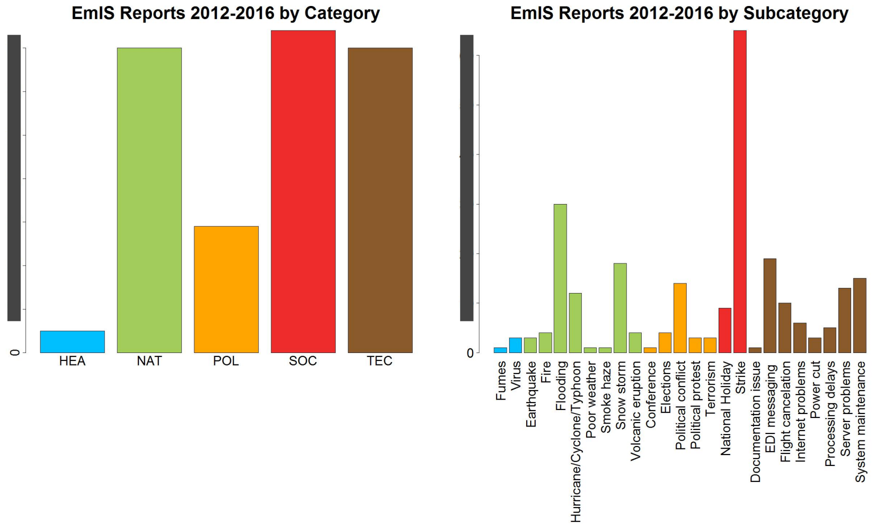


FIGURE 4.6: Categories of EmIS messages (2012-2016). HEA = Health, NAT = Natural, POL = Political, SOC = Social, TEC = Technical issues.

the maximum and minimum dot size in the graduated symbol map for comparing the CWs among each other. It further allows to present several CWs beside each other on the same screen for comparison. Tableau was preferred for visualising the time series in the form of a line graph. R also offers packages for creating interactive time series visualisations (e.g. R package 'dygraphs'). After testing R and Tableau, Tableau was chosen for this visualisation, as it offers a time scale (x-axis) on a daily basis, whereas the implementation with R works for a yearly or monthly scale. For the available data sets, yearly or monthly scales did not lead to a satisfactory result. Tableau enables linking several plots in a dashboard, which was found to be very convenient for analysts.

R was chosen for creating ODMs, flow maps and static chord diagrams. The tool allows customising all of the elements of the visualisation, but more technical skills are needed to create applicable visualisations than for existing tools like ArcGIS Online or Tableau. In this thesis, only static visualisations were implemented with R, however there are tools like Shiny[1] for example, which provide features for some interactive implementations. JavaScript libraries such as D3[2] could be used for creating an interactive chord diagram.

Graph visualisations were implemented with both R and Gephi. With R, it was only possible to create static graph visualisations, whereas Gephi provides options for adding geography for mapping the connections on a world map and for adding

[1]https://shiny.rstudio.com
[2]https://d3js.org

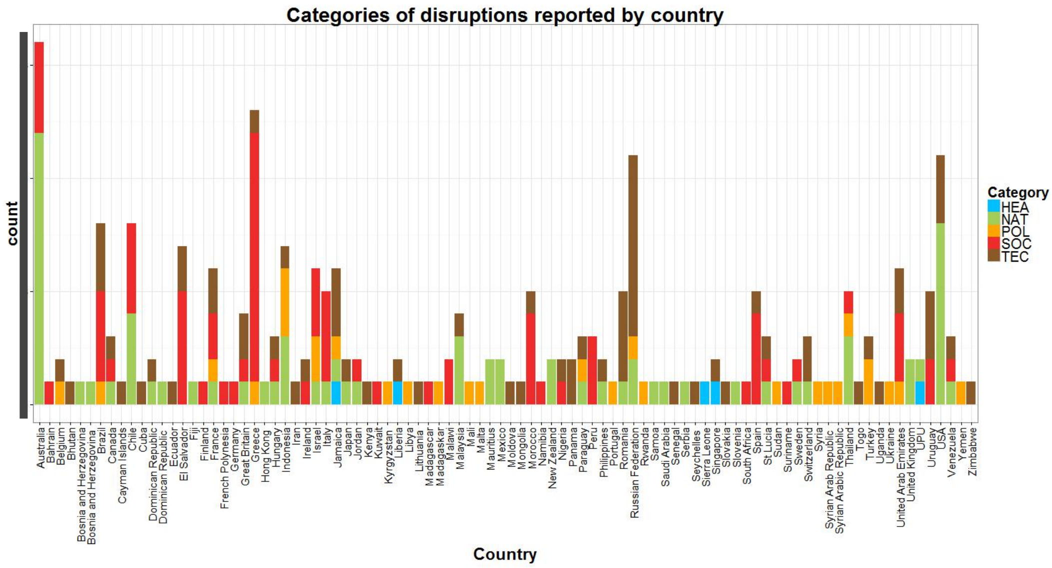


FIGURE 4.7: Number of EmIS messages submitted by country (2012-2016, categorised). During this time span, Australia reported the maximum number of disruptions (16).

interactivity by using specific plugins that are freely available (e.g. Sigmajs Exporter[3], GeoLayout[4]). Gephi is a powerful tool for visualising networks, but since the two case studies focused on two countries, the data structure was found to be suboptimal for this tool.

4.5 Time Difference between Planned and Actual Arrival

As was discussed in 3.7.1, the time difference between planned and actual arrival time was computed for the CL inbound data, however this analysis was not pursued further for the US data and not included in the usability study. The mean, median standard deviation values per CW are displayed in table 4.4. The year 2013, when the strike took place, appears different compared to the other two years (see also the histogram, figure 3.11). However, for a more reliable interpretation, a longer time span should be analysed.

4.6 Usability Study

In order to evaluate the implemented visualisation types regarding their usability in terms of detecting meaningful patterns and anomalies, a usability study was conducted with postal experts working in different fields within the UPU. The results of the different parts of the usability study are described in the following sections.

[3] http://sigmajs.org
[4] https://marketplace.gephi.org

TABLE 4.4: Mean, median and standard deviation (in days) of the time difference between planned and actual arrival (*Event D*, inbound) of CL data

	Mean (d)			Median (d)			Std. Dev. (d)		
CW	**2012**	**2013**	**2014**	**2012**	**2013**	**2014**	**2012**	**2013**	**2014**
CW 31	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CW 32	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CW 33	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CW 34	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CW 35	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
CW 36	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

4.6.1 System Usability Scale (SUS)

The SUS was conducted at the end of the task session but before the guided interview. The participants answered ten questions that were modified from the original version by Brooke (1996). The full SUS questionnaire can be found in Appendix E.4 and the results can be seen in figure 4.8. The results show that the variance is rather high, and there are several outliers. As the sample consists only of six people, this cannot be used as reliable evidence and should be conducted with a larger group of people.

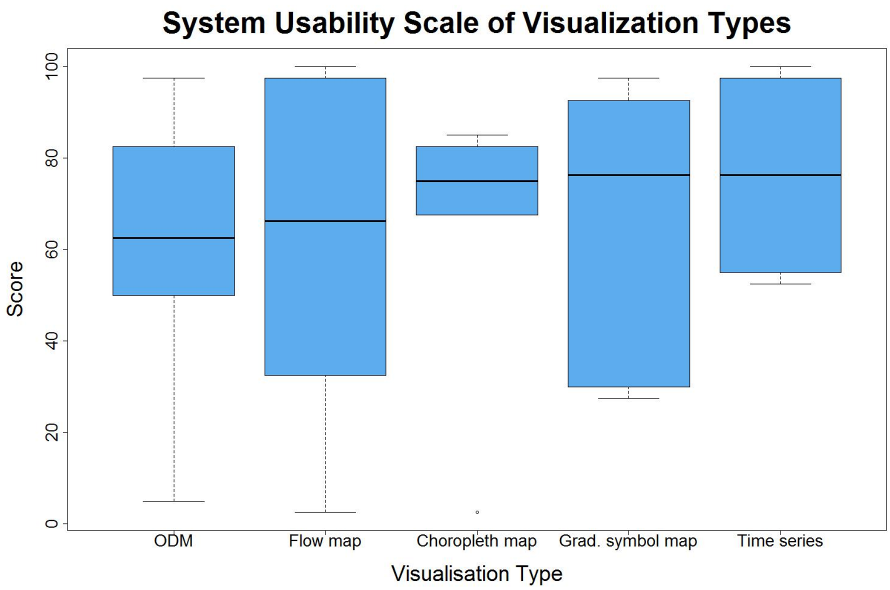
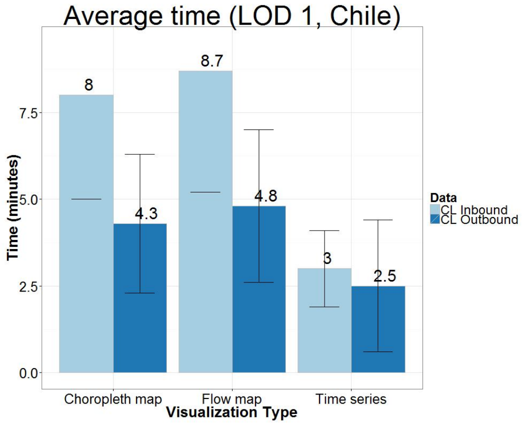


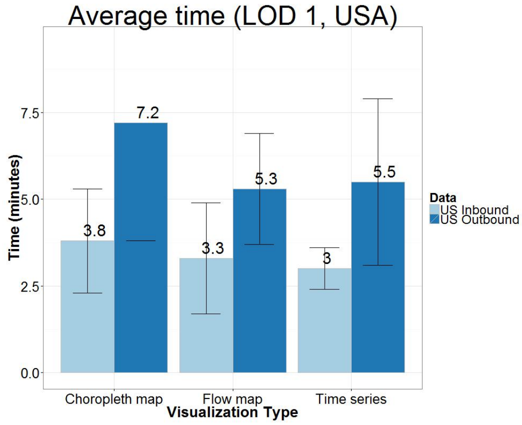
FIGURE 4.8: System Usability Scale: 100 = Maximum

4.6.2 Time

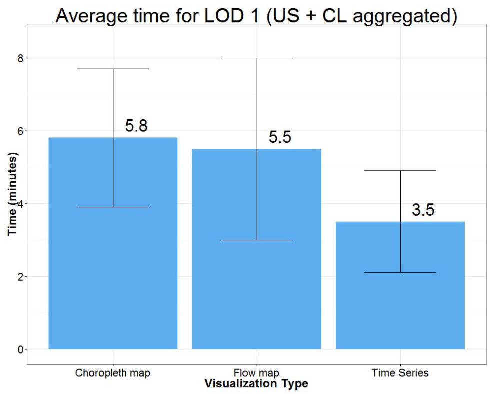
As was mentioned earlier, the time needed by the participants to complete each task was recorded. Plots (a) and (b) in figure 4.9 show the average time needed for exploring the inbound and outbound data for LOD 1, plot (c) shows aggregated



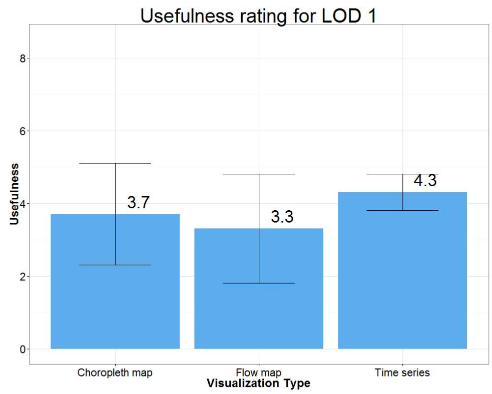
(a) Average time needed for LOD 1, CL



(b) Average time needed for LOD 1, US



(c) Average time needed across all data sets in LOD 1



(d) Usefulness rating for visualisations in LOD 1

FIGURE 4.9: Time needed for task performance per visualisation type and usefulness rating for visualisations in LOD 1

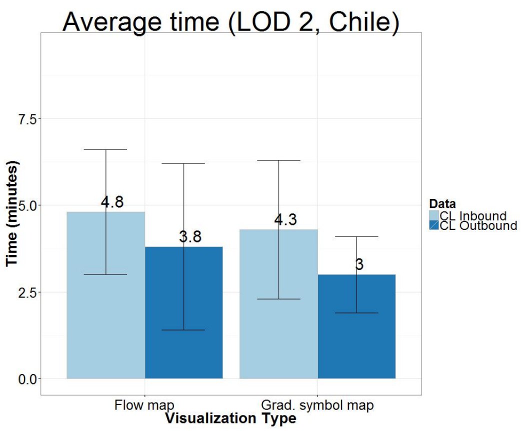
values of time needed for both data sets (US and CL). The same order of plots applies for LOD 2 in figure 4.10.

4.6.3 Usefulness

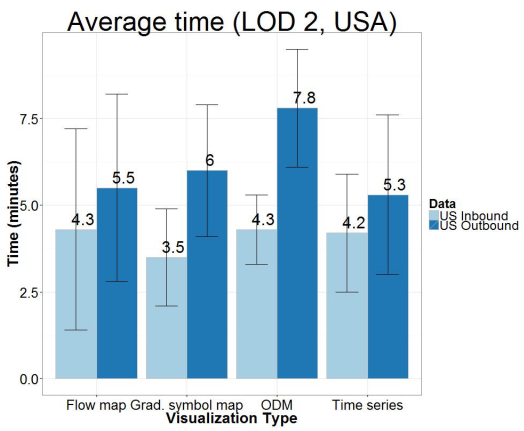
After completing each section of tasks, the participants were asked to rate the visualisations they had just seen. The question they were asked was *How useful did you find this visualisation type?*. They could choose between 1 (Not useful at all) and 5 (very useful). The results of this rating can be seen in plot (d) in figure 4.9 as well as figure 4.10 for LOD 1 and LOD 2, respectively.

4.6.4 Qualitative feedback

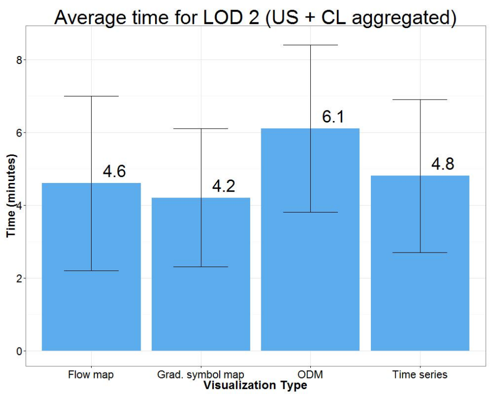
After each section of visualisations, the participants were asked to share their opinion about the main advantages and disadvantages of the specific visualisation types. A selection of the main comments regarding the advantages and disadvantages is shown in table 4.5. At the end of the whole usability study, the participants could



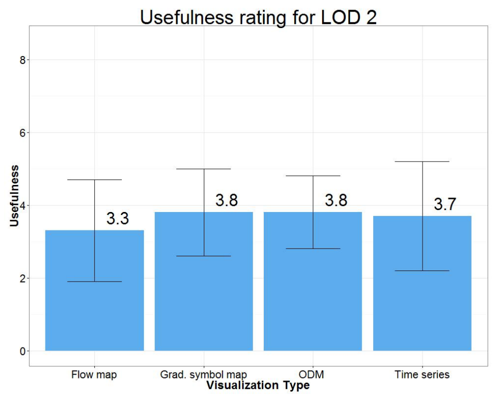
(a) Average time needed for LOD 2, CL



(b) Average time needed for LOD 2, US



(c) Average time needed across all data sets in LOD 2



(d) Usefulness rating for visualisations in LOD 2

FIGURE 4.10: Time needed for task performance per visualisation type and usefulness rating for visualisations in LOD 2

state their opinion in more detail in a guided interview. The following section summarises the statements that were made during the interviews regarding individual preference, interactive versus static visualisation, the benefit of the spatial dimension as well as difficulties that were encountered during the tasks.

Preference

All participants stated that they would over all prefer visualisations over other forms of reporting, such as text or tables. Some were absolutely in favour of visualisations because they could immediately see if something seemed unusual or wrong in the data. It was also mentioned that visualisations helped to remember what they had seen. Some stated, however, that it depends on the purpose. They would always prefer visualisations to obtain a comprehensive, quick overview, draw preliminary conclusions and see patterns and anomalies. However, in order to understand more about the reasons, they would need to look at tables. All participants agreed that they would prefer visualisations over other forms of reporting for the type of tasks that they were confronted with in the usability study (detecting patterns and anomalies).

The preferences regarding visualisation type varied between the six participants, with many of them favouring a combination of several visualisation types in order to look at different things. Two participants preferred the flow map overall. They stated that it depicted the connectivity of the postal network and its intensity in a richer and more telling way than the others. The main features that were perceived as helpful in the case of the flow map was that there were lines showing the connection between origin and destination on a map as well as the line thickness that gave the users an indication about the volumes.

Those two participants who preferred the choropleth map overall mentioned that, as a lay person, it was very easy to see which countries interact with each other and how strong. It was perceived as very user friendly by those who gave it a high rating. For one participant it was the easiest and fastest to read and he felt that the choropleth did not not mislead conclusions compared to others.

Two participants found that the time series visualisation was the most preferable. One gave familiarity with this visualisation type as the main reason. The other participant would prefer to combine the time series with the origin destination map. The time series was seen as very helpful in terms of getting information about mail volumes. However, the participants who favoured the time series visualisation over others also mentioned that the connectivity information was missing and that perhaps they would use a map visualisation first to get a quick overview and then use the time series (or origin destination map) to get more detailed information about the actual numbers.

When asked which visualisations they would chose if they had to combine two or three of them, the choropleth map was mentioned by almost all participants as one of the visualisations, along with the time series.

TABLE 4.5: Selection of statements made by participants regarding advantages and disadvantages of visualisation types.

Note that participants made contradictory statements; what was seen as advantage for some could be a disadvantage for others, hence similar elements can be found in both respective columns.

Visualisation type	Advantages	Disadvantages
Flow map	• Large flows/patterns are easily detectable	• Small flows/patterns difficult to detect
	• Much faster detection of patterns and anomalies than with figures and numbers	• Only the thickness of the lines points out changes
	• Good for identifying main sending and receiving countries	• Generally rather difficult to read
	• A very effective way of visualising the intensity of connectivity	• Can easily become a messy map when many routes to visualise at once
	• Useful for anomalies	• Less useful for patterns
Choropleth map	• The colours make it easy to identify changes in destinations and origins	• Not clear what actual amounts are
	• Quick assessment, easy to read	• Difficult to identify trends over the years
Graduated symbol map	• Easy to read, quick capture of trends	• Not easy to read
	• Fast detection of anomalies through the number of points	• Overlapping circles
	• Useful for patterns	• Less useful for anomalies
	• Possibility to dive deeper into data (exact values through clicking)	• Difficult to see patterns for one individual IMPC
OD matrix	• Good and quick overview	• Not mapped (no geographic reference)
	• Detailed information and patterns	• Difficult to detect patterns over time
	• Very useful when IMPCs are on the left side of the table (not on top).	• A series of one period on the same screen at the same time is preferred
	• Easy to see importance of countries and contributions of the different IMPCs	
Time series	• Easy to detect trends and seasonalities	• Not necessarily easy to detect anomalies
	• Good to identify disruptions between the patterns in certain years	• No spatial/geographic reference
	• Easy to identify trends and draw conclusions	

Most participants had the same preference in the visualisation type they would use for either of the tasks (finding patterns or anomalies), however there were some that would prefer one visualisation type for detecting patterns and another one for detecting anomalies. For example, one participant stated that he would use the choropleth map for patterns, but first the graduated symbol map for the anomalies because he felt that they were easier to detect with the graduated symbol map.

Interactive vs static visualisations

The participants were asked whether they would prefer static or interactive visualisations. Four participants stated that they would prefer interactive visualisations because they can get more detailed information out of them. The possibility of filtering information according to the needs or interests was seen as very useful. One participant preferred static visualisations for comparison (e.g. between years or weeks), but interactive visualisations for a more in-depth analysis. One participant found it very easy to detect patterns with the static visualisations, but for some of the anomalies it was more difficult – in this case, interactive solutions could be useful to confirm what has been detected with the static implementation.

Benefit of spatial dimension

All participants agreed that the spatial dimension added value for the tasks they were asked to perform. They found it more comprehensive and mentioned that it enabled a wider field of use, e.g. for analysis with regard to connectivity. It was mentioned that it helped to remember trends and/or anomalies in certain parts of the world and the detection of main gaps in certain areas or relations between specific countries.

Difficulties

The participants were also asked whether they had encountered any difficulties with any of the visualisations. Some mentioned the general problem of mapping such a high amount of information on a map and the danger of information overload and 'messiness' that comes with it, which makes it difficult to detect anything. The overlapping of the dots in the graduated symbol map was perceived as rather difficult by some participants. One participant found the flow map particularly cumbersome because it was static, another participant found the choropleth map very difficult to use, because he felt that he had to spend too much time to find out how it worked and it seemed to be too much work to find patterns and anomalies.

Some participants also perceived the normalisation of the values on country level as problematic. They did not see any benefits in showing numbers relative to population size. However, some agreed that normalisation makes sense.

TABLE 4.6: Known patterns and anomalies in the data

Data	Patterns	Anomalies
US (6 weeks)	• Strong connections with main partners (e.g, CA, CN, AU, BR)	• Major technical issue in 2013 (not reported in EmIS reports)
2012/13/14/15	• High global connectivity (outbound) low global connectivity (inbound)	• Snow storms in CW 4 and CW 6 of 2014
	• Growth in overall traffic over time	• Technical issue in CW 7/8 in 2015
	• Weekly and monthly pattern (outbound)	• Drop in flow from Canada in CW 6 2015
CL (6 weeks)	• Strong connections with main partners (US, LatAm, ES, SE, HK)	• Strike in CW 32-34 of 2013
2012/13/14	• Overall low global connectivity	• Backlog and catch up in CW35/36 of 2013
	• Growth in traffic over time	

4.6.5 Detection of patterns and anomalies

This section presents the insights the participants of the usability study obtained during the evaluation of the visualisation types. Table 4.6 provides an overview of the patterns and anomalies that are present in the two data sets. The patterns are generally assumed to be flows between the main trading partners of the respective countries, with a focus on e-commerce trading, as the data is mostly related to consumer behaviour.

Detected Patterns and Anomalies and Underlying Reasons

This section describes the patterns and anomalies that were detected by the participants. It is structured as follows: for each visualisation type and LOD, the patterns and anomalies that were detected are presented. The first part describes the results for the US data, the second part focuses on the CL data. After each of those sections, there is a short paragraph highlighting the potential reasons for the anomalies (and patterns) that were mentioned by the participants. An overview of the detected patterns and anomalies can be found at the end of this section in tables 4.7 and 4.8 for the US and CL data, respectively.

US data

Patterns and anomalies

With the **choropleth map for LOD 1** (country level, normalised to population), all participants commented on the stability of the flows. Particularly for the incoming flows, they stated that there were only few countries sending items to the US (CA, CN, AU, BR) and that flows e.g. from Europe, RU and Africa were missing, which should not be true based on their prior knowledge. Some also noted a growth in the postal flows from CN to the US, however at least one person mentioned that it was not very clearly detectable. Most participants commented on the big global connectivity of the US, pointing out some of the particularly large flows from the US to e.g. CA, MX, AU, RU, UK, South Africa, South America and in general the English speaking countries. Some noted a general shift in outbound flows from North America (in 2012) to the Asia-Pacific region (in 2015).

Regarding anomalies, the participants found it difficult to detect any in the inbound data. Some pointed out that CZ was only present in 2012, and BG started to appear, but generally they found it very strange that there was not more inbound traffic from Europe. One person mentioned a drop in traffic from CA to the US in calendar week 6 in 2015. As for the outbound data, all participants mentioned the large drop in volumes and destination countries in 2013. Changes in RU in particular caught their attention, because of the large space it covers on the map.

By looking at the **flow map for LOD 1**, most participants mentioned the big stability of the data and the flows, especially pointing out that the large flows to the main postal partners stayed very stable within the six weeks as well as across the years. For one person, the stability of the flows seemed to be very striking, judged by his expression. Some participants mentioned the somewhat less stable situation of the smaller flows, pointing to lines that disappeared and reappeared over time. All participants mentioned a very stable flow to and from CA, as well as a large inbound flow from CN, growing over time. A growing inbound flow was also mentioned for AU, as well as a constantly large outbound flow to AU. The outbound flow to Europe was mentioned by most participants as stable or growing over time, and some of them mentioned the flows to South America and the Caribbean. Some participants pointed out the strong global connectivity for the US when looking at the outbound data. Most participants also mentioned that there was almost no traffic to and from Africa.

All participants noted the anomaly in 2013, stating that there was much less data compared to the other years. Some were very surprised by the few connection in the incoming data, as this was not what they had expected based on their knowledge about postal traffic. Some participants detected an anomaly for the inbound data from CA in calendar week 6 in 2015.

For the **flow map for LOD 2** (IMPC level), the stability of the flows was again mentioned, as well as the growth in the overall traffic. The growth in the inbound flow from CN was perceived as particularly impressive by most participants. As for LOD1, the missing flows from Europe to the US were noted as surprising. The participants found strong connections with CA, AU, Europe and Latin America (BR in particular) and increasing flows to Asia over the years. They also stated that there were almost no flows to and from Africa. It was mentioned by several participants that the flows from CN reached mainly three IMPCs (Chicago (ORD), LAX and JFK). LAX, ORD and JFK were noted to be the most important IMPCs for outbound traffic, whereas one participant noted a particular growth in flows from Miami to AU. It was also noted that there was not one specific hub or group of IMPCs connected to just one region, but that all of the IMPCs were connected to various parts in the world.

All participants again noted a decrease in traffic in 2013 for the inbound and outbound data, however it was perceived as more striking in the outbound data than in the inbound data. Some participants mentioned that Europe was almost missing in the inbound flows. Two participants explicitly mentioned an anomaly in calendar week 6 of 2014 and three of them reported one in calendar weeks 7 and 8 of 2015.

With the **graduated symbol map for LOD 2**, the participants again found that especially the large flows were very stable within the six weeks and across the years, with main flows coming from CA, AU, BR and CN. For the smaller flows (the small dots), they mentioned to see a higher volatility. Some participants pointed out a growth in inbound traffic in 2015 and confirmed this by clicking in order to read the exact values from the dots. For the outbound data, the global connectivity was again mentioned, the participants commented on the fact that a lot of the US IMPCs were active (overlapping circles). One participant mentioned that the distribution was not evenly balanced across the map, showing some very big corridors and smaller ones. As the main destinations of outbound flows CA, GB, CN, AU, MX, BR, RU were mentioned.

All participants mentioned the drop in flows (less dots on the map) in 2013 with a recovery in 2014, some noted that the main hubs remained even in 2013. A couple of participants found this very striking. Some looked at specific regions, e.g. mentioned that there was postal exchange with Africa in 2012 and 2015, but none in 2013/14. One person detected a drop in inbound flows from CA in CW6, 2015.

When examining the data with the **ODM for LOD 2**, the participants immediately noted the limited number of connections in the inbound data. One person called it a big asymmetry with few, but very dark blue cells. As the dominating sending countries, CA and CN were mentioned by all participants, some of them additionally mentioned BR and AU. The participants also noted differences between the IMPCs, stating that there were some IMPCs that receive most of the traffic (JFK, LAX, ORD). The participants all agreed in the point that there were no big changes over the weeks and years, however one person mentioned that ORD seemed to become more important over the years. For the outbound data, again the big global connectivity as well as the stability of flows over time were mentioned. Also, it was pointed out that some IMPCs were hardly used, while others were used very strongly (JFK, LAX, ORD, MIA). A general growth trend was detected (particularly strong for CN), and strong connections to CA, AU, GB, MX, JP, Common Wealth, CN, RU, BR. Some participants reported that it was easy to see the importance of countries and the contributions of the different IMPCs.

All participants immediately noted again the drop in the number of origins in 2013. Some noted that in some weeks, there were more postal partners than in others. One person mentioned that some countries like JP, PO, DE were completely missing as sending countries, which would not correspond with reality. One person mentioned that one IMPC (JEC) completely disappeared in 2013.

When looking at the **time series for LOD 1**, most participants found it very difficult to detect any patterns due to the high volatility in the data. However, they all agreed in seeing an overall growth in volumes over the years, with a particularly large change in volumes between 2013 and 2014. Most participants looked at the numbers to confirm this impression. One participant noted a peak at the beginning of the six weeks period (in all years). For the outbound data across all years, all participants noted a weekly pattern in the data, with additional monthly peaks. Some also mentioned a surge every time after the drop.

Some participants mentioned a strong peak in CW 3 of 2015 for the inbound data. For the outbound data, several participants reported an anomaly in CW 7/8 2015 in form of an unusual drop in volumes. Two participants mentioned lows around 27 and 29 January 2015 (CW 5) as anomaly. Most participants mentioned a significant drop in volumes in 2013, especially for the outbound data. Also in the outbound data, one person noted that the monthly peak appeared only once instead of twice in 2015, and one person mentioned that the first of the large monthly peaks appeared to happen one week later in 2013, compared to the other years.

With the **time series for LOD 2**, the postal experts again mentioned the overall increase in volumes in inbound data, as well as the growth in volatility in the data. They now focused more on the individual IMPCs, stating for example that ORD and JFK were the leaders in receiving items. For the outbound data, the participants again mentioned the weekly pattern in the data as well as the monthly peaks. Most of them also noted that ORD was the IMPCs that handled the biggest amount of data, but other IMPCs were catching up and getting stronger over time (JFK in particular, but also LAX). Some participants also mentioned that HNL was constantly very low.

For the inbound data, most participants stated that it was impossible to detect any anomalies because of the high volatility of the data. One person, however, reported some particularly high peaks for JFK on 17, 20 and 30 January 2014 (CW 3, 4, 5). This person also reported a huge peak for JFK on 13 January (CW 3) and a sudden drop on 27 January 2015 (CW 5). For the outbound data some participants again mentioned the overall drop in volumes in 2013. Two people reported a sudden drop on 13 February 2014 (CW 7) for JFK, another one reported an anomaly on 5 February (CW 6) for JFK. For 31 January 2015, a strong surge following a drop was reported by one person.

Underlying reasons

As reasons for the patterns, i.e. the strong connections between certain countries as well as the overall growth, the postal experts mentioned trade, particularly e-commerce, as the main reason. Further reasons that were mentioned for the patterns include the shared language or migration between countries.

For the weekly patterns in the outbound data, most participants mentioned the opening hours of post offices as reason. For the monthly peaks, it was not so clear. Some assumed one of the reasons could be pay day, leading to more online shopping. The peak at the beginning of the time series for the inbound data was linked to the Christmas peak. The reasons for other large peaks were not clear, some participants assumed special offers in online shopping, etc.

For all participants it was clear that the big anomaly in 2013 must have been a technical problem or a reporting issue of some sort. Two main reasons were mentioned for this assumption. First of all, the flows were consistently low across the whole time period, second some of the connections to other countries seemed to be completely missing, while others stayed at relatively 'normal' numbers.

The reason for the large amount of missing countries in the inbound data was not clear to the experts. It was mentioned that it could be because only tracked items were shown. Others thought it might be because some of the industrialised countries collect their own data about postal traffic but do not share it with the UPU. This could particularly be the case for Europe, it was argued.

Especially for the anomalies, also for the smaller ones mentioned by some participants, they stated that it would be helpful to see the mirror data, e.g. if there was a drop in the inbound flows from CA to the US, it would be interesting to check the flows during this time from CA to other countries.

CL data

Patterns and anomalies

With the **choropleth map for LOD 1**, participants reported an overall growth in the inbound data. As main inbound connections, they saw US, ES, NO, SE, HK, AU, DE and South America. Africa was mentioned as being basically inexistent. One participant stated that one had to be careful about making conclusions, because the overall volumes were very low. For the outbound data, the years 2012 and 2014 were perceived as very steady in terms of flows. As main outbound connections, the participants mentioned the US, ES, DE, SE, CA, NO, AU, PE, UR.

All participants reported an anomaly regarding extremely low traffic in CW 32/33/34 2013, growing again in CW 35 and showing the highest values in CW 36. One person stated that this was better visible in the outbound than in the inbound data.

When analysing the data with the **flow map for LOD 1**, all participants noted a dominant inbound flow from HK, but most of them also mentioned that there were plenty of connections with other countries all over the world. Another main stream of inbound data was reported from Europe, the US was also mentioned by some participants. For the outbound data, one person states that the main corridors

TABLE 4.7: Detected patterns and anomalies in the US data

US **CW 3-8, 2012-15**	**Patterns**	**Anomalies**
Choropleth map **LOD1** interactive	• Stable flows from CA, CN, AU, BR, growing from CN • Flows from Europe, RU, Africa constantly missing • Big global connectivity with a shift from North America to Asia-Pacific • Large flows to CA, MX, AU, RU, UK, S-Africa, S-America	• Inbound flows from CZ only in 2012 • Inbound flows from BG appearing • Drop in traffic from CA in CW 6, 2015 • Very low outbound traffic in 2013
Flow map **LOD1** static	• Strong connections with CA, CN, AU • High global connectivity • Growth in traffic over time	• Missing data in 2013 • Few connections in inbound data • Less data from CA in CW 6, 2015
Flow map **LOD2** static	• Stable and growing inbound flows particularly from China • Strong connections with CA, AU, Europe, BR, Asia • Most outbound flows from LAX, ORD (Chicago), JFK • No flows to and from Africa • Flows from CN mostly to LAX, SFO, JFK	• Missing inbound flows from Europe • Lower inbound flows in CW 6 in 2014 • General low flows in 2013 • Lower inbound flows in CW 7/8 in 2015
Graduated symbol map LOD2 interactive	• Stable inbound flows, mainly from CA, AU, BR, CN • Overall high global connectivity with all of the US IMPCs active • Higher volatility in smaller flows • Growth in traffic in 2015	• General low flows in 2013 • No traffic to/from Africa in 2013/14 • Drop in traffic from CA in CW 6, 2015
ODM **LOD2** static	• Main inbound flows from CA, CN, AU, BR • Limited number of inbound connections, but stable • Main flows to CA, AU, GB, MX, JP, Common Wealth, CN, RU, BR • Main flows into LAX, ORD, JFK • High global connectivity • Some outbound IMPCs used very strongly (JFK, LAX, ORD, MIA)	• Missing inbound flows from JP, PO, DE • Some weeks have fewer partners than others • Fewer origins in 2013
Time series **LOD1** interactive	• Generally extremely volatile inbound data, but growth over the years • Weekly drops and monthly peaks • Surges after weekly drops	• Drop in volumes in 2013 and in CW 5 and CW 7/8 of 2015 • 'Late' first monthly peak in 2013, missing second monthly peak in 2015 • Strong peak in CW 3 of 2015
Time series **LOD2** interactive	• Extremely volatile inbound data, but growth over the years • ORD and JFK handle most inbound flows • Weekly pattern and monthly peaks in outbound data • ORD handles most outbound flows but JFK (and LAX) catching up	• Inbound: high peaks for JFK in CW 3/4/5 in 2014 • Inbound: high peaks for JFK CW 3 and a sudden drop in CW 5 2015 • General low volumes in 2013 • Outbound: drop in CW 6 and 7 in 2014

were varying over time and space because of volatility in the data. However, most participants reported traffic to SE, CA, NZ, AU and Europe. Almost all participants were surprised to see NZ, since they did not notice NZ in other visualisations. UR was further mentioned to be appearing as a new route in 2013.

For the inbound data, an anomaly was mentioned by most participants in CW 35 2013, as the flows from HK disappeared. Bermuda was also mentioned by two participants as suddenly appearing with a rather thick line in CW 31 2012. Most participants further noted a thick line for IL in CW 36 2013 in the inbound data. For the outbound data, all participants immediately reported an anomaly with a dramatic drop in CW 32/33/34 2013 and a rise in values again in CW 35 reaching the highest amounts across all years and weeks in CW 36.

For the **flow map for LOD 2**, many of the participants mentioned that there was a high volatility, especially for the inbound data. Still, they stated that the main sending countries (US, DE, ES, (HK)) stayed rather constant over time, whereas Africa was missing completely. One person mentioned that the strength of the US diminished somewhat because the flows were split between the IMPCs. Most participants further mentioned an increase of items sent from Miami (MIA) and DE in 2014. For the outbound data, 2012 and 2014 were seen as stable and very similar, with an increase of volume over time. The US and Europe were mentioned as main outbound connections.

Regarding anomalies, CW 31 2013 was mentioned by many of the participants with MIA appearing as a very strong connection. CW 32/33/34 2013 stroke all participants as they mentioned that there was almost no traffic, especially in the outbound data. A surge in the data was noted in CW 36 2013.

With the **graduated symbol map for LOD 2**, most participants reported an increase in inbound volumes over time. US (especially MIA) and Europe were mentioned as the main partners for inbound traffic by all the participants. Africa was again mentioned by some participants as mostly absent. For the outbound data, the participants reported an increase in the number and the size of the circles. Again, they mentioned that the flows to the main partners (MIA, Europe (ES), South America (UR)) stayed rather constant over time, with an overall growth trend, particularly in the connection with MIA.

All participants reported an anomaly in 2013, with a large decrease in volumes in CW 32/33/34, an increase again in CW 35 and a surge in CW 36, which was particularly high for inbound traffic from LAX. The focus on the anomaly in the outbound data lay on the US, especially MIA. One person stated that in CW36 2013, there was an extreme increase in outbound traffic to the US (volume) and Europe (number of destinations).

With the **time series for LOD 1**, all participants stated that it was extremely difficult to see any patterns in the inbound data, as it appeared to be very volatile. It was mentioned that overall the values were very small, which made it difficult to draw conclusions. In the outbound data, some participants found that a slight seasonality might be visible, but others stated that no patterns were visible because there was too much variance in the data. It was mentioned by most participants that 2012 and 2014 looked very similar, with an overall increase of volume between 2012 and 2014 and more volatility in 2014 than in 2012.

Regarding anomalies, all participants reported 2013 as a 'crazy year' where something must have happened. Two participants specifically mentioned a drop in the inbound data to almost 0 between 5-10 August 2013, one person reported the 7-24 August 2013. One participant compared the number of peaks in the inbound data, stating that in 2013, the 1000 mark was only hit once, whereas it was hit four times in 2014. For the outbound data, very high peaks were reported following a period with a very low number of items.

Underlying reasons

It was mentioned that the major e-commerce hubs are located in the US and CN, which could be one of the reasons for strong connections with those countries, especially for the inbound data. The strong links to the Americas in the outbound data was attributed to a regional effect, whereas the strong connection with Europe was attributed to cultural ties. It was concluded by at least one participant that regional and cultural ties seemed to be more important than just trade. The reason for the strong connection with SE was assumed to have something to do with the Chilean diaspora there, which is supposed to be rather large. The particular increase in MIA was linked to e-commerce, as Correos de Chile has a virtual address in Florida that will collect and send goods to CL that can only be ordered to a US address.

Regarding the large anomaly in 2013, some stated that it must be an error in the data. Others thought that something must have happened that blocked the infrastructure. One person reasoned that because it lasted quite a long time, it could have been a strike or a natural disaster that destroyed infrastructure. Another participant mostly focused on the clearly visible catch up at the end of the period. This person argued that if it was a pure reporting problem, no catch up would have been visible (the data simply would have gone back to normal once the problem was fixed). As the other person, he took into account the duration of the event and therefore concluded that most likely it had been a strike.

4.7 Concluding Remarks

Preferences regarding visualisation types varied rather strongly between the postal experts who participated in the usability study. Measures like the system usability

TABLE 4.8: Detected patterns and anomalies in the CL data

CL **CW 31-36, 2012-14**	**Patterns**	**Anomalies**
Choropleth map **LOD1** interactive	• 2012 and 2014 very steady, overall growth in flows • Main inbound connections: US, ES, NO, SE, HK, AU, DE, South America • Main outbound connections: US, ES, DE, SE, CA, NO, AU, PE, UR • Africa basically inexistent	• Extremely low traffic in CW 32/33/34 in 2013, back in CW 35, highest in CW 36
Flow map **LOD1** static	• Dominating inbound flow from HK, but also from Europe and US • Main outbound traffic to SE, CA, NZ, AU and Europe • New route to UR appearing in 2013	• Inbound flows from HK disappear in CW 35 2013 • Bermuda appearing in CW 31 2012 and IL in CW 36 2013 • Drop in CW 32/33/34 2013, rise in CW 35, peak in CW 36
Flow map **LOD2** static	• High volatility in inbound data. • Main inbound connections: US, DE, ES, (HK) • Main outbound flows to the US and Europe • 2012 and 2014 stable outbound flows, with growing volumes	• Large increase in connection with Miami • Almost no traffic in CW 32/33/34 2013 • Surge in CW 36 2013
Graduated symbol map LOD2 interactive	• US and Europe main inbound connections • Africa mostly absent • 2012 and 2014 stable outbound flows, overall growth trend • MIA, Europe, South America main outbound connections	• Huge problem in CW 32/33/34 2013 • Catch up in CW 35/36 2013
Time series **LOD1** interactive	• Overall increase of traffic, very volatile data • 2012 and 2014 very similar	• Very low inbound traffic in August 2013 • Large peaks following very low number of items

score should therefore be treated with caution, as there was an outlier for the majority of the evaluated visualisation types (see figure 4.8). Since there were only six participants, it is recommended that the tests are run with more participants to get a more comprehensive picture.

Some visualisation types helped the participants better than others detect certain types of patterns and anomalies. This section summarises the main patterns and anomalies that were discovered and table 4.9 provides an overview of the patterns and anomalies that were reported during the exploration of the data with the different visualisation types.

The **connections to the main postal partners** were described with all of the visualisation types except the time series, for both the US and the CL data set. Comments about the **global connectivity** of both the US and CL were made with all of the visualisation types that included geographic space, i.e. using a map to display the data. For CL, only one person made a remark regarding global connectivity by using the graduated symbol map. One person commented on global connectivity when using the ODM for the US data. Global connectivity can mean the amount of connections of the US or CL to the whole world or also notions about missing connections from certain regions, e.g. Africa. The **overall growth trend** in the US data was reported when working with the flow map, the graduated symbol map, the ODM (only one person), the time series for LOD 1 and the time series for LOD 2 (only one person). For the CL data, it was reported with all of the available visualisation types (only one person for the flow map LOD 1). Some participants even specified between growth in volume and growth in the number of origins or destinations. A **growth trend in specific regions, countries, IMPCs** for the US data was detected with the choropleth map, the flow map (LOD 1 and 2) and the ODM (only one person), as well as the time series for LOD 2. For the CL data, such patterns were reported when using the choropleth map (one person only), the flow map, and the graduated symbol map. The **weekly and monthly patterns** in the US outbound data were reported by using the time series (LOD 1 and 2) for analysing the data.

As for the anomalies, the following ones were reported by analysing the data with the different visualisation types. For the US data, the **technical issue in 2013** was detected with all of the visualisation types. The **snow storms in CW 4/6 2014** were only reported with the choropleth map (only one person), the flow map (LOD 2) and the time series (LOD 2). Similarly, the **technical issue in CW 7/8 2015** was reported by using the choropleth map (one person), the flow map (LOD 2) or the time series (LOD 1, one person). The **low traffic from CA in CW6 2015** was detected when using the choropleth map (one person), the flow map (LOD 1) and the graduated symbol map (one person). A few of the participants mentioned the appearance and disappearance of BG and CZ when analysing the data with the choropleth map. **Peaks in CW 3 and low flows in CW 5 2015** were mentioned by various participants when analysing the data with the time series (LOD 1 and 2) and by one person when looking at the flow map (LOD 2).

The **strike** in CL in form of a drop in postal flows (CW32–34 2013), a backlog and a catch up (CW35/36 2013) was reported with all of the available visualisation types. Additionally, irregular flows in the CL data set to or from **Israel, Bermuda and Polynesia** were mentioned by some participants when analysing the data with the flow map (LOD 1).

TABLE 4.9: Reported patterns and anomalies and the visualisation types with which they were detected. If only one participant reported a certain pattern or anomaly, this is marked with (1 p.)

Pattern US	**Visualisation type**
Connections to main partners	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map, ODM
Global connectivity	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map , ODM (1 p.)
Overall growth trend	Flow map LOD 1, Flow map LOD 2, Graduated symbol map, ODM (1 p.), Time series LOD 1, Time series LOD 2 (1 p.)
Growth trend of specific regions, countries, IMPCs	Choropleth map, Flow map LOD 1, Flow map LOD 2, ODM (1 p.), Time series LOD 2
Weekly/monthly pattern	Time series LOD 1, Time series LOD 2
Anomaly US (previously known or reported by the participants)	
Technical issue 2013	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map, ODM, Time series LOD 1, Time series LOD 2
Snow storms CW 4/6, 2014	Choropleth map (1 p.), Flow map LOD 2, Time series LOD 2
Technical issue CW 7/8, 2015	Choropleth map (1 p.), Flow map LOD 2, Time series LOD 1 (1 p.)
Less traffic from Canada CW6, 2015	Choropleth map (1 p.), Flow map LOD 1, Graduated symbol map (1 p.)
BG and CZ appearing/ disappearing	Choropleth map
Peak in CW 3, 2015 drop in CW 5, 2015	Flow map LOD 2 (1 p.), Time Series LOD 1, Time series LOD 2

Pattern CL	
Connections to main partners	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map
Global connectivity	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map (1 p.)
Overall growth trend	Choropleth map, Flow map LOD 1, Flow map LOD 2 (1 p.), Graduated symbol map, Time series LOD 1
Growth trend of specific regions, countries, IMPCs	Choropleth map (1 p.), Flow map LOD 2, Graduated symbol map
Anomaly CL (previously known or reported by the participants)	
Strike in CW 32-34 2013	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map, Time series LOD 1
Backlog/catch up in CW 35/36 2013	Choropleth map, Flow map LOD 1, Flow map LOD 2, Graduated symbol map, Time series LOD 1
IL, Bermuda, Polynesia appearing/disappearing	Flow map LOD 1

Chapter 5

Discussion

The visual representation of large data sets is a crucial part in the field of visual analytics, as it helps analysts to investigate the data and extract relevant information from it and thus obtain knowledge (Andrienko and Andrienko, 2010). This thesis contributes to the discussion in the field of geovisualisation and visual analytics in general by analysing various visualisation types regarding their applicability for visualising global postal OD data for the visual discovery of patterns and anomalies. It further contributes by conducting a user study with a real-life use case to evaluate the usability of selected visualisation types for practitioners (postal experts, in this particular case). The study evaluated five different visualisation types: the choropleth map, flow map, graduated symbol map, ODM, and time series. The first four sections of this chapter summarise and discuss the main findings of this study with respect to the research questions and hypotheses as well as previous literature. The fifth section discusses unexpected results while the sixth section synthesises the findings and provides recommendations for postal experts regarding the visual representation of their data. The seventh section briefly discusses the limitations of this study.

5.1 Influence of Spatial Properties and their Encoding

The first research question of the thesis aimed at assessing how spatial properties of the postal network and the encoding of origins and destinations affect analysts' ability to visually detect and analyse patterns and anomalies in the postal OD data. The research question was addressed by investigating the aspects of the *encoding* of origins and destinations, *connectivity* within the network as well as *distance* between origins and destinations.

There is a trade-off between the generalisation for better visual perception and information loss

The *encoding* of origins and destinations into two different LODs (country level and IMPC level) showed that the participants could extract different degrees of information content, depending on the LOD they were using. LOD 1 provided them with generalised information for a broad overview, for example about the main connections between countries and the overall connectivity of the US or CL, respectively. LOD 2 allowed them getting a more differentiated view on the data, they could for example find out about the dynamics and relative changes in flow volumes to or from individual IMPCs. This was for example observed for the US data, where the experts noted in the time series visualisation that the IMPC USORD was clearly dominating the traffic, but USJFK was catching up later (see figure 5.1). Further, LOD 2 enabled them to detect shifts in the routing of shipments, e.g. if some of the IMPCs disappeared or suddenly started to handle a lot more items. The visualisation type influenced which data characteristics were mainly noted: the ODM was helpful for the participants to note the disappearance of certain IMPCs, and in general the spatial visualisation types (geographical and non-geographical) helped the participants to make remarks about the connectivity of specific IMPCs as well as the overall activity of all IMPCs. Growing relations between certain IMPCs and specific countries could be detected well with the flow map and the graduated symbol map.

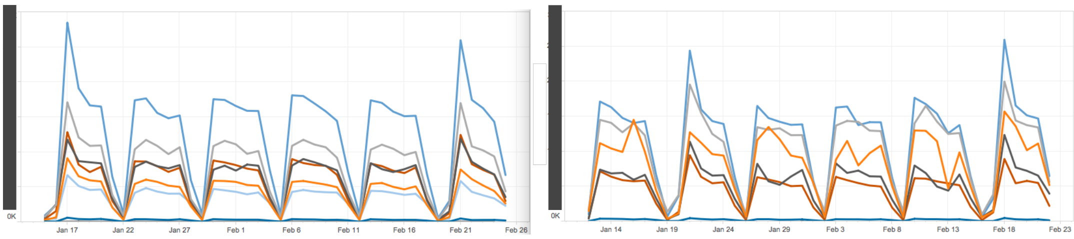


FIGURE 5.1: Time series visualisation for LOD 2 of US outbound data, 2012 (left) and 2014 (right). Participants mentioned a 'catch up' of the IMPCs USLAX (grey) and USJFK (orange) relative to USORD (blue).

It is difficult to visualise the large number of origins and destinations in LOD 1, but even more challenging to visualise the flows in the more detailed LOD 2 (for the US data, a threshold was set because of visual clutter). The findings in this regard are in line with previous literature and also apply to large OD data sets and networks in other fields (Herman, Melancon, and Marshall, 2000; Wong et al., 2006). In accordance with other studies (Thomas and Cook, 2005; Andrienko and Andrienko, 2010), there is a trade-off between aggregation for easier depiction and perception of data and information loss, or between providing a detailed-view and visual clutter and information overload. The assumption that spatial properties such as granularity of the spatial units have an impact on the visualisation and detection of patterns and anomalies is thus supported by the findings (Hypothesis 1a).

The structure and the encoding of the data are linked to the choice of tools for creating visual representations of the data. Choosing an adequate tool for creating a data visualisation is an important part of the process for analysts. Results of the tests with several tools show that there are differences regarding the flexibility of a tool, the technical knowledge needed and the level of customisation that is provided by the respective tool. It is thus important for the analyst to assess the different possibilities that are offered by the tools with regard to his or her requirements, resources (e.g. technical knowledge) and the data structure. If the users are not familiar with programming, for example, then an existing tool is certainly the better solution for them than a tool where they have to do some coding to produce a visualisation.

The level of connectivity influences the visualisation and the analysis of the data

In general, high *connectivity* is an inherent characteristic in global networks like the postal OD data, as such networks connect locations across the globe through their consistent flow of people, goods, etc. In the global postal network, as most likely also in many other global networks, there are differences in the level of connectivity between the individual countries (or other entities representing the nodes). Thus, to address the aspect of connectivity, two case studies with different levels of global connectivity were chosen in this study (see 3.3). Findings show that the level of global connectivity has consequences for visualising and analysing the data (Hypothesis 1b).

The results show that data with a high global connectivity are more difficult to visualise and run the risk of creating information overload, especially in visualisations with overlapping features such as flow maps or graduated symbol maps. However, this is also true in the case of ODM, where a high connectivity leads to many rows or columns (Andrienko and Andrienko, 2010; Wood, Dykes, and Slingsby, 2010). In contrast, data sets with a low global connectivity are easier to visualise because visual clutter is less of a problem (see for example figures 5.2 and 5.3).

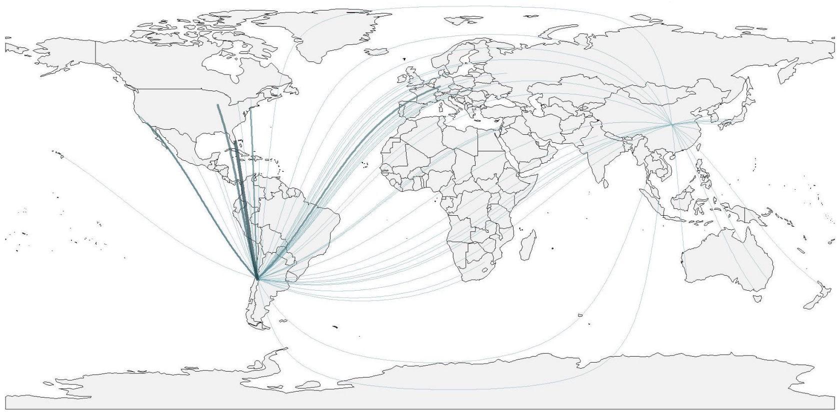


FIGURE 5.2: Flow map of CL inbound flows, CW35 2014. Low connectivity compared to the US data.

In countries with a high global connectivity, there are also often large flow volumes, whereas flow volumes are lower in countries with lower global connectivity. Results show that large volumes allow a better analysis of the patterns and anomalies, because the flows are overall more stable and, hence, large changes are noticed

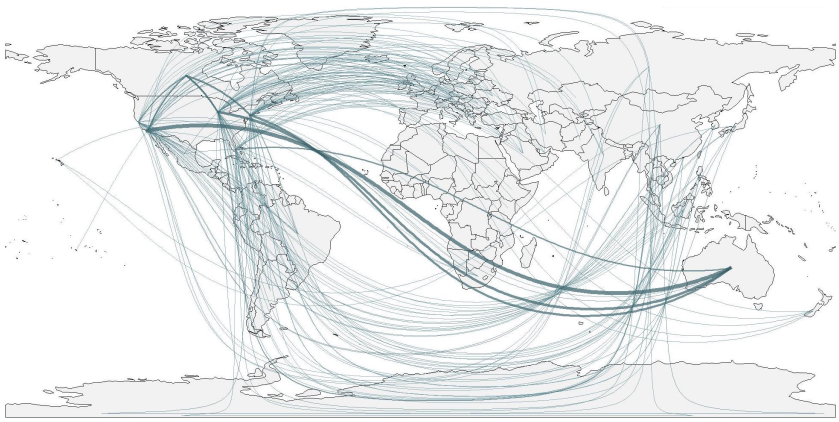


FIGURE 5.3: Flow map of US outbound flows, CW4 2015. High connectivity, but more difficult to visualise because of visual clutter.

rather quickly. In the case of low overall volumes in flows, sudden changes in volumes can be related to the natural volatility in the data rather than to actual disruptions. Hence, even if the visualisation might be more challenging in one case, this does not necessarily mean that this has an overall negative effect on the detection of patterns and anomalies. This finding can have implications for other network visualisations, as it is assumed that this difference in the level of connectivity and its association with varying flow volumes can also be found in other networks.

Spatial correlation exists in postal OD data, but other factors are also relevant

As a third aspect, the impact of *distance* between origins and destinations was analysed. The assumption was that spatial autocorrelation exists in the postal traffic data, as it is found in many other geographic flows (Tobler, 1970; Wood, Dykes, and Slingsby, 2010).

Results show that there is spatial correlation in the postal data, as can be seen in the regional ties that were reported by the participants during the usability study (e.g. strong connections between CL and other countries in South America or between the US and CA). However, there are also many and strong connections between very distant places, hence, Hypothesis 1c can only partly be supported. From the comments made during the usability study as well as the interview sections, it can be seen that space is most likely not the strongest element for defining the postal flows. Cultural ties (language, historical connections, migration and diasporas) are also relevant (e.g. connections between the US and English speaking countries or between CL and Spain or Sweden (diaspora)) and especially trade is important (flows between the US and CN and between CL and HK). These findings are in line with the previous study by Hristova et al. (2016). Some trade agreements correlate with proximity, e.g. in the case of NAFTA[1], thus it would be interesting to see if there was such a strong postal connection even without a trade agreement. A stronger focus

[1] North American Free Trade Agreement

on the temporal dimension, i.e. the changes over time regarding the connections between countries could be very interesting for further investigation. Additionally, it would be worth analysing developments over time in other networks such as migration, flight or trade to assess and compare the changes between those different networks.

5.2 Applicability of Visualisation Types

It was assumed that visualisation types which have been used for analysing other flow networks are also applicable in the analysis of postal OD data. Thus, the second research question aimed at evaluating the selected visualisation methods by applying them to the postal OD data and conducting a usability study with postal experts. An exploratory approach was applied in order to evaluate several visualisation types regarding their potential and limitations as well as their complementarity.

Potential and limitations of the implemented visualisation types

The advantages and limitations of the applied visualisation types were assessed in order to identify the visualisations that perform best at the tasks of visually discovering patterns and anomalies in the postal OD data. The following sections discuss the potential and limitations of each of the evaluated visualisation types.

Choropleth map

Most of the participants liked to work with the choropleth map and performed well with it during the tasks, especially for detecting the major flows. However, they found it difficult to detect an overall growth trend in the data with this visualisation type. Because of the skewness of the data, a classification based on natural breaks would be the better choice for this representation (Guo et al., 2009), however having separate class limits for each map (each CW) would make it impossible to compare changes between the CWs. With the classification that was chosen in this study, some classes contain a large range of values. This makes it very difficult to visualise changes, since the values will still fall into the same class unless the changes are very big. Apart from easily detecting the major connections between postal partners, the participants also made comments about the general global connectivity of the two countries used in the case studies.

Overall, participants performed worse with the choropleth map in detecting anomalies than they did in detecting patterns. This is especially true for smaller anomalies like a disruption that only lasted one or two weeks. It is assumed that this lack in detection of anomalies is largely due to the fact that was mentioned above regarding classification schemes. Further, results show that participants strongly focused on specific countries during the tasks. Especially countries with a large area (CA, AU, US, BR, RU) are very dominant in this type of visualisation and thus caught their attention (see figure 5.4). Therefore, they almost only noticed changes or anomalies if those could be seen in the large countries. This focus on specific countries, along with the issue of classification scheme, also explains that they reported a growth trend in certain countries, but not the overall growth trend. In contrast to other visualisation types, there are no overlapping features in this representation because of the administrative units. This helped the participants notice the sudden appearance and disappearance of BG and CZ (see for example figure 5.5), which was not reported with any other visualisation type (cf. existential changes in spatio-temporal data, as defined by Andrienko, Andrienko, and Gatalsky (2003)).

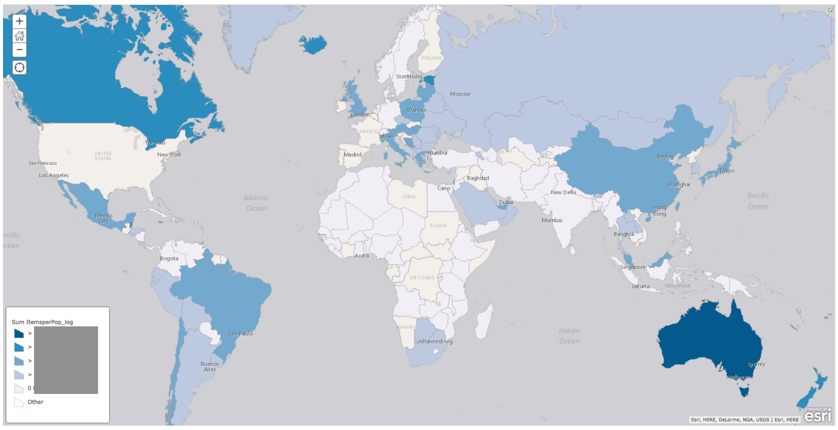


FIGURE 5.4: Choropleth map for US outbound data, CW4 2015. Countries with a large area are very dominant and catch the user's attention.

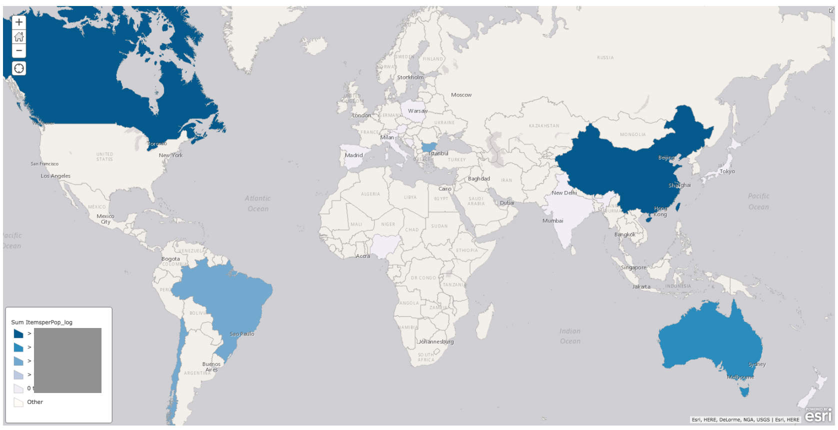


FIGURE 5.5: Choropleth map for US inbound data, CW4 2014. Surprisingly, there is almost no information about tracked items sent from Europe to the US. Bulgaria is thus standing out.

In accordance with the studies by Sander et al. (2014) and Kelly et al. (2012), it is questionable how well choropleth map visualisations would perform with a more comprehensive data set, i.e. for showing flows between several countries instead of focusing on flows to or from one country. In order to overcome this issue, a filter mechanism could be installed to enable selecting and displaying one-to many or many-to-one flows for one country at a time. This aspect of Hypothesis 2a can thus not be fully supported by this study, further tests would be necessary. As can be seen from the results, this visualisation might hinder analysts to see even larger changes in the number of items because they could still fall into the same class. Thus, although connectivity related issues might be easily detectable, traffic trends and seasonal effects are more difficult to discover.

Flow map

Overall, participants performed well with the flow map in both detecting patterns and anomalies in the data. Results from the interview section show that the participants find flow maps intuitive, which is in accordance with Yang et al. (2016). Regarding the individual preference of the participants, the flow map visualisation did not rank particularly high. Part of this can be attributed to the way it was presented rather than to the actual visualisation type. This was confirmed during the interview sections, where some of the participants stated that they think the flow map would be more useful for them if they did not have to switch between the representations of the single CWs, but could see them side by side on the same screen.

As opposed to the choropleth map, the flow map enabled the participants to detect flows between very distant as well as very small places (e.g. connections between CL and NZ or connections to small islands). As was seen above, the participants focused on large countries with the choropleth map, whereas this was not the case for flow maps.

FIGURE 5.6: Flow map for CL outbound data, CW32 2014 (normalised to populations size of destination country). Flows to NZ become visible, whereas participants did not see them with other visualisation types.

The flow map visualisation (for both LODs) helped the participants to make a quick assessment of the main flows as well as commenting on the global connectivity of the two analysed data sets, which is in accordance with Wood, Slingsby, and Dykes (2011) and supports Hypothesis 2a. For example, the participants noted changes in the number of flows that were displayed (cf. existential changes mentioned by Andrienko, Andrienko, and Gatalsky (2003)). For LOD 2, the participants reported on the overall activity of the different IMPCs.

However, the flow map visualisation also has its drawbacks. Results show that it was difficult for the participants to distinguish the lines in regions with many flows as well as many small countries (e.g. Europe). They could thus report that there were many flows to Europe for example, but in contrast to the exploration with the choropleth map, it was impossible to say to which countries exactly. The problem of overlapping information and visual clutter is a known issue in previous research about spatio-temporal data visualisation (Wood, Dykes, and Slingsby, 2010; Andrienko and Andrienko, 2010; Cui et al., 2008).

Cui et al. (2008) show a promising approach in dealing with visual clutter by applying their edge-clustering method (see also 2.1). However, the flow map could also be implemented in an interactive way, hence a specific flow of interest and its corresponding origin and destination could be highlighted. Additional information about the volume of the selected flow could be displayed on the screen. For LOD 2, colouring the flows according to their origin or destination IMPC would be a possibility, however this would add yet another piece of information to the map and thus making it more complex. Whether such approaches would help analysts to better explore the data or rather hinder them would again have to be tested in user studies, but would be worth investigating.

OD matrix

The results show that ODM visualisations helped the participants detect overall patterns of flows between the US IMPCs and countries. Results from observations during the usability show that the colouring of the cells was particularly helpful to quickly note the major connections between origins and destinations. The displayed values in the cells helped to confirm those first observations. Large changes in the number of columns or rows were quickly noted, either as anomaly or existential change (Andrienko, Andrienko, and Gatalsky, 2003), for example if one IMPC disappeared completely (see figures 5.7 and 5.8).

In contrast to the choropleth map or flow map visualisations, the participants generally reported flows to or from specific countries (CA, JP, etc.), but never mentioned flows or missing flows to or from certain larger regions like Africa, Europe or Asia-Pacific. Thus, the ODM visualisation, as it was presented in this study, is not very useful for detecting regional patterns. This can be attributed to the lack of geographic information, which was also indicated during the interview sessions.

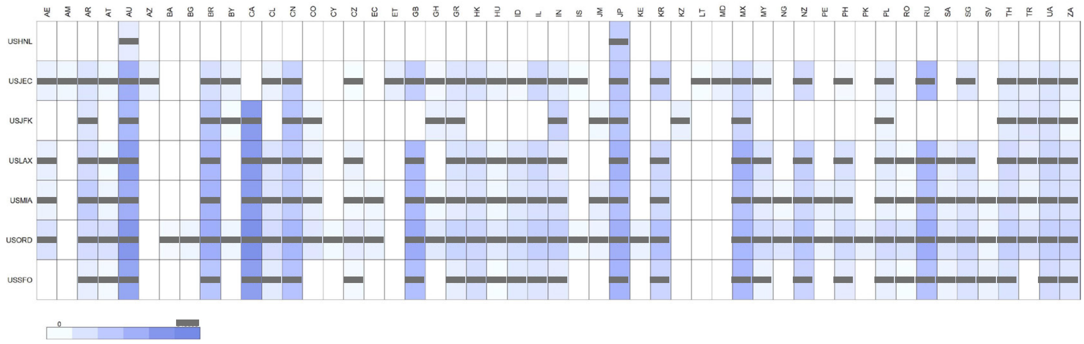


FIGURE 5.7: ODM for US outbound data, CW 7 2012. Rows = origins, columns = destinations. The IMPC USJEC (second row from top) disappeared after 2012.

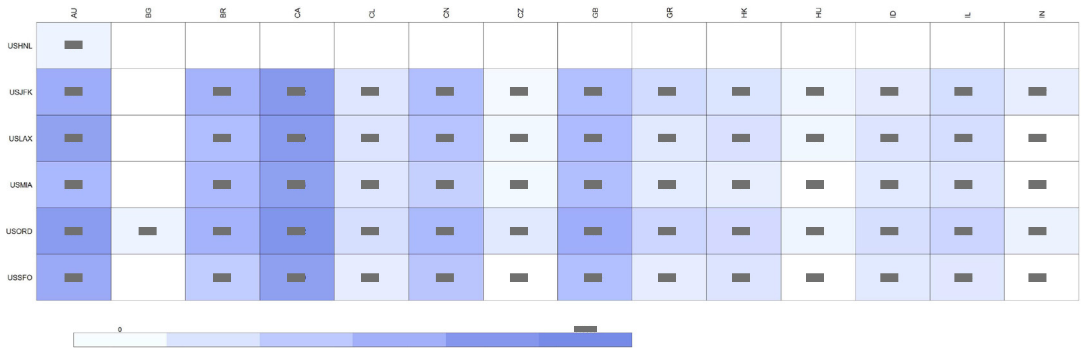


FIGURE 5.8: ODM for US outbound data, CW 7 2013. Rows = origins, columns = destinations. There are clearly fewer destinations than in the previous year (figure 5.7). The IMPC USJEC was only present in 2012, but not in any of the following three years.

Results further show that the large amount of origins and destinations is challenging for ODM visualisations. In contrast to flow maps, where overlapping information was the main problem, the large number of rows and columns are the main issue in the case of ODMs. This could partly be overcome by grouping the countries, for example according to their respective world region (Guo et al., 2006). One ODM per region could be created in order to allow an analysis of this particular data set. Other possibilities include interactive implementations, including a filter function to select the information to display.

Missing values were also found to be an issue, as this resulted in empty columns or cells, which were not displayed. This resulted in a changing order of columns and rows between the ODMs of the CWs, which complicated comparing changes between the CWs. This problem could potentially also be solved by showing all rows and columns, including empty ones, if the countries were grouped to regions and thus one ODM had to display a smaller number of rows or columns. As participants generally perceived this visualisation type as being helpful but difficult to read because of the above mentioned issues, further tests with adjusted implementations are recommended to address the reported problems.

The related method of OD maps proposed in previous studies (Slingsby et al., 2012; Wood, Dykes, and Slingsby, 2010) was not further pursued or included in the usability study, because results from the requirements analysis show that OD maps were too complicated and thus not practical for this target group. Additionally, whereas they were implemented on a national scale in previous studies, global data is thought to be rather difficult to visualise with this approach.

Graduated symbol map

With the graduated symbol map, the participants were able to quickly assess the spatial distribution of origins and destinations. Results show that this was largely due to the strong saliency of the dots, which helped the participants to note changes, especially existential changes and location changes (cf. Andrienko, Andrienko, and Gatalsky, 2003). Changes in volume were also noted, e.g. the overall growth in traffic, however overlapping circles made this notion more difficult for participants. The overlapping of dots generally made it more challenging and time consuming for participants to explore the data (see figure 5.9), as this depiction seemed to be confusing for some of them. Large anomalies in the data were noted quite well with the graduated symbol map, due to striking changes in the number of dots and their size during the anomaly in contrast to the other CWs (see for example figure 5.10. However, smaller anomalies (snow storms in the US) were not reported by any of the participants with this visualisation type.

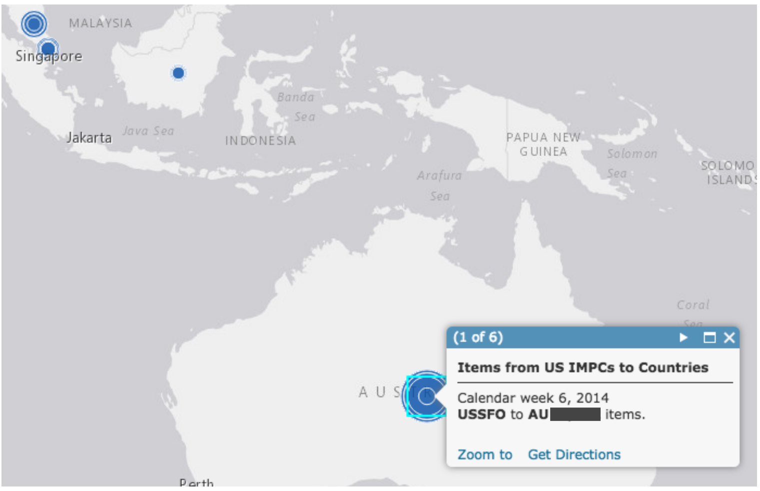


FIGURE 5.9: Graduated symbol map for US outbound data, detail view for Australia. Users need to switch back and forth between the different IMPCs to see their respective number of items sent to Australia.

Interestingly, most of the participants seemed to like this visualisation type, even though many of them did not bother to interact with it a lot in order to investigate

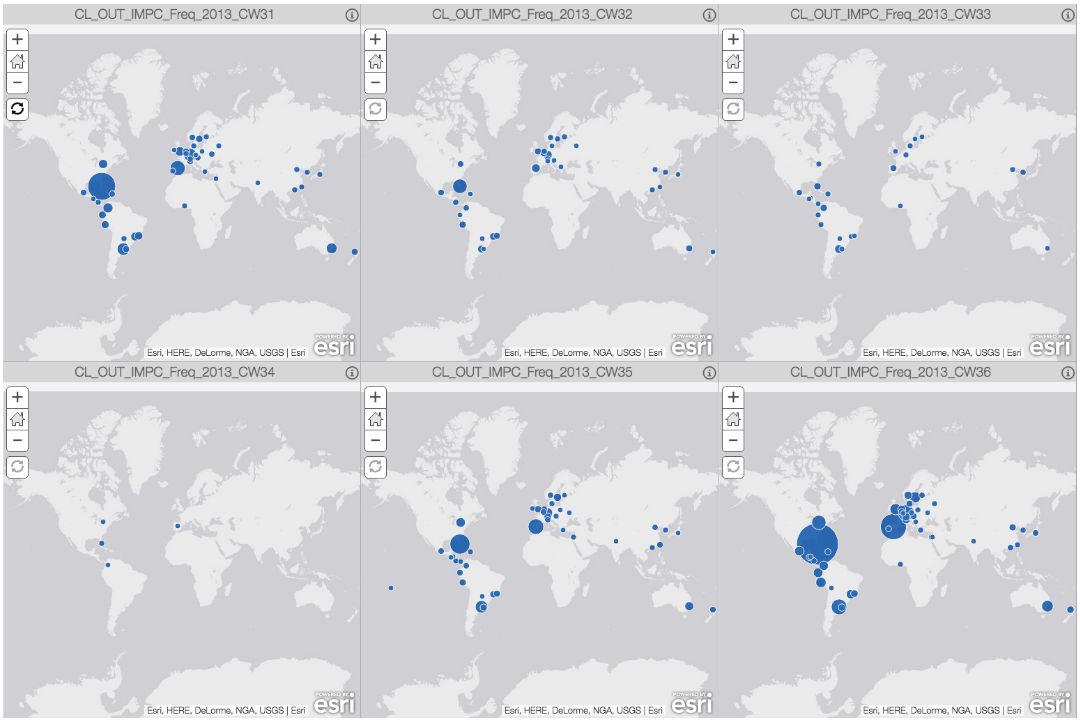


FIGURE 5.10: Graduated symbol map for CL outbound data, CW31–36, 2013. The drop in destinations and volumes during the strike could easily be detected by the participants.

the data further. Perhaps a different form of grouping the circles and a possibility to dis-aggregate them when zooming in more closely could help the participants overcome the difficulties they encountered. Using colours to discern flows to or from a specific IMPC or country could help to provide the analyst with more information. Using different colours for large and small volumes, as seen in the study by Rae (2009), could further help to highlight the main flow connections. However, because of the nature of the data at hand, this might result in a more confusing depiction because of the high number of origins and destinations. Thus, the suitability of such an approach would have to be investigated further. Additional possibilities like filtering could further help to analyse specific flows. Such implementations would require further testing.

Time series

Results show that all of the participants liked the time series visualisation for exploring the data, in spite of the fact that it was presented to them on a smaller computer screen than the other visualisations. Many of them liked it because they were familiar with this visualisation type, as was stated in the interview sessions. They were able to detect an overall growth in volumes for both data sets as well as large anomalies such as the low values in 2013 in the US and the strike during a few weeks in CL. Results further show that they reported weekly and monthly patterns in the US data (see figure 5.1). However, no patterns were reported for the CL data, due to the high volatility in the data compared to the overall values (see for example figure 5.11). The high volatility was noted for both data sets and it was mentioned that the volatility seemed to grow with growing overall volumes, particularly in the US. The volatility in the data was not mentioned as much with any of the other visualisation types, which can at least partly be attributed to the difference in temporal aggregation. Participants attributed this development to the growth in e-commerce and thus people's shopping behaviour. The interactive features (extracting exact values when hovering over the line or filtering of specific IMPCs) helped the participants to study the data in more depth, which is in line with previous studies (Andrienko and Andrienko, 2005). Showing the time series graphs of the different years beside each other was seen as helpful by the participants because it enabled them to compare the lines directly.

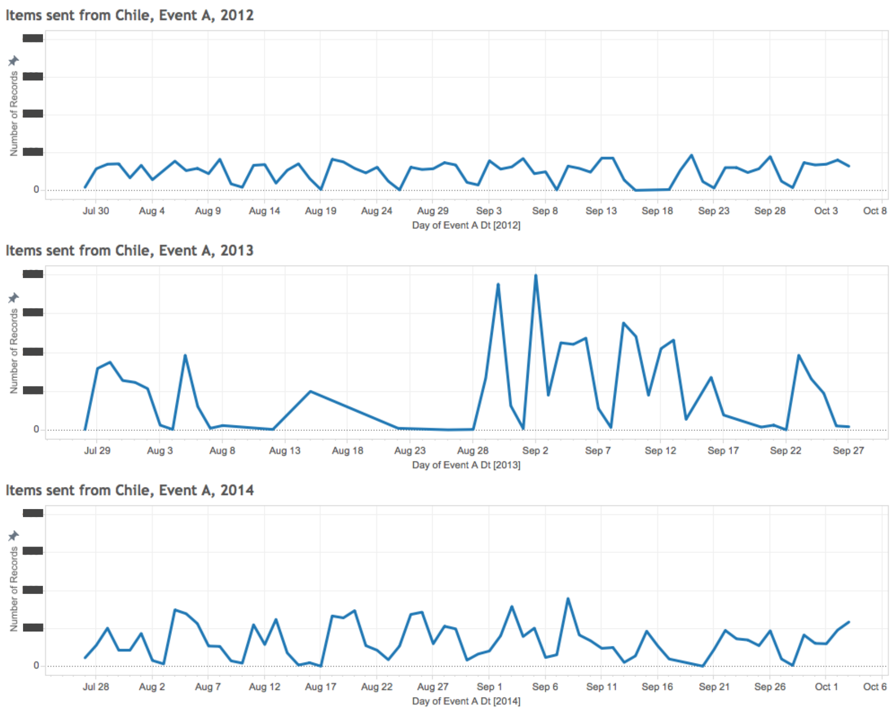


FIGURE 5.11: Time series visualisation for CL outbound data for 2012–2014. The participants could not detect any patterns in the data, but many reported that there was an anomaly in 2013. Most participants also mentioned the volatility in volumes, especially for 2013.

The results show that the biggest drawback of the time series visualisation for the study's tasks was the missing geographic reference. Participants could not comment on a country's overall connectivity or on connections to its main partners. Thus, a dynamic link to a different visualisation type such as a map would be necessary in order to draw any conclusions about connectivity related tasks. This enhancement has also been mentioned by Andrienko and Andrienko (2005).

The time series visualisation of LOD 2 enabled the participants to discover changes in the behaviour of specific IMPCs, they mentioned for example that one IMPC was the strongest first, whereas another one caught up over time and stayed important after. This led the participants to reason about changes in routing decisions, etc. Results from the interview sessions show that a line showing the overall volume would be seen as useful for comparing values.

As mentioned above, some of the participants filtered the data, e.g. to look at the behaviour of individual IMPCs. However, similar to other visualisation types, this option will prove difficult if there are too many objects to filter from, e.g. all the countries. Queries with the help of timeboxes, as suggested by Hochheiser and Shneiderman (2004) might be a promising way for selecting the data of interest.

Time series visualisations help analysts to explore the data mainly in terms of traffic volumes and seasonality effects and less regarding connectivity. The temporal dimension is important for analysing flows, as it allows to see, for example, how the spatial distribution or routing of flows changes over time, or how disruptions affect the flows. This can, for example, help analysts to explore the data regarding bottlenecks and network resilience, which is important from an operational point of view.

As can be seen by the above discussion, each visualisation type has its advantages and drawbacks. The next section thus shifts the focus to the potential of combining several visualisation types.

Visualisation types should be combined because they complement each other, but which ones should be combined depends on the analysts' interest

The analysis of advantages and limitations of the different visualisation types clearly shows that some of the evaluated visualisations are better suited for investigations with a focus on connectivity (e.g, map visualisations and ODM to a certain degree), traffic volumes (e.g., flow map, graduated symbol map, time series) or seasonality (time series). Some are better for a quick overview (static implementations, especially the flow map), others allow to derive more in depth information (interactive implementations). These findings are important because it allows analysts to chose the tools that are the best fit to their specific interests and needs. To address different aspects in data exploration, various methods should be combined, as they can complement each other.

Some visualisations were found to be very useful for detecting large anomalies (e.g. the graduated symbols map). The detection of anomalies in the data can help the analysts to identify problems in the IPN, whether they are caused by an external factor like a natural disaster or a reporting problem or technical issue. This detection of irregularities in the postal network is of particular interest in recent developments of prediction analysis of postal shipments, where accurate baseline values are crucial for such prediction models. In the case study at hand, the large anomaly of 2013 in the data for the US would for example have an effect on such models. It has to be mentioned that such large anomalies itself affect the detection of other irregularities. It is for example assumed that the mentioned large data issue in 2013 affected the ability of the participants to note the much smaller anomalies in 2014, because the first one was so dominant in this particular time span.

Results from observations made during the usability study as well as the interview sections show that some of the participants were overwhelmed by the choropleth map or the graduated symbol map, which both contained interactive features. Although those visualisations allowed them to extract more detailed information (e.g. extract absolute values) than the static representations, they found it was too much 'work' and they wanted to get all the necessary information at first glance. Other participants were very keen on exploring the data with the interactive maps by zooming, clicking, etc. Hence, interactive features are only useful if the analyst is willing to spend time for querying. Many of the participants stressed in the interviews that they would prefer to combine different types of visualisations, as some provided them with a quick first overview and others seemed very useful to them for further exploration of the data. These findings and observations are thus in accordance with Hypothesis 2b and the idea behind the "Visual Information Seeking Mantra: Overview first, zoom and filter, then details-on-demand" presented by Shneiderman (1997) or the meaning of the three "levels of reading" defined by Bertin (1983), as well as other studies by Roberts (2007) and Keim et al. (2010).

Results from the requirements analysis as well as the qualitative feedback and the observations made during the usability study show that the individual preferences regarding visualisation type differ among the participants. Previous experience and familiarity with certain visualisation types explain part of this variation in preferences, as was found during the interviews of the usability study. Results from the interview sections further indicate that the preferences are also related to the different interests and uses of visual representations by the participants. While some of the participants publish printed reports, others are more interested in interactive tools with which they can investigate the data in more depth. There was no clear overall preference for interactive visualisations, contrary to what was expected in Hypothesis 2c based on previously reported findings (Hegarty et al., 2009). Due to the individual foci regarding their work, some participants are more interested in gaining a quick overview and therefore liked static visualisations like the flow map, while others had difficulties to extract relevant information that served there needs from this static representation. It is important to take into account the users' needs and preference regarding visualisation types, because only if a visualisation helps them tackle the issues that are of interest to them and they like to work with it, they will be able to benefit from its full potential.

The decision about which visualisation types should be applied and potentially combined also depends on individual preferences, but some recommendations about the suitability for certain needs or tasks can be made based on the findings of this thesis. It can be said that for a first general overview, static implementations such as the flow map are very useful. Visualisation types like the ODM are useful if one is interested in analysing the main connections between specific countries and wants to see more detailed information or exact values at first sight. However, if the connectivity within the whole postal network is of interest, a map provides additional valuable information. If the analyst is interested in big countries and main players, a choropleth map could be a fair choice. However, if he or she is also interested in seeing flows to or from smaller places, either a graduated symbol map or a flow map is the better choice. Keeping in mind that the flow map could also be implemented in an interactive way, this would allow to investigate the data in more detail if needed. However, the usability of such an implementation would again have to be tested with the users.

The number of connections that need to be visualised also plays a crucial role in the selection of the visualisation type. As could be seen with all (geo-)spatial visualisations, the high number of origins or destinations was often a critical factor. In the case of ODMs, it would be possible to show only a selected number of origins or destinations, respectively. Depending on the interest of the analyst, a grouping of the visual representation for example by world region could be a solution, thus creating separate ODMs for showing flows to and from countries in the Americas, the Asia-Pacific region, or Europe, etc. (cf. Guo et al., 2006).

Rather than in the form of animation, or generally if the visualisations are used in a printed form, snap shots of the defined time intervals can be shown side by side. This has the advantage that the analyst does not have to remember the previously seen sequences, which is a known difficulty when using animation as an exploratory tool, especially if it is used in a non-interactive way (Hegarty et al., 2009; Brunsdon, Corcoran, and Higgs, 2007).

For a stronger focus on changes over time, a time series visualisation was perceived as very helpful. Depending on where the main interest of the analyst lies, a time series visualisation could also be combined with one of the spatial visualisations.

5.3 Detecting Cyclical Patterns and Anomalies

It is known that there are cyclical patterns in the postal OD data which indicate weekend/weekday patterns, seasonal effects, special holidays, etc. The third research question aimed at assessing the applicability of the selected visualisation methods with regard to cyclical patterns and anomalies.

Time series graphs are well suited for detecting weekly patterns

In accordance with Hypothesis 3 and previous literature (Keim et al., 2010), results show that out of the tested visualisation methods, the time series graph proofed to be the most useful visualisation type for discovering cyclical patterns (see for example figure 5.1 for weekly patterns). All of the participants were able to detect cyclical patterns such as a weekly low as well as a monthly peak in the US outbound data for *Event A* (post offices). However, none of the participants reported any cyclical patterns for the CL data set with any of the visualisation types, which is largely due to the high volatility in the data.

However, it needs to be said that due to the available data sets and the time intervals that were chosen for most of the implementations, results obtained for the detection of cyclical patterns are limited in this study. A larger time span should be analysed to see whether seasonal effects can be detected. Disruptions due to natural hazards such as floods could potentially also be detected on a larger scale, as they can also show cyclical behaviour (Keim et al., 2010). Aggregation to other temporal units, e.g. days or hours, could show patterns and anomalies on different scales. Literature suggest that other visualisation types such as spiral visualisations (Keim et al., 2010), circular dot plots (Brunsdon, Corcoran, and Higgs, 2007) or ringmaps (Zhao, Forer, and Harvey, 2008) are more useful for analysing cyclical patterns in the data. As this was not possible within the scope of this study, it is suggested that such visualisation types are implemented and evaluated with a further user study, since the temporal dimension is a crucial part for analysing this type of data, its spatio-temporal characteristics and changes.

5.4 Ability to Explain the Detected Anomalies

The fourth research question focused on the underlying reasons for the patterns and anomalies in the postal data and aimed at assessing whether the selected visual representations can help analysts explain those discovered patterns and anomalies.

Visualisations help to reflect about reasons for discovered patterns and anomalies, but need additional data for explanation and confirmation

The results from the usability study show that the participants were able to draw some conclusions about the underlying reasons for patterns and anomalies they discovered in the visualisations of the postal data. However, the prior knowledge of the participants as well as the focus of their respective work played an important part in their reasoning. The participants who are more focused on technological and operational aspects in their work tended to rather attribute all anomalies to data or reporting issues, while others who are more interested in network activity, legal or sustainability issues thought of reasons like natural disasters, political events, economic problems or strikes as being potential reasons for anomalies. The reasons underlying the detected patterns seemed to be easier for them to define: all of the participants associated the detected patterns to a large extent with e-commerce development and hence trade. Overall, the findings thus support statements from previous literature (Andrienko and Andrienko, 2010) and Hypothesis 4: visualisations can help to understand underlying processes, but additional information is needed in order to confirm assumptions or enable a reliable interpretation. It is important to recognise that visualisations cannot fully explain underlying reasons for discovered patterns or anomalies, but that they are a helpful tool for guiding analysts towards events in the data that may be worth investigating in more detail.

Results further show that the general interest of the participant and the time spent for exploring the data played a role in reflecting about underlying reasons. The participants who looked at the data most closely found some hints that lead them to the correct explanation for at least some of the detected anomalies. In particular, it was mentioned by some of the participants that the visualisations were able to show the difference between a purely technical issue and other disruptions such as a strike or a natural disaster. A purely technical issue will show missing data, but will return back to normal after the issue has been resolved, whereas with other issues, there is a backlog and a large catch up subsequently (similar to the regular peaks found after the weekends). It was also pointed out by some of the participants that the duration of an event can provide some information about the underlying reason, as a natural disaster often lasts for a few days, whereas events like a strike can last much longer. However, as resulted from the requirements analysis, the time until the flows are 'back to normal' not only depends on the type of the disruption itself, but also on the resilience of the affected country and the network as a whole.

A combination of the postal data with additional information such as the reports from the EmIS system, weather reports or news articles could help investigate the reasons behind detected anomalies, as was also suggested by the participants during the interview sessions. Such an approach would be worth investigating in more depth because i can have implications not only for analysing the postal network but potentially also other flow networks.

5.5 Unexpected Results

Normalised values

Some of the participants, mostly the experts working in a more technical field concerned with operational issues, were confused by the visualisations using normalised values (LOD 1), i.e. number of items per capita. It was not the visualisation types that confused them, but the use of normalised values, as they did not see any advantage of using those values. They would clearly prefer to see absolute values. Other experts with a stronger focus on global networks and network connectivity found it very helpful to see relative numbers.

Outliers in SUS

The result for the System Usability Scale shows that there was not a huge difference regarding the overall score between any of the visualisation types. However, except for the time series visualisation, the variance of the scores was very high, with extreme outliers for the ODM, the flow map as well as the choropleth map. This indicates that the participants had very different opinions about the usability of the different visualisation types. To get a more comprehensive result, more people would be needed for rating the visualisation types. It was surprising that some found e.g. the choropleth map extremely intuitive and helpful, while others felt completely overwhelmed by this representation and its interactive features.

Difference in reaction regarding the form of visual representations

It was further interesting to see how differently the participants reacted regarding the presentations of the visualisation types. While some had no problems at all to switch between the CWs for the static visualisations (ODM and flow map), others felt like they could not deal with this at all. However, even when they thought they had problems seeing anything at all because of the way the visualisation was presented, they still performed rather well.

5.6 Recommendations for the Visual Exploration of Postal OD Data

Results from the usability study and the interviews clearly show that interests, preferences and needs differ between the experts, which indicates that flexibility is key for designing a solution that caters for those various needs and interests. Thus, it is evident that data visualisation for pattern and anomaly detection in global postal data requires a customised approach and customised solutions. In the case study at hand, it became clear that the postal experts are interested in 1) connectivity issues, 2) traffic trends and 3) seasonal effects in terms of pattern visualisation and anomaly discovery. Therefore, any customised solution must tackle these three needs with respect to flow visualisation for enabling the visual detection of meaningful patterns and anomalies. Options for such an approach may include a toolbox with different types of visualisations that the users can choose from. Ideally, they can combine different visualisation types according to their needs in a dashboard-like manner with linked plots to enable a multiple linked view of the data (Roberts, 2007; Andrienko and Andrienko, 2005). The ability to switch between absolute and normalised values was for example mentioned in the interviews as a desired feature. It is further suggested that the temporal dimension is added either with a time series visualisation or in the form of a time slider to move back and forth in time in the other visualisations. Collecting the data with additional information about the time zone or directly converting the collected timestamps to UTC time would facilitate analysis with a focus on the temporal dimension.

An initial combination of visualisations could be recommended for certain tasks. For example, a combination of a flow map, a choropleth map and a time series graph could serve for getting a first quick spatial overview (flow map), a more detailed view (choropleth map) and information about temporal changes (time series graph). Additionally, to each visualisation type, recommendations about its use should be provided, e.g. information about which visualisation type is the best fit for which type of analysis, or example use cases for each visualisation type.

Ideally, the dashboard would provide access to additional data sources such as the EmIS reports, news feeds, flight data, weather reports, etc. in order to enrich the data and analysis potential.

Enabling to directly select the so-called 'mirror data'[2] if the analyst thinks that he or she has detected something irregular in the flows from one country would be an additional benefit.

5.7 Study Limitations

The data that could be extracted for this study was limited because of its size. Thus, two case studies and a certain period of interest were selected to conduct the study. For some points that were addressed, especially the temporal and cyclical component, a larger time period would probably lead to further insights.

Because of time limitations regarding the usability studies, a selection had to be made in terms of visualisation types and data to show. Since some of the visualisations were static and others interactive, the concept of 'informationally equivalent' visualisations by Larkin and Simon (1987) could only partly be followed. For each LOD, the same amount of information was used to create the respective visualisations, but the participants could access some more information in the interactive visualisations. However, many of the participants actually did not use the interactive features often, thus the effect of this partial lack of equivalency did not seem to have a major influence in the outcomes.

Not all of the visualisations that were tested could be presented in the exact same way (static vs. interactive, 6 CWs on one screen vs. switching between CWs). This may have influenced some of the results that were obtained during the study. However, the participants generally made very differentiated statements about whether they did or did not like a visualisation as such or the way it was presented to them.

[2] If analysts discover an anomaly in flows to or from one specific country, for example, drops in flows from CA to the US when exploring the US data, they might be interested in looking at the CA data for the respective time period.

Since this thesis focused on two case studies, a learning effect during the usability study could not be avoided. The effect was sought to be minimised by rotating the data sets within and between the sessions, but could not be eliminated completely.

As mentioned before, quantitative measures like the System Usability Scale have limited implications in the frame of this study, as there was an outlier for each visualisation type and there were only six participants. To improve results, it should be conducted with more participants. For the study at hand, qualitative feedback was found to be much more informative.

Log scale was used for some of the visualisation because of the value distribution. This was slightly confusing for some of the participants. However, they were always informed about it and the more used they were to it, the less it seemed to bother them.

Data quality

The postal data collected by the UPU only reflects a part of the global postal traffic. A substantial part of the postal shipments is handled by private companies or entities that do not share their data with UPU, hence it is not possible to show a comprehensive picture of global postal flows with the data used in this study. It is, for example, not clear whether the data about flows from European countries to the US were missing because they had not been collected by the UPU itself or whether there was another reason for it.

Missing spatial information about IMPCs, erroneous data in the EDI messages or unidentifiable codes (e.g. of post offices) are other data quality problems. The mentioned issues did probably not severely influence the outcomes of this specific study, however the issues remain for potential further studies with other data sets.

The communication of disruptions via the EmIS database is not compulsory for UPU member countries. Thus, it is not a fully reliable source for explaining the reasons for certain anomalies appearing in the data. However, it can serve as a source of information and is expected to be helpful in a combination with the visualisation of the postal data as well as data from other sources. If the communication of disruptions was compulsory, it would improve the situation.

Chapter 6

Conclusion and Outlook

6.1 Conclusion

In this thesis, a first assessment of the potential of selected visualisation types for the visual exploration of OD data with a focus on pattern and anomaly discovery in global postal data was conducted. It followed an exploratory approach, i.e. testing several tools and visualisation types that had been discussed in previous work regarding network- and flow visualisation. Apart from the theoretical background from literature research, the thesis generated new empirical knowledge obtained from a requirements analysis and a usability study conducted with postal experts. The requirements analysis was used for selecting the visualisation methods that were applied to two use cases of countries within the IPN for further testing. With an exploratory analysis of the tools and in-depth expert interviews, the thesis contributes to current research in geovisualisation and visual analytics. It provides results from the assessment of advantages and drawbacks of several visualisation types for visualising large OD data sets. Furthermore, the study contributes to ongoing research by providing insights about the real-world usability of the selected visualisation types for the visual detection of patterns and anomalies in postal data, which were obtained through a usability study conducted with experts from the actual target group.

Analysing the advantages and limitations of visualisation methods from a theoretical side and a 'top-down' judgement is an often followed approach in geovisualisation research and studies about visual exploration of flow- or movement data. However, conducting expert user studies for evaluating the visualisation types regarding their usability and effectiveness in daily work of analysts helps to define the actual requirements of data visualisation techniques and improve them according to the identified needs and requirements. This is in line with findings by Brunsdon, Corcoran, and Higgs (2007) for instance.

The following paragraphs summarise the findings with regard to the four main research questions of this thesis.

Results from this study show that the encoding (granularity) of the spatial units, the level of connectivity as well as the distance between origins and destinations affect a) the visualisation of large OD data sets and b) the analysts' ability to analyse the data with respect to the visual detection of patterns and anomalies. In line with many other studies in geovisualisation (e.g., Andrienko and Andrienko, 2010), results show that there is a trade-off between generalisation for better visual perception and information loss *(RQ 1)*.

This has implications for the analysts regarding the choice of LOD in which the data is displayed and the choice of visualisation type, depending on what he or she is most interested in investigating in the data.

Each of the tested visualisation types has its advantages and drawbacks. This needs to be taken into account by analysts when selecting a method or a combination of methods for visualising the data. It is important for the analysts to be aware of the purpose of the visualisation. If they want to get a quick overview, a static flow map or an ODM might be a good fit, but if they seek to explore the data in more depth, an interactive map approach would be the better choice. Further, analysts need to be aware of what they want to learn from the visualisation, whether they are interested in finding out about temporal patterns or traffic trends, or whether they are more interested in the global connectivity of a specific country or location and its changes in space and over time. All these points need to be addressed by the analysts as they have implications for choosing a visualisation that meets their requirements *(RQ 2)*.

Out of the visualisation methods tested in this study, participants could only detect cyclical patterns with the time series visualisation. Partly, this can be attributed to the aggregation levels that were chosen for the visualisations in this thesis. It is suggested that other visualisation methods that are suggested in visual analytics literature are tested in a next step *(RQ 3)*.

Results from this study further show that, depending on the interest and the background knowledge of the analysts, visualisations can help them reflect about reasons underlying the discovered patterns and anomalies. However, it also became clear that the postal data on its own cannot provide sufficient information for confirming the analysts assumptions and reasoning regarding the drivers of such patterns or disruptions. However, one can argue that including additional information from other sources could improve this situation and help the analysts to strengthen or reject certain assumptions *(RQ 4)*.

With the real-world case study using global postal data, this thesis could demonstrate the benefits of visualising large OD data sets, but also investigate the challenges of such a task. The identified limitations and potential of the evaluated visualisation methods are not necessarily specific to visualising postal data, but are rather related to the size and structure of the visualised data. Hence, other large (global) OD data sets from, for example, migration or trade research are likely to be confronted with similar challenges and might benefit in the same way from visualisation. Thus, the findings can potentially be generalised to similar cases in other fields where flow data is relevant.

6.2 Outlook

This study was the first to analyse several methods for visually representing global postal OD data collected by the UPU. Its findings lead to new questions that are worth addressing in further research, some of which are presented below.

The study investigated two different LODs and a temporal aggregation of calendar weeks (years for the time series graph). Results show that users could discover certain patterns and anomalies with the selected visualisations, however results indicate that using different time intervals could lead to additional insights, especially with regard to detecting anomalies that only lasted for a few days. Therefore, the thesis suggests to conduct further studies with other temporal intervals, for example days or even hours to investigate whether short term disruptions are better detectable at such a temporal scale.

The thesis further suggests to combine and test selected visualisation types in an environment that provides the possibility to link several plots dynamically. Allowing the users to filter the data in order to focus on specific subsets was only possible in the time series visualisation, but it showed promising results and filtering was mentioned as desirable by many of the participants. The thesis thus suggests to test a dashboard-like environment that provides features for combining plots, filtering as well as moving back and forth in time. Such an implementation should again be tested with the actual user group to assess its effectiveness in the operational work.

Further, as findings from a study with postal OD data might also have indications for other studies with global flow data, the thesis suggests that such a visualisation environment should be tested with other types of flow data to asses whether it is suitable for other fields of research. This could be of particular interest with regard to the previous study with postal data by Hristova et al. (2016), which compared several global networks.

New technical possibilities and ever more realtime data that is available (Sun et al., 2013) open up new possibilities for visualising and linking data from different sources. Results from this study, particularly of the interview sessions, show that such an approach could be very interesting for further investigating postal data. Especially for getting a more comprehensive picture and explaining discovered anomalies, it could be beneficial to combine data from other sources such as EmIS reports, weather data, news feeds, etc. with the postal data.

As was stated in the discussion, results regarding the suitability of the selected visualisation types for detecting cyclical patterns were limited. However, this is a very interesting angle to look at in the postal data. The thesis thus suggest to undertake further investigation of the postal OD data with other visualisations like spiral visualisations (Keim et al., 2010), circular dot plots (Brunsdon, Corcoran, and Higgs, 2007) or ringmaps (Zhao, Forer, and Harvey, 2008).

The disruptions that were reported in the EmIS tool were categorised during the work on this thesis. It would be interesting to analyse whether there is a way to detect any differences between the different types of anomalies in and with the visualisations. Some results suggest that there could, for example, be a differentiation between technical issues and disruptions due to strikes or natural disasters. Another interesting approach would be to analyse the time needed for the network to go 'back to normal', e.g. obtain knowledge about network resilience. Additional investigation of the temporal information in the data regarding time difference between planned and actual arrival or departure time of items could be one way for addressing this issue.

As can be seen, numerous new questions emerge from discussing the findings from this thesis. The insights gained from this study can hopefully contribute to a better understanding of the potential and limitations of different visualisation types for the visual detection of patterns and anomalies in large OD data sets.

Literature

Ad hoc group on Disaster Risk Management (2015). "UPU Policy and Strategy Paper on disaster risk management".

Andrienko, G. et al. (2010). "A framework for using self-organizing maps to analyze spatio- temporal patterns , exemplified by analysis of mobile phone usage". In: *Location Based Services* 4, pp. 200–221. ISSN: 1748-9725. DOI: 10.1080/17489725.2010.532816.

Andrienko, Gennady and Natalia Andrienko (2005). "Visual Exploration of the Spatial Distribution of Temporal Behaviors". In: *Ninth International Conference on Information Visualisation (IV'05)*, pp. 571–576. ISSN: 15591662. DOI: 10.1145/2505395.2505398. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1509132$\backslash$npapers2://publication/doi/10.1109/IV.2005.83.

— (2010). "A General Framework for Using Aggregation in Visual Exploration of Movement Data". In: *The Cartographic Journal* 47.1, pp. 22–40. ISSN: 00087041. DOI: 10.1179/000870409X12525737905042.

Andrienko, Natalia and Gennady Andrienko (2007). "Designing Visual Analytics Methods for Massive Collections of Movement Data". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 42.2, pp. 117–138. ISSN: 0317-7173. DOI: 10.3138/carto.42.2.117.

— (2011). "Spatial Generalization and Aggregation of Massive Movement Data". In: *IEEE Transactions on Visualization and* 17.2, pp. 205–219. ISSN: 1077-2626. DOI: 10.1109/TVCG.2010.44. URL: http://www.computer.org/portal/web/csdl/doi/10.1109/TVCG.2010.44.

— (2013). "A visual analytics framework for spatio-temporal analysis and modelling". In: *Data Mining and Knowledge Discovery* 27.1, pp. 55–83. ISSN: 13845810. DOI: 10.1007/s10618-012-0285-7.

Andrienko, Natalia, Gennady Andrienko, and Peter Gatalsky (2003). "Exploratory spatio-temporal visualization: An analytical review". In: *Journal of Visual Languages and Computing* 14.6, pp. 503–541. ISSN: 1045926X. DOI: 10.1016/S1045-926X(03)00046-6.

Anson, José and Matthias Helble (2014). "Development strategies for the postal sector: an economic perspective". In: Delft: Universal Postal Union, p. 237.

Ben-Ayed, Omar and Salem Hamzaoui (2012). "Multiobjective multiproduct parcel distribution timetabling: a real-world application". In: *International Transactions in Operational Research* 19.4, pp. 613–629. ISSN: 1475-3995. DOI: 10.1111/j.

1475-3995.2012.00852.x. URL: http://dx.doi.org/10.1111/j.1475-3995.2012.00852.x.

Bertin, Jacques (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, Wisconsin: The University of Wisconsin Press, p. 415.

Bleich, Erik and Robert Pekkanen (2013). *How to Report Interview Data*. Ed. by Mosley Layna. 1st ed. Ithaca: Cornell University Press. Chap. 4, pp. 84–105. ISBN: 0-8014-6796-9.

Boyandin, Ilya et al. (2010). "Using Flow Maps to Explore Migrations Over Time". In: *Geospatial Visual Analytics Workshop in conjunction with The 13th AGILE International Conference on Geographic Information Science*. Vol. 2. DOI: 10.1.1.455.7957. URL: https://diuf.unifr.ch/main/diva/sites/diuf.unifr.ch.main.diva/files/jflowmap-geova10.pdf.

Brooke, John (1996). "SUS - A quick and dirty usability scale". In: *Usability evaluation in industry* 189.194, pp. 4–7. ISSN: 1097-0193. DOI: 10.1002/hbm.20701. URL: http://hell.meiert.org/core/pdf/sus.pdf.

Brunsdon, C. (2001). "The comap: exploring spatial pattern via conditional distributions". In: *Computers, Environment and Urban Systems* 25.1, pp. 53–68. ISSN: 01989715. DOI: 10.1016/S0198-9715(00)00042-9.

Brunsdon, Chris, Jonathan Corcoran, and Gary Higgs (2007). "Visualising space and time in crime patterns: A comparison of methods". In: *Computers, Environment and Urban Systems* 31.1, pp. 52–75. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2005.07.009.

Carr, Daniel B., Denis White, and Alan M. MacEachren (2005). "Conditioned choropleth maps and hypothesis generation". In: *Annals of the Association of American Geographers* 95.1, pp. 32–53. ISSN: 00045608. DOI: 10.1111/j.1467-8306.2005.00449.x.

Courage, Catherine and Kathy Baxter (2005). *Understanding Your Users: A Practical Guide to User Requirements Methods, tools, and Techniques*. Ed. by Stuart Card, Jonathan Grudin, and Jakob Nielsen. San Francisco: Elsevier Inc., p. 781. ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Cui, Weiwei et al. (2008). "Geometry-based edge clustering for graph visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 14.6, pp. 1277–1284. ISSN: 10772626. DOI: 10.1109/TVCG.2008.135.

Dou, Wenwen et al. (2010). "Comparing different levels of interaction constraints for deriving visual problem isomorphs". In: *VAST 10 - IEEE Conference on Visual Analytics Science and Technology 2010, Proceedings*, pp. 195–202. DOI: 10.1109/VAST.2010.5653599.

Ferreira, Nivan et al. (2013). "Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12, pp. 2149–2158. ISSN: 10772626. DOI: 10.1109/TVCG.2013.226.

Fredrikson, Anna et al. (1999). "Temporal , Geographical and Categorical Aggregations Viewed through Coordinated Displays : A Case Study with Highway Incident Data". In: *Workshop on New Paradigms in information Visualization and Manipulation* 99, pp. 26–34. DOI: `10.1145/331770.331780`.

Gilbert, M et al. (2005). "Cattle movements and bovine tuberculosis in Great Britain." In: *Nature* 435.7041, pp. 491–496. ISSN: 0028-0836. DOI: `10.1038/nature03548`.

Guo, D. (2007). "Visual analytics of spatial interaction patterns for pandemic decision support". In: *International Journal of Geographical Information Science* 21.8, pp. 859–877. ISSN: 1365-8816. DOI: `10.1080/13658810701349037`.

Guo, Diansheng et al. (2006). "A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP)". In: *IEEE Transactions on Visualization and Computer Graphics* 12.6, pp. 1461–1474. ISSN: 10772626. DOI: `10.1109/TVCG.2006.84`. arXiv: `arXiv:1011.1669v3`.

Guo, Diansheng et al. (2009). "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach". In: *Cartogr Geogr Inf Sci.* 32.2, pp. 113–132. DOI: `10.1559/1523040053722150.Multivariate`.

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky (2010). "A tour through the visualization zoo". In: *Communications of the ACM* 53.6, p. 59. ISSN: 00010782. DOI: `10.1145/1743546.1743567`. URL: `http://portal.acm.org/citation.cfm?doid=1743546.1743567`.

Hegarty, Mary et al. (2009). "Naïve Cartography: How Intuitions about Display Configuration Can Hurt Performance". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 44.3, pp. 171–186. ISSN: 0317-7173. DOI: `10.3138/carto.44.3.171`. URL: `http://utpjournals.metapress.com/openurl.asp?genre=article{\&}id=doi:10.3138/carto.44.3.171`.

Heil, Erich and Dominik Walch (2015). "Analysis and V isualization of Austria ' s Social Insurance System". In: *CEUR Workshop Proceedings*. Vol. 1481, pp. 105 – 109.

Henry, Nathalie and Jean-Daniel Fekete (2006). "MatrixExplorer: a dual-representation system to explore social networks." In: *IEEE transactions on visualization and computer graphics* 12.5, pp. 677–84. ISSN: 1077-2626. DOI: `10.1109/TVCG.2006.160`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/17080787`.

Herman, I., G. Melancon, and M.S. Marshall (2000). "Graph visualization and navigation in information visualization: A survey". In: *IEEE Transactions on Visualization and Computer Graphics* 6.1, pp. 24–43. ISSN: 10772626. DOI: `10.1109/2945.841119`. URL: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=841119`.

Hernandez, Tony (2007). "Enhancing retail location decision support: The development and application of geovisualization". In: *Journal of Retailing and Consumer Services* 14.4, pp. 249–258. ISSN: 09696989. DOI: `10.1016/j.jretconser.2006.07.006`.

Hochheiser, Harry and Ben Shneiderman (2004). “Dynamic query tools for time series data sets: Timebox widgets for interactive exploration”. In: *Information Visualization* 3.1, pp. 1–18. ISSN: 1473-8716. DOI: 10.1057/palgrave.ivs.9500061. URL: http://www.palgrave-journals.com/doifinder/10.1057/palgrave.ivs.9500061.

Höld, Regina (2009). *Zur Transkription von Audiodaten*. Ed. by Buber Renate. 2nd ed. Wiesbaden: Gabler. Chap. 6, pp. 655–668. ISBN: 978-3-8349-0976-3. DOI: 10.1007/978-3-8349-9441-7.

Hristova, Desislava et al. (2016). “The International Postal Network and Other Global Flows as Proxies for National Wellbeing”. In: *PLoS ONE* 11.6, pp. 1–19. DOI: 10.1371/journal.pone.0155976.

Keim, Daniel et al. (2010). *Mastering the Information Age – Solving Problems with Visual Analytics*. Ed. by Daniel Keim et al. Vol. 1. Eurographics Association, p. 168. ISBN: 978-3-905673-77-7. URL: http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf.

Kelly, Mary et al. (2012). “Historical Internal Migration in Ireland”. In: *City Research Online - City University London*, pp. 14–16. DOI: 10.1016/j.cpr.2010.02.004.

Larkin, Jh and Ha Simon (1987). “Why a diagram is (sometimes) worth ten thousand words”. In: *Cognitive Science* 99, pp. 65–99. ISSN: 0364-0213. DOI: 10.1016/s0364-0213(87)80026-5. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1551-6708.1987.tb00863.x/abstract.

Lee, Sukwon et al. (2016). “How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice’s Information Visualization Sensemaking”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 499–508. ISSN: 10772626. DOI: 10.1109/TVCG.2015.2467195.

Lewalter, D. (2003). “Cognitive strategies for learning from static and dynamic visuals”. In: *Learning and Instruction* 13.2, pp. 177–189. ISSN: 09594752. DOI: 10.1016/S0959-47520200019-1.

Li, Songnian et al. (2016). “Geospatial big data handling theory and methods: A review and research challenges”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 115, pp. 119–133. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2015.10.012. arXiv: 1511.03010. URL: http://dx.doi.org/10.1016/j.isprsjprs.2015.10.012.

Lloyd, Robert (1997). *Spatial cognition: geographic environments*. Dordrecht: Kluwer Academic Publishers, p. 287.

Marković, Dejan, Biljana Grgurović, and Slavica Štrbac (2011). “The Use of Spatial Data for Segmentation of the Postal Service Market”. In: *Technological and Economic Development of Economy* 17.1, pp. 87–100. ISSN: 2029-4913. DOI: 10.3846/13928619.2011.554016. URL: http://www.tandfonline.com/doi/abs/10.3846/13928619.2011.554016.

Nagel, Till and Christopher Pietsch (2015). *cf. city flows*. URL: https://uclab.fh-potsdam.de/cf/.

Nielsen, Jakob (2001). "Usability Metrics".

Nieto Corredera, Daniel (2015). *Measuring postal e-services development. A global perspective*. Tech. rep. Bern.

Paci, Raffaele and Stefano Usai (2009). "Knowledge flows across European regions". In: *Annals of Regional Science* 43.3 SPEC. ISS. Pp. 669–690. ISSN: 05701864. DOI: 10.1007/s00168-008-0256-5.

Rae, Alasdair (2009). "From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census". In: *Computers, Environment and Urban Systems* 33.3, pp. 161–178. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2009.01.007. URL: http://dx.doi.org/10.1016/j.compenvurbsys.2009.01.007.

Roberts, Jonathan C. (2004). "Exploratory Visualization with Multiple Linked Views". In: *Exploring Geovisualization*, pp. 149–170.

— (2007). "State of the Art: Coordinated & Multiple Views in Exploratory Visualization". In: *Proceedings of the 5th International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)* Cmv, pp. 61–71. DOI: 10.1109/CMV.2007.20. URL: http://dx.doi.org/10.1109/CMV.2007.20.

Sander, N et al. (2014). "Visualising Migration Flow Data with Circular Plots".

Sander, Nikola, Guy J. Abel, and Ramon Bauer (2014). *The Global Flow of People*. URL: http://www.global-migration.info/.

Severtson, Dolores J and Robert E Roth (2015). "Does Interactivity Enhance Reasoning with Maps that Visualize Uncertain Information ? Preliminary Results from an Empirical Study". In: *IEEE VIS 2015 - IEEE Computer Society*. October.

Shneiderman, Ben (1997). "Next generation of graphical user interfaces: Information visualization and better window management". In: *Displays* 17.3-4, pp. 125–129. ISSN: 01419382. DOI: 10.1016/S0141-9382(97)00005-X.

Singh, S. Paul et al. (2006). "Measurement and analysis of the next-day air shipping environment for mid-sized and lightweight packages for DHL, FedEx and United parcel service". In: *Packaging Technology and Science* 19.April, pp. 227–235. ISSN: 08943214. DOI: 10.1002/pts.726.

Sinton, D. (1977). "The Inherent Structure of Information as a Constraint to Analysis: Mapped Thematic Data as a Case Study." In: *First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems.* Vol. 7. G. Dutton. Dedham, MA, Laboratory for Computer Graphics and Spatial Analysis., pp. 1–17.

Slingsby, Aidan, Mary Kelly, and Jason Dykes (2014). "OD maps for showing changes in Irish female migration between 1851 and 1911". In: *Environment and Planning A* 46.12, pp. 2795–2797. ISSN: 0308-518X. DOI: 10.1068/a140112g. URL: http://epn.sagepub.com/lookup/doi/10.1068/a140112g.

Slingsby, Aidan et al. (2012). "OD Maps for Studying Historical Internal Migration in Ireland". In: *Poster presented at the IEEE Conference on Information Visualization (InfoVis)*. Seattle, Washington, US.

Sun, Guo Dao et al. (2013). "A survey of visual analytics techniques and applications: State-of-the-art research and future challenges". In: *Journal of Computer Science and Technology* 28.5, pp. 852–867. ISSN: 10009000. DOI: 10.1007/s11390-013-1383-8.

Tabak, P and T Kljak (2009). "Traffic Flows Model of Postal Items Based on Input-Output Demands in the Public Postal Network: Case Study of Croatia". In: *Promet-Traffic & Transportation* 21.6, pp. 407–414. ISSN: 03535320. URL: {\textless} GotoISI{\textgreater}://000273468000004.

Thomas, James J. and Kristin A. Cook (2005). "Illuminating the path: The research and development agenda for visual analytics". In: *IEEE Computer Society*, p. 184. ISSN: 1664302X. DOI: 10.3389/fmicb.2011.00006. URL: http://vis.pnnl.gov/pdf/RD{_}Agenda{_}VisualAnalytics.pdf.

Tobler, Waldo R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region". In: *Economic Geography* 46, pp. 234–240.

— (1987). "Experiments In Migration Mapping By Computer". In: *Cartography and Geographic Information Science* 14.2, pp. 155–163. ISSN: 15230406. DOI: 10.1559/152304087783875273.

Tufte, Edward Rolf (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut.

United Nations (2007). *International Strategy for Disaster Reduction Hyogo Framework for Action 2005-2015: Building the Resilience of Nations and Communities to Disasters*. Tech. rep. Geneva: United Nations, p. 25.

Widloecher, Patrick (2013). "Mandate, members and work plan of the Ad Hoc Group on Disaster Risk Management".

Wong, Pak Chung et al. (2006). "Have green -A visual analytics framework for large semantic graphs". In: *IEEE Symposium on Visual Analytics Science and Technology 2006, VAST 2006 - Proceedings*, pp. 67–74. DOI: 10.1109/VAST.2006.261432.

Wood, J., A. Slingsby, and J. Dykes (2011). "Visualizing the dynamics of London's bicycle hire scheme". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 46.4, pp. 239–251. ISSN: 0317-7173. DOI: 10.3138/carto.46.4.239.

Wood, Jo, Jason Dykes, and Aidan Slingsby (2010). "Visualisation of Origins, Destinations and Flows with OD Maps". In: *The Cartographic Journal* 47.2, pp. 117–129. ISSN: 0008-7041. DOI: 10.1179/000870410X12658023467367.

Yang, Yalong et al. (2016). "Many-to-Many Geographically-Embedded Flow Visualisation: An Evaluation". In: *IEEE Transactions on Visualization and Computer Graphics*, p. 1. ISSN: 1077-2626. DOI: 10.1109/TVCG.2016.2598885. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7539669.

Zhang, Hantao (2010). *Latin Square Design*. DOI: `10.4135/9781412961288`.

Zhao, Jinfeng, Pip Forer, and Andrew S Harvey (2008). "Activities, ringmaps and geovisualization of large human movement fields". In: *Information Visualization* 7.3-4, pp. 198–209. ISSN: 1473-8716. DOI: `10.1057/palgrave.ivs.9500184`. URL: `http://www.palgrave-journals.com/doifinder/10.1057/palgrave.ivs.9500184$\backslash$nhttp://ivi.sagepub.com/lookup/doi/10.1057/palgrave.ivs.9500184`.

Appendix A

Requirement Analysis Materials

A.1 Interview Guide

Introductory information for the interviewee

We will have about 30 min together, and we will go through a number of questions. Many of the questions require short answers, but some of them open conversations. We anticipated about 1-min response time for these, if we go over, I might need to remind that we need to move on.

In the course of a Master's thesis conducted at the Department of Geography at the University of Zurich, we aim at examining and enhancing methods for visualizing large origin-destination data such as postal traffic data for analysis. This questionnaire is designed to gather information and opinions from experts in this area on selected topics as presented below. The results are expected to help document how commonly visual analytics methods are used in this domain, how they are perceived by experts, and what could be improved. The participant information will be treated with full confidentiality.

<u>GENERAL INFORMATION</u>

1. Participant information
    (a) Name
    (b) City/Country
    (c) E-Mail
    (d) Gender
2. What is your first language?
3. What is your highest level of education?
4. What is your profession?
5. What is your position within UPU?
6. Since when are you working at UPU?
7. How experienced or familiar are you with information-, geo-, statistical-, network visualization?

<u>EXPERIENCE</u>

1. What is the main goal of your group?
2. What are your group's main projects within UPU?
3. What are the main target groups of your group?
4. What is the main purpose or contribution of data on postal traffic for you/your group? What are you most interested in?
5. Which regions are you mostly focusing on?

6. Which spatial scales regarding postal data are the most relevant ones for you/your group? Why?

   (Global, regional, national, local, other)

7. Which temporal scales regard. postal data are the most relevant ones for you/your group? Why?

   (Decennial, annual, weekly, daily, other)

8. Do you work with the postal data directly or rather with results of analyses conducted by others? How do you work with the data?

9. Which practices are currently used in your group for analyzing postal traffic data?

   (Statistical data analysis, data visualization, spatial/geographical data visualization, none, other)

10. Which practices are currently applied by you for analyzing postal traffic data?

11. In which form does your group publish its analyses?

    (Internal reports (written), public reports (written), online publications, none)

12. How often do you use data on postal traffic for statistical analysis, generating general data visualizations, or geovisualizations?

13. How often do you use results from statistical analysis or (geo-)visualizations by others?

14. What are you interested in detecting in the postal data? (Focus: patterns/anomalies ? what kind of patterns/anomalies?)

15. Which spatial and/or temporal patterns are you aware of in data about postal traffic?

    (e.g.: periodically recurrent peaks (e.g., Christmas), seasonal patterns, strong connections between countries/regions/cities, steady growth/decline of postal traffic over time in specific regions, ??))

16. Why would you call them "patterns"?

17. Which spatial and/or temporal anomalies are you aware of, which anomalies do you expect? Where have you seen anomalies so far?

    (e.g.: short/medium/long peak/disruption, country/regional/local level, disruptions on the way to/from a specific location, ?)

18. How would you define an "anomaly"?

19. What do you expect are the reasons for those anomalies?

20. How long does it take to return to the regular situation/pattern in the postal data/traffic? Are there clear differences? What do you know/expect?

21. What is the impact of natural disasters and/or conflicts on postal traffic?

22. Do you think visualization methods could help to detect and analyze spatial and/or temporal patterns in the data?

23. Why (not)?

24. Do you think visualization methods could help to detect and analyze spatial and/or temporal anomalies in the data?

25. Why (not)?

26. How important is data visualization in your work?

27. Which advantages do you see regarding the visualization of large postal traffic data?

    (e.g. information extraction, complexity reduction, pattern/anomalies detection, ?)

28. Which issues do you see concerning the visualizations?

29. What do you suggest as improvement regarding data visualization for analysis? Please explain.

30. Have you ever seen any of the following data visualization types?

    (Different figures shown on a computer screen)

31. Have you ever used any of the above-mentioned data visualization types?

32. How useful do you find them?

33. If you haven't seen/used such visualizations, do you anticipate that one or several of them could be useful for your work? Why, how?

34. What (other) solutions are you currently working with?

    (e.g. software, analysis type, data (other than postal data), ?)

35. What is your opinion about the solutions you are currently using?

    (perfect, could be improved, could be combined with other solutions, ?)

A.2 Online Questionnaire

On the following pages, the online survey is included as it was shown to the participants (online). It was downloaded directly from surveymonkey.com. The questions marked with a star are compulsory.

AW UPU Questionnaire

General Information

Dear participant,

Welcome and thank you for accepting our invitation.

In the course of a Master's thesis conducted at the Department of Geography at the University of Zurich, we aim at examining and enhancing methods for visualizing large origin-destination data such as postal traffic data for analysis. This questionnaire is designed to gather information and opinions from experts in this area on selected topics as presented below. The results are expected to help document how commonly visual analytics methods are used in this domain, how they are perceived by experts, and what could be improved.

The participant information will be treated with full confidentiality.
If you are willing to be contacted once more towards the end of the thesis to answer a few more questions on the topic, please provide your e-mail address in the beginning of the survery, we would highly appreciate this.

Completing the questionnaire will take you about 15 minutes. For some of the questions, answering is compulsory. However, most often you can also select "other", "none", "I don't know" or "N/A" if you're not sure what to answer.

Many thanks for your participation.

If you would like to receive reports or papers resulting from this study, or for any other questions you may have, please do not hesitate to contact us.

Arzu Coltekin and Anne Wegmann
{arzu.coltekin@geo.uzh.ch, annemarie.wegmann@geo.uzh.ch}
http://www.geo.uzh.ch/~arzu

1. About you:

Name: [ ]

Age (optional): [ ]

City/Country: [ ]

E-mail: [ ]

*** 2. Gender:**

- ○ Female
- ○ Male

3. What is your first language?

- ○ English
- ○ German
- ○ French

Other (please specify) [ ]

4. If English is not your first language, please rate your English level:

- ○ Beginner
- ○ Intermediate
- ○ Proficient

*** 5. Please mark your highest level of education:**

- ○ Basic education
- ○ High school or equivalent
- ○ Bachelor's degree
- ○ Master's degree
- ○ Doctoral degree

*** 6. Please specify your profession.**

[ ]

*** 7. What is your position within UPU?**

[ ]

*** 8. How experienced (knowledge or training) are you in the following categories?**

	1. Not at all	2.	3.	4.	5. Professional/daily exposure
Cartography	○	○	○	○	○
Geovisualizations (e.g. thematic maps)	○	○	○	○	○
Geography related disciplines	○	○	○	○	○
Statistical visualizations	○	○	○	○	○
Network visualisations	○	○	○	○	○
Information visualizations	○	○	○	○	○
Graphic design	○	○	○	○	○

Thank you! We will now move to the next section.

*** 12. Please rank the spatial scales regarding postal data according to their relevance for you/your group (1 = most important, 4 = least important).**

- Global ☐ N/A
- Regional ☐ N/A
- National ☐ N/A
- Local ☐ N/A

*** 13. Please rank the temporal scales regarding postal data according to their relevance for you/your group (1 = most important, 6 = least important).**

- Decennial ☐ N/A
- Annual ☐ N/A
- Monthly ☐ N/A
- Weekly ☐ N/A
- Daily ☐ N/A
- Hourly ☐ N/A

*** 14. Which practices are currently used in your <u>group</u> for analyzing postal traffic data?**

- ☐ Statistical data analysis
- ☐ General data visualization
- ☐ Spatial/geographical data visualization
- ☐ None
- ☐ Other (please specify)

*** 15. Which practices are currently applied by <u>you</u> for analyzing postal traffic data?**

- ☐ Statistical data analysis
- ☐ General data visualization
- ☐ Spatial/geographical data visualization
- ☐ None
- ☐ Other (please specify)

*** 12. Please rank the spatial scales regarding postal data according to their relevance for you/your group (1 = most important, 4 = least important).**

- ☐ Global ☐ N/A
- ☐ Regional ☐ N/A
- ☐ National ☐ N/A
- ☐ Local ☐ N/A

*** 13. Please rank the temporal scales regarding postal data according to their relevance for you/your group (1 = most important, 6 = least important).**

- ☐ Decennial ☐ N/A
- ☐ Annual ☐ N/A
- ☐ Monthly ☐ N/A
- ☐ Weekly ☐ N/A
- ☐ Daily ☐ N/A
- ☐ Hourly ☐ N/A

*** 14. Which practices are currently used in your group for analyzing postal traffic data?**

- ☐ Statistical data analysis
- ☐ General data visualization
- ☐ Spatial/geographical data visualization
- ☐ None
- ☐ Other (please specify)

*** 15. Which practices are currently applied by you for analyzing postal traffic data?**

- ☐ Statistical data analysis
- ☐ General data visualization
- ☐ Spatial/geographical data visualization
- ☐ None
- ☐ Other (please specify)

*** 16. In which form does your group publish its analyses?**

- ☐ Internal reports (written)
- ☐ Public reports (written)
- ☐ Online publications
- ☐ None
- ☐ Other (please specify)

17. If these publications are available online, please provide an URL.

URL 1:

URL 2:

URL 3:

Comments:

*** 18. How often do you use data on postal traffic in one of the following ways?**

	1. Never	2.	3.	4.	5. On a daily basis
Conduct statistical data analysis myself	○	○	○	○	○
Generate general data visualizations myself	○	○	○	○	○
Generate spatial/geographical data visualizations myself	○	○	○	○	○
Use statistical results from data analysis for written reports	○	○	○	○	○
Use (geo-) data visualizations for written reports	○	○	○	○	○
Use statistical results from data analysis for online applications	○	○	○	○	○
Use (geo-) data visualizations for online applications	○	○	○	○	○
Use the data for detecting patterns and/or anomalies in the data	○	○	○	○	○

Comments

19. Regarding patterns and anomalies: What are you interested in detecting in the postal traffic data?

*** 20. Which spatial and/or temporal patterns are you aware of in data about postal traffic? Please explain the patterns in the *Comments* field.**

- ☐ Periodically recurrent peaks (e.g., Christmas)
- ☐ Seasonal patterns
- ☐ Strong connections between specific countries
- ☐ Strong connections between specific regions
- ☐ Strong connections between specific cities
- ☐ Steady growth/decline of postal traffic over time
- ☐ Other (please specify)
- ☐ I don't know
- ☐ Comments:

*** 21. Which spatial and/or temporal anomalies are you aware of in data about postal traffic? Please explain the anomalies in the *Comments* field.**

- ☐ Short peak/disruption (1-2 days)
- ☐ Medium peak/disruption (3 days - 1 week)
- ☐ Long peak/disruption (longer than 1 week)
- ☐ Anomalies/disruptions on country level
- ☐ Anomaliesy/disruptions on regional level
- ☐ Anomalies/disruptions on local level
- ☐ Disruptions on the way to a specific location
- ☐ Disruptions on the way from a specific location
- ☐ Other (please specify)
- ☐ I don't know
- ☐ Comments:

22. What do you expect are the reasons for those anomalies? Please explain.

* **23. Do you think visualization methods could help to detect and analyse spatial and/or temporal <u>patterns</u> in the data?**

- ◯ Yes
- ◯ No
- ◯ I don't know

* **24. Do you think visualization methods could help to detect and analyse spatial and/or temporal <u>anomalies</u> in the data?**

- ◯ Yes
- ◯ No
- ◯ I don't know

* **25. Which advantages do you see regarding the visualization of large postal traffic data?**

- ☐ Information extraction
- ☐ Complexity reduction
- ☐ Pattern and anomalies detection
- ☐ Communication
- ☐ Other (please specify)

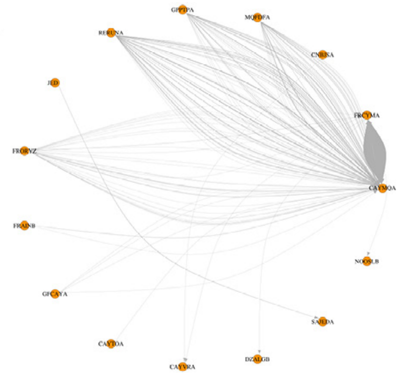
26. Which problems do you see concerning the visualizations? Please describe.

27. What do you suggest as improvement regarding data visualization for analysis? Please explain.

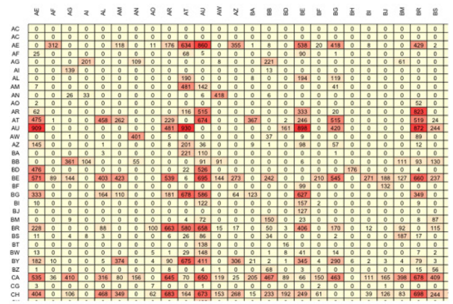
* **28. Have you ever <u>seen</u> any of the following data visualization <u>types</u>? If yes, please proceed to question 29. If no, please proceed to question 30.**

- ◯ Yes
- ◯ No

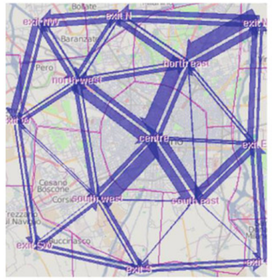
a



b

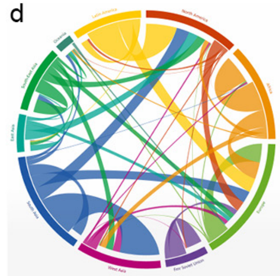


c



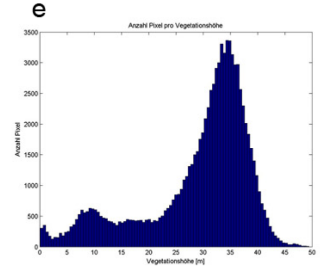
Source: Andrienko, G. & Andrienko, N. (2010): A General Framework for Using Aggregation in Visual Exploration of Movement Data. The Cartographic Journal, 47 (1), 22-40.

d



Source: http://www.global-migration.info/

e



29. Have you ever <u>used</u> any of the above-mentioned data visualization <u>types</u>? If yes, how useful do you find them? If no, please proceed to question 30.

	1. Not useful at all	2.	3.	4.	5. Very useful	N/A
a	☐	☐	☐	☐	☐	☐
b	☐	☐	☐	☐	☐	☐
c	☐	☐	☐	☐	☐	☐
d	☐	☐	☐	☐	☐	☐
e	☐	☐	☐	☐	☐	☐

Please explain for which tasks in your work routine these visualization type(s) could be useful.

30. If you are not used to work with data visualizations of the types seen above, do you anticipate that one or several of them could be useful for your work? Please specify.

31. What other solutions are you currently working with?

Software: ______

Data analysis type: ______

Data (including data other than postal traffic data): ______

Other (please specify): ______

32. What is your opinion about the solutions you are currently using?

- ◯ I am completely happy with the current solutions, I don't need anything else for my work.
- ◯ I can work well with the current solution, but I have some ideas about how they could be improved (please specify below).
- ◯ I would be happy to have other tools at hand to work with, or to combine with currently used solutions (please specify below).
- ◯ Comments:

Thank you for participating! Have a nice day.

Appendix B

EmIS Reports

Figure B.1 depicts the spreadsheet in which the EmIS reports were prepared for further use and categorisation. Figure B.2 shows a visualisation of the reported EmIS events per category.

EmIS = Emergency Information System

EmIS No.	EmIS entr	Start Date	End Date	Country	Post Code	Category	Reason	Comments	Related Em	First entry
13	2014-01-24	2014-01-22	2014-01-23			NAT	Snow storm	[...] return to normal dispatch procedures for mail to JFK International Airport in New York (OEJFK), effective Thursday, 23 January 2014.	12	N
14	2014-01-29	2013-12-24	2014-01-29	Suriname		SOC	Strike	[...] the strike by some of the country's postal workers has officially ended and all mail operations have returned to normal.	100 (2013), 2 (2014)	N
15	2014-02-03	2014-01-31	2014-02-26	Slovenia		NAT	Snow storm	On Sunday, 2 February, the Slovenian Environment Agency (ARSO) issued a red alert for the whole country, signalling a high risk of major damage to property and danger to human lives. Traffic in the country is hampered and some roads have been completely blocked by fallen trees. Also, many areas are without power due to broken power lines. ARSO has also issued an orange alert for the northwest of the country for Monday, 3 and Tuesday, 4 February. As a result of this inclement weather, mail operations have been completely paralyzed in certain areas, and severely hampered in others.	28	Y
16	2014-02-05	2014-02-03	2014-02-09	USA	New York Exchange Office **(USJFK)** from **3 to 9 February** Chicago International Service Centre **(USORD)** from **5 to 9 February**	NAT	Snow storm	USPS is experiencing major transportation and operational delays, as a result of winter storms currently affecting the Eastern Seaboard of the United States. Snowfalls of up to 25 cm, freezing rain and ice have led to the cancellation of thousands of flights.	17, 18	Y
17	2014-02-07	2014-02-05	2014-02-09	USA	New York Exchange Office **(USJFK)** from **3 to 9 February** Chicago International Service Centre **(USORD)** from **5 to 9 February**	NAT	Snow storm	[...] winter storms are now also affecting mid-western states. Heavy snow and poor visibility have led to the cancellation of hundreds of flights and have negatively impacted postal operations at Chicago International Service Centre (USORD).	16, 18	N
18	2014-02-11	2014-02-05	2014-02-09	USA	New York Exchange Office **(USJFK)** from **3 to 9 February** Chicago International Service Centre **(USORD)** from **5 to 9 February**	NAT	Snow storm	[...] the winter storms that severely affected the New York Exchange Office (USJFK) from 3 to 9 February, and Chicago International Service Centre (USORD) from 5 to 9 February, have ended, and normal services have been restored in the impacted areas.	16, 17	N

FIGURE B.1: Screenshot of the EmIS collection in Microsoft Excel.

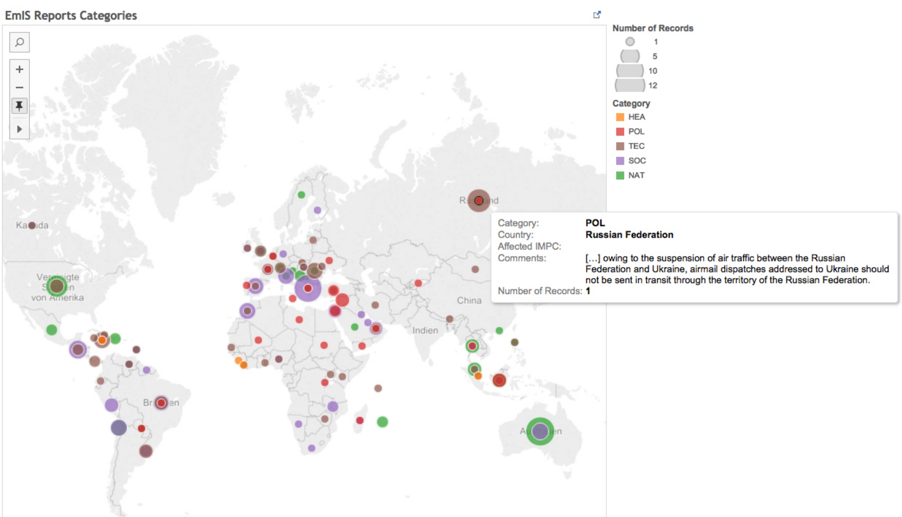


FIGURE B.2: Screenshot of the EmIS visualization in Tableau

Appendix C

Examples of R code used for creating the visualisations

```
# -------------------------------------------------------------------
# ODM
# -------------------------------------------------------------------

f.odm.vis.logcol.OUT <- function(odm, odm1, threshold, threshold2, titleName){ # odm= real values, odm1 = log values for color
  library(plotrix)
  par(mar = c(17, 13, 18, 5)) # unten, links, oben, rechts
  odm[odm <= threshold] <- NA # fill all cells < threshold with NA
  odm <- odm[rowSums(is.na(odm)) != ncol(odm), colSums(is.na(odm)) != nrow(odm)] # delete cols with NA
  odm1[odm1 <= threshold2] <- NA # fill all cells < threshold with NA
  odm1 <- odm1[rowSums(is.na(odm1)) != ncol(odm1), colSums(is.na(odm1)) != nrow(odm1)] # delete cols with NA
  color2D.matplot(odm,
                  show.values = T,
                  axes = FALSE,
                  xlab = "",
                  ylab = "",
                  cellcolors = color.scale(odm1, c(260),c(30, 80),c(100, 60),
                                           color.spec="hcl",xrange=c(threshold2,log(29030))), # color.scale-Werte: c(min), c(middle), c(max)
                  vcex = 2,
                  vcol = "black",
                  pty = "m",
                  Hinton=F
  )
  title(ylab="Origin", line=+8, cex.lab=2.5)
  title(xlab="Destination", line=-76, cex.lab=2.5)
  title(titleName, line = +14, cex.main=3)

  axis(3, at = seq_len(ncol(odm)) -0.5,
       labels = colnames(odm), tick = FALSE, las = 2, cex.axis = 2.2)
  axis(2, at = seq_len(nrow(odm)) -0.5,
       labels = rev(rownames(odm)), tick = FALSE, las = 2, cex.axis = 2.2)
  testcol = color.scale(c(log(1.1), log(10), log(100), log(1000), log(10000), log(29030)),c(260),c(30, 80),c(100, 60), color.spec="hcl")
  col.labels<-c("0","540308")
  color.legend(0.5,-1,8,-0.7,col.labels,testcol, cex = 2)
}
```

FIGURE C.1: Example code for the ODM

```
checkDateLine_US.CNTRY.IMPC.OUT <- function(l){
  n<-0
  k<-length(l)
  k<-k-1
  for (j in 1:k){
    n[j] <- l[j+1] - l[j]
  }
  n <- abs(n)
  m<-max(n, rm.na=TRUE)
  ifelse(m > 30, TRUE, FALSE)
}
clean.Inter_US.CNTRY.IMPC.OUT <- function(p1, p2, n, addStartEnd){
  inter <- gcIntermediate(p1, p2, n=n, addStartEnd=addStartEnd)
  if (checkDateLine_US.CNTRY.IMPC.OUT(inter[,1])){
    m1 <- midPoint(p1, p2)
    m1[,1] <- (m1[,1]+180)%%360 - 180
    a1 <- antipode(m1)
    l1 <- gcIntermediate(p1, a1, n=n, addStartEnd=addStartEnd)
    l2 <- gcIntermediate(a1, p2, n=n, addStartEnd=addStartEnd)
    l3 <- rbind(l1, l2)
    l3
  }
  else{
    inter
  }
}
add_lines_US.CNTRY.IMPC.OUT <- function(freq.table, nodes, nodesCNTRY, maxFreq){
  pal <- colorRampPalette(c("#8ac7d5", "#123037")) # only blue
  colors <- pal(nrow(nodes))
  fsub <- freq.table[order(freq.table$Freq),]
  maxfreqlog <- max(fsub$Freq.log1)
  for (j in 1:nrow(fsub)) {
    u1 <- nodes[nodes$Code == fsub[j,]$Source,]
    u2 <- nodesCNTRY[nodesCNTRY$ISO2 == fsub[j,]$Target,]
    p1 <- c(u1[1,]$lng, u1[1,]$lat)
    p2 <- c(u2[1,]$LON, u2[1,]$LAT)
    inter <- clean.Inter_US.CNTRY.IMPC.OUT(p1,p2,n=100, addStartEnd=TRUE)
    colindex <- round((fsub[j,]$Freq.log1) * (length(colors)/15))
    lwdFreq <- (fsub[j,]$Freq/(maxFreq/10))
    lines(inter, col=alpha(colors[colindex], 0.7), lwd=lwdFreq, lend = 2) # amount = line width
  }
}
map_world_US.CNTRY.IMPC.OUT <- function(freq.table, nodes, nodesCNTRY, title, maxFreq, threshold){
  library(maps)
  library(geosphere)
  library(scales)
  freq.table <- subset(freq.table, freq.table$Freq > threshold)
  par(mar = c(0, 0, 8, 0)) |
  plot(worldmap, col="#f2f2f2",  bg="white", lwd=0.05, main = title, cex.main=4)
  legend.minline <- (min(freq.table$Freq)/(maxFreq/10))
  legend.maxline <- (maxFreq/(maxFreq/10))
  legend("topright", inset=.05, title="Postal flows from US IMPCs to Countries", cex=1.5,
         c("100 items", "29030 items (maximum over all years)"), horiz=F, lty=c(1,1), bty="o",
         lwd=c(legend.minline, legend.maxline),col=alpha(c("#8ac7d5", "#123037"),0.7))
  add_lines_US.CNTRY.IMPC.OUT(freq.table, nodes, nodesCNTRY, maxFreq)
}
```

FIGURE C.2: Example code for the flow map for LOD 2

Appendix D

Requirement Analysis Results

Additional results of the online survey conducted for the requirement analysis are displayed below.

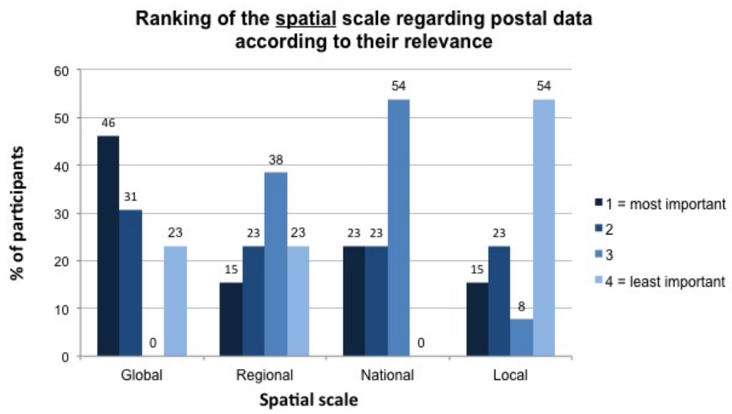


FIGURE D.1: Ranking of spatial scale regarding postal data, according to their relevance

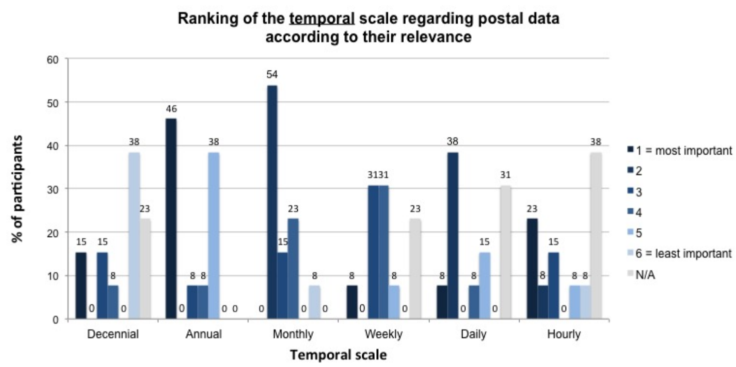


FIGURE D.2: Ranking of temporal scale regarding postal data, according to their relevance

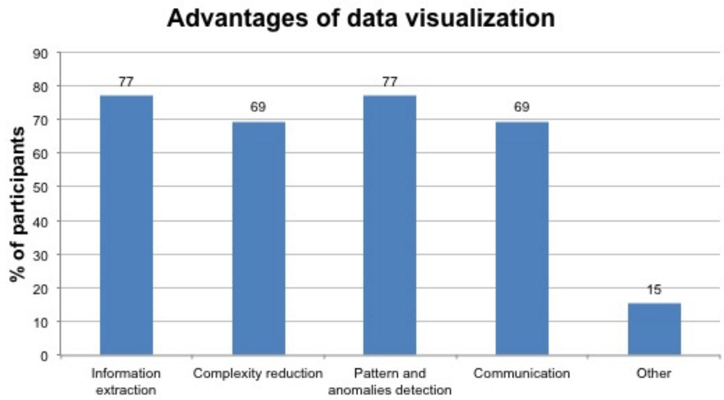


FIGURE D.3: Advantages of data visualisation

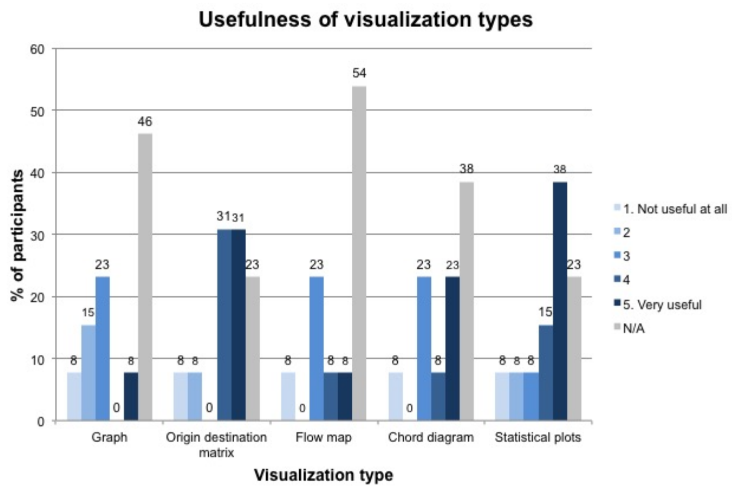


FIGURE D.4: Opinion about the usefulness of visualisation types

Appendix E

Usability Study

E.1 Setup of the usability study

FIGURE E.1: Setup Usability Study

E.2 Study Design

This section depicts the order of the visualisations shown to participants I and II. Participant III started with "SECTION 3", participant IV again with "SECTION 1'.

TABLE E.1: Examples of usability study structure

Participant IV started again with "SECTION 1', as did participant I. The visualisations and data sets were rotated within the sections, thus participant I and IV looked at the them in a different order.

Participant I	**Participant II**
Preliminary questionnaire	**Preliminary questionnaire**
SECTION 1: LOD 1 (Country)	**SECTION 2: LOD 2 (IMPC)**
Chile IN; Flow Map	US OUT; OD Matrix
Chile IN; Choropleth Map	US OUT; Flow Map
US OUT; Choropleth Map	US OUT; Graduated Symbol Map
US OUT; Flow Map	Chile IN; Flow Map
Chile OUT; Flow Map	Chile IN; Graduated Symbol Map
Chile OUT; Choropleth Map	US IN; Flow Map
US IN; Choropleth Map	US IN; OD Matrix
US IN; Flow Map	US IN; Graduated Symbol Map
Questionnaire for section 1	Chile OUT; Graduated Symbol Map
SECTION 2: LOD 2 (IMPC)	Chile OUT; Flow Map
US OUT; OD Matrix	**Questionnaire for section 2**
US OUT; Flow Map	**SECTION 3: Time Series**
US OUT; Graduated Symbol Map	US OUT; Time Series, split into IMPCs
Chile IN; Flow Map	US OUT; Time Series
Chile IN; Graduated Symbol Map	Chile IN; Time Series
US IN; Flow Map	US IN; Time Series
US IN; OD Matrix	US IN; Time Series, split into IMPCs
US IN; Graduated Symbol Map	Chile OUT; Time Series
Chile OUT; Graduated Symbol Map	**Questionnaire for section 3**
Chile OUT; Flow Map	**SECTION 1: LOD 1 (Country)**
Questionnaire for section 2	Chile IN; Flow Map
SECTION 3: Time Series	Chile IN; Choropleth Map
US OUT; Time Series, split into IMPCs	US OUT; Choropleth Map
US OUT; Time Series	US OUT; Flow Map
Chile IN; Time Series	Chile OUT; Flow Map
US IN; Time Series	Chile OUT; Choropleth Map
US IN; Time Series, split into IMPCs	US IN; Choropleth Map
Chile OUT; Time Series	US IN; Flow Map
Questionnaire for section 3	**Questionnaire for section 1**
SUS Questionnaire	**SUS Questionnaire**
Semi-structured interview	**Semi-structured interview**

TABLE E.2: Previous experience and knowledge of participants

Experience with tools and data (1 = none, 5 = expert)

Participant	Non-geo. inform. vis.	Geogr. inform. vis.	GIS	ArcGIS Online	Tableau	R	Spatial data	Temp. data	Spatio-temp. data	Postal data (EDI)
I	4	4	4	4	2	3	4	4	3	5
II	3	2	2	1	1	4	2	3	2	3
III	2	3	1	1	1	1	1	1	1	2
IV	3	3	3	3	3	3	3	3	3	3
V	3	4	1	1	1	1	1	1	1	3
VI	3	2	1	2	1	1	1	1	1	1

Familiarity with tools (1 = never work with, 5 = daily use)

Participant	GIS/ ArcGIS Online	Tableau	R	Spatial data	Temp. data	Spatio-temp. data	Postal data (EDI)
I	3	1	3	4	4	4	5
II	1	1	3	1	2	1	3
III	1	1	1	1	1	1	1
IV	2	2	1	1	1	1	1
V	1	1	1	1	1	3	3
VI	2	1	1	1	1	1	1

E.3 Pre-Experiment Questionnaire

The participants were asked about their previous experience and knowledge regarding different visualisation tools and data types. They could choose between 1 (no experience) and 5 (expert) and 1 (never work with) and 5 (daily use), respectively. Results are displayed in table E.2.

E.4 System Usability Scale (SUS) Questionnaire

The following pages show the SUS questionnaire as it was presented (online) to the participants after completing the tasks with the visualisations. It was downloaded directly from surveymonkey.com.

AW UPU Usability Study

Post-experiment questionnaire (SUS)

* 11. I think I would like to use this visualization frequently

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	◯	◯	◯	◯	◯
Flow Map	◯	◯	◯	◯	◯
Choropleth Map	◯	◯	◯	◯	◯
Counts Map	◯	◯	◯	◯	◯
Time Series	◯	◯	◯	◯	◯

* 12. I found the visualization unnecessarily complex

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	◯	◯	◯	◯	◯
Flow Map	◯	◯	◯	◯	◯
Choropleth Map	◯	◯	◯	◯	◯
Counts Map	◯	◯	◯	◯	◯
Time Series	◯	◯	◯	◯	◯

* 13. I thought the visualization was easy to use

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	◯	◯	◯	◯	◯
Flow Map	◯	◯	◯	◯	◯
Choropleth Map	◯	◯	◯	◯	◯
Counts Map	◯	◯	◯	◯	◯
Time Series	◯	◯	◯	◯	◯

* 14. I think that I would need the support of a technical person to be able to use this visualization

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	◯	◯	◯	◯	◯
Flow Map	◯	◯	◯	◯	◯
Choropleth Map	◯	◯	◯	◯	◯
Counts Map	◯	◯	◯	◯	◯
Time Series	◯	◯	◯	◯	◯

* 15. I found this visualization type very useful for detecting patterns and anomalies

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	◯	◯	◯	◯	◯
Flow Map	◯	◯	◯	◯	◯
Choropleth Map	◯	◯	◯	◯	◯
Counts Map	◯	◯	◯	◯	◯
Time Series	◯	◯	◯	◯	◯

* 16. Finding patterns and anomalies in the data was very challenging with this visualization type

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	◯	◯	◯	◯	◯
Flow Map	◯	◯	◯	◯	◯
Choropleth Map	◯	◯	◯	◯	◯
Counts Map	◯	◯	◯	◯	◯
Time Series	◯	◯	◯	◯	◯

* 17. I would imagine that most people would learn to use this visualization very quickly

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	○	○	○	○	○
Flow Map	○	○	○	○	○
Choropleth Map	○	○	○	○	○
Counts Map	○	○	○	○	○
Time Series	○	○	○	○	○

* 18. I found the visualization very cumbersome to use

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	○	○	○	○	○
Flow Map	○	○	○	○	○
Choropleth Map	○	○	○	○	○
Counts Map	○	○	○	○	○
Time Series	○	○	○	○	○

* 19. I felt very confident using the visualization

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	○	○	○	○	○
Flow Map	○	○	○	○	○
Choropleth Map	○	○	○	○	○
Counts Map	○	○	○	○	○
Time Series	○	○	○	○	○

* 20. I need to learn a lot of things before I could get going with this visualization

	1. Strongly disagree	2.	3.	4.	5. Strongly agree
Origin-Destination Matrix	○	○	○	○	○
Flow Map	○	○	○	○	○
Choropleth Map	○	○	○	○	○
Counts Map	○	○	○	○	○
Time Series	○	○	○	○	○

E.5 Post-Experiment Interview

1. If you were asked to choose one of the visualization types, which one would you choose (overall) and why?
   (a) Would you choose the same visualization type for the specific tasks (patterns/anomalies) that you solved in this experiment? Why?
2. How would you order (rank) the visualizations in terms of your preference? Why?
   (a) How would you order/rank them per task type?
3. If you need to combine 2 or 3, which ones would you choose?
   (a) How would you combine them per task type?
4. Do you overall prefer visualizations over tables, or other forms of reporting such as text?
   (a) Do you *always* prefer visualizations over tables or other forms of reporting such as text?
   (b) Would you prefer visualizations over tables or other forms of reporting for this type of tasks?
5. Which visualization was most pleasing to work with (aesthetically)?
6. Did you encounter any difficulties with any of the visualizations and why? What were those difficulties per visualization type you worked with?
   (a) Were these difficulties the same for all tasks or are they task-specific? Why?
7. Do you prefer static or interactive visualizations for examining the data overall? Why?
   (a) Do you prefer static/interactive for the particular tasks you solved in this experiment? Why?
8. For what other type of questions (in relation to your expertise) do you think these visualization types might be useful? Is a specific visualization type fitting well to a specific task type? Why do you think so?
9. Do you think the spatial dimension helped you for detecting patterns and anomalies? Did it add valuable information to you? In which way?
10. Do you have any general remarks?

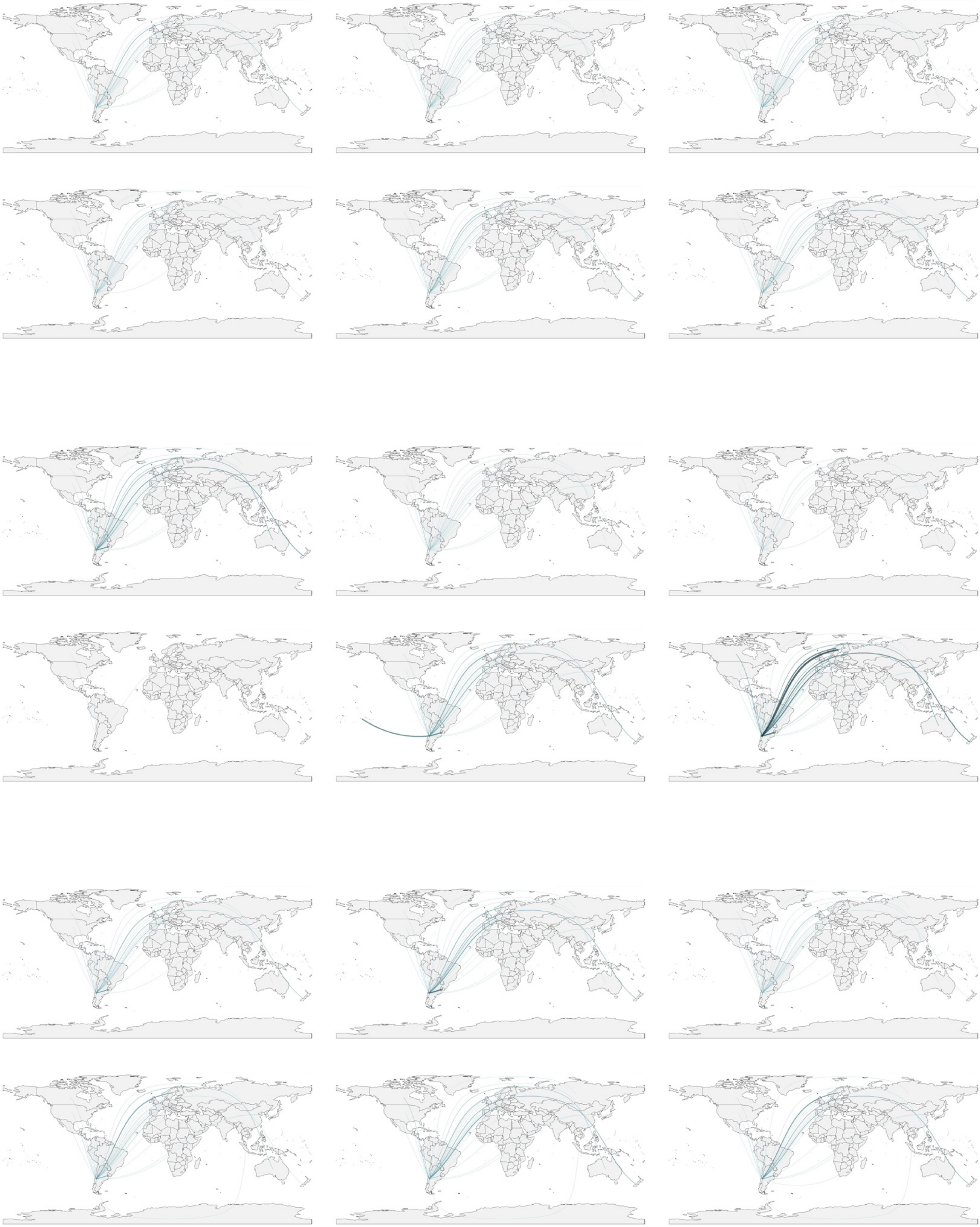
Appendix F

Thumbnails of Visualisations

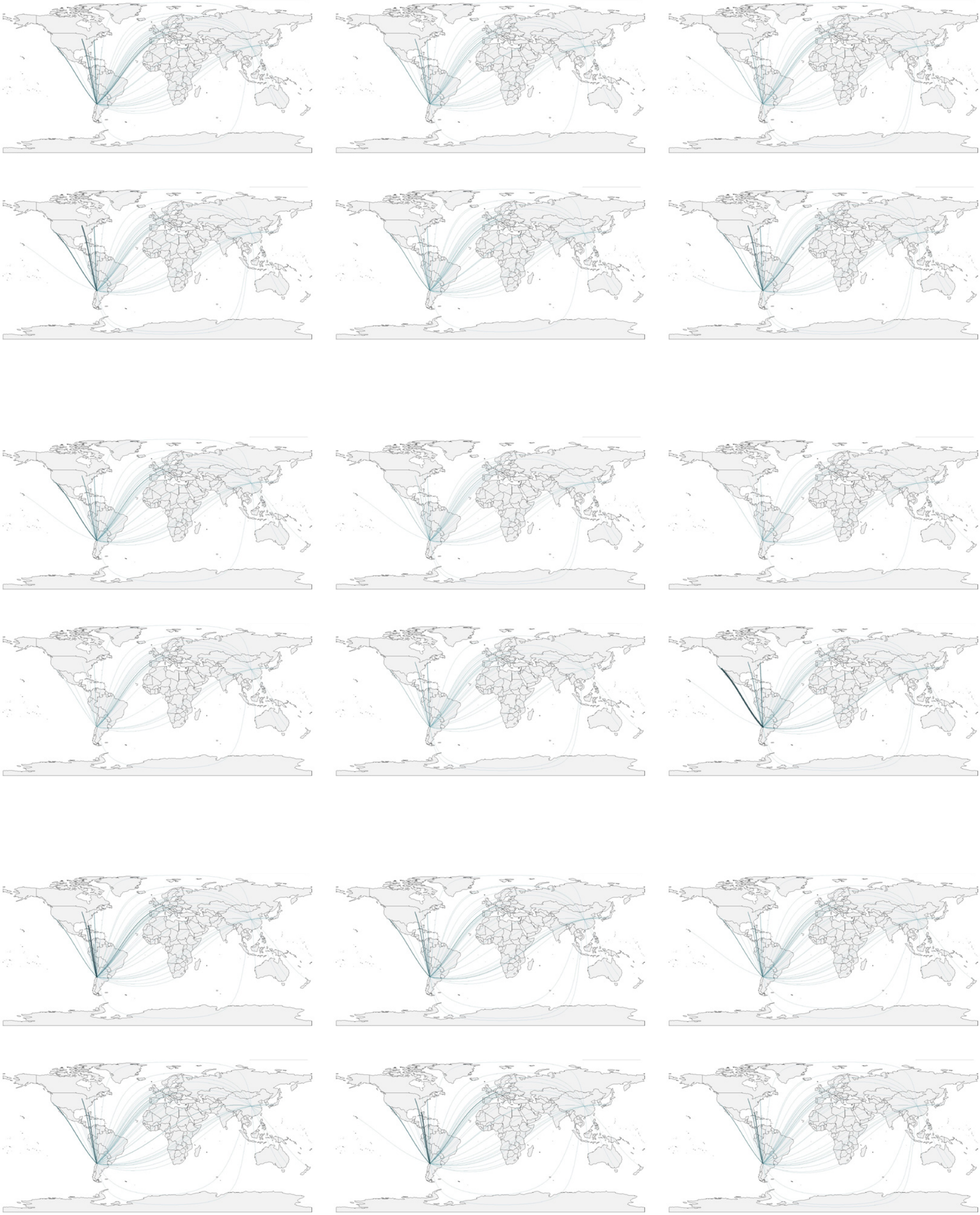
The following pages include thumbnail images of all the visualisations created from the postal data. Legends that were shown to the participants in the usability study were excluded for readability reasons.

CL inbound LOD 1. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)

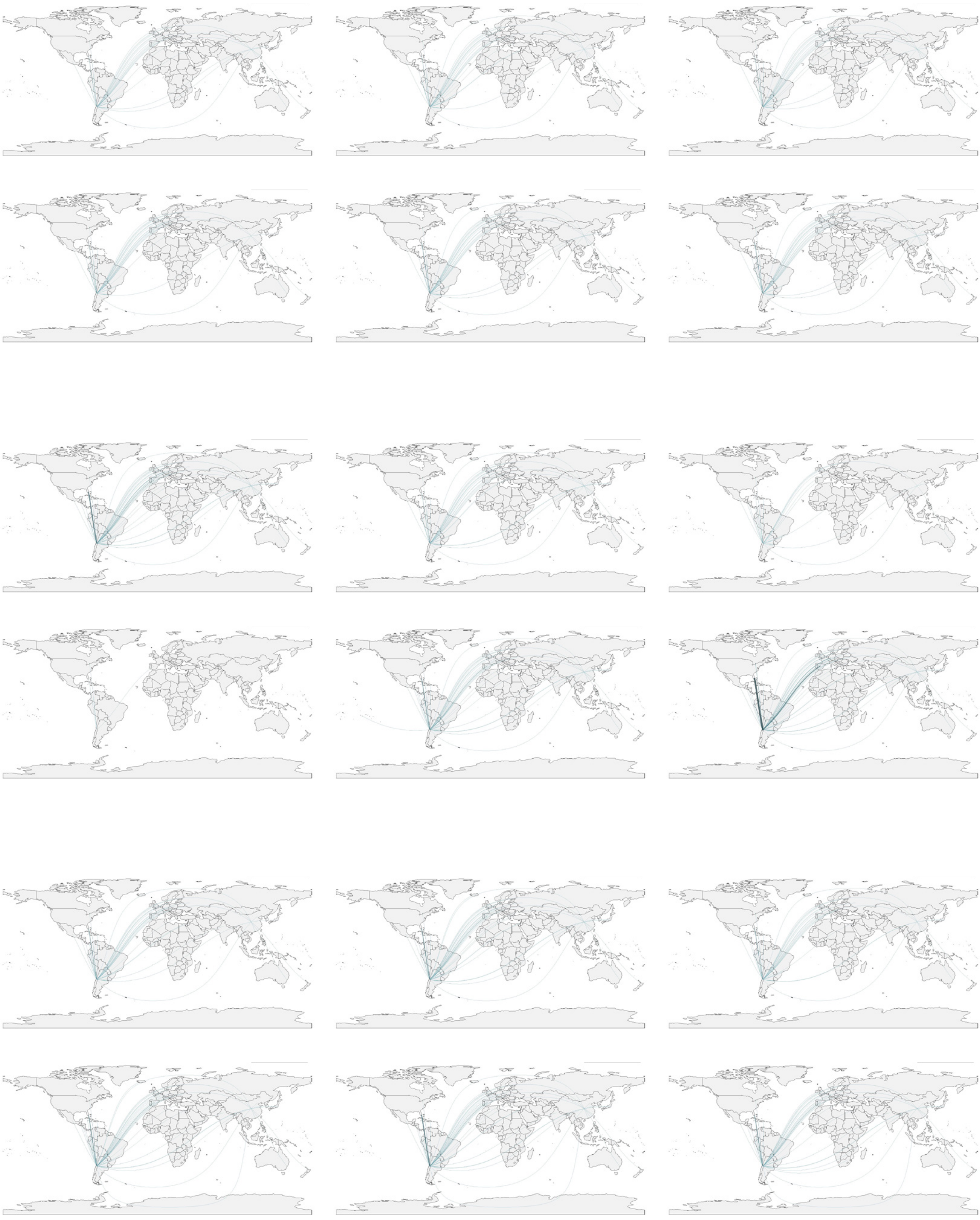
CL outbound LOD 1. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



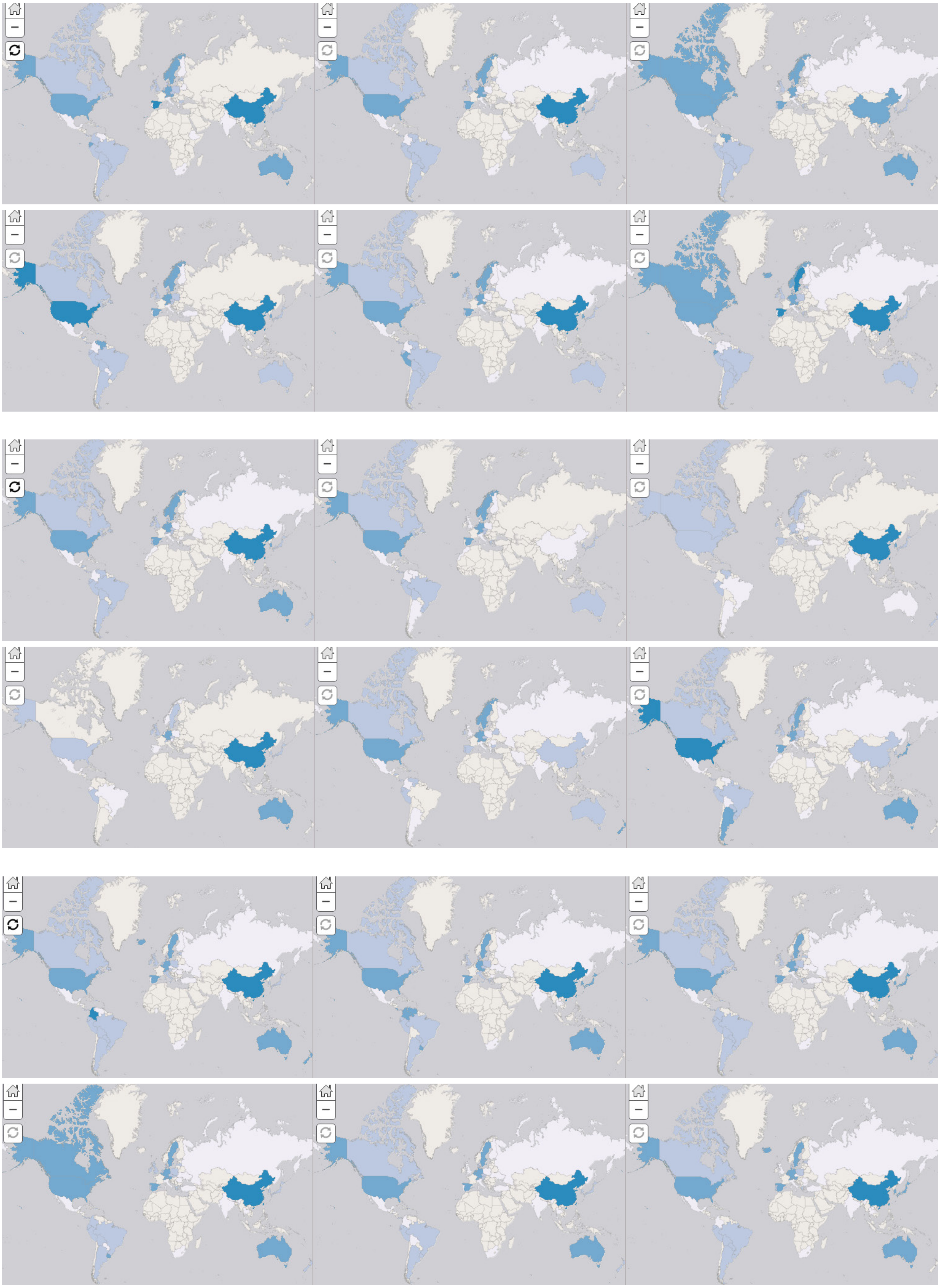
CL inbound LOD 2. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



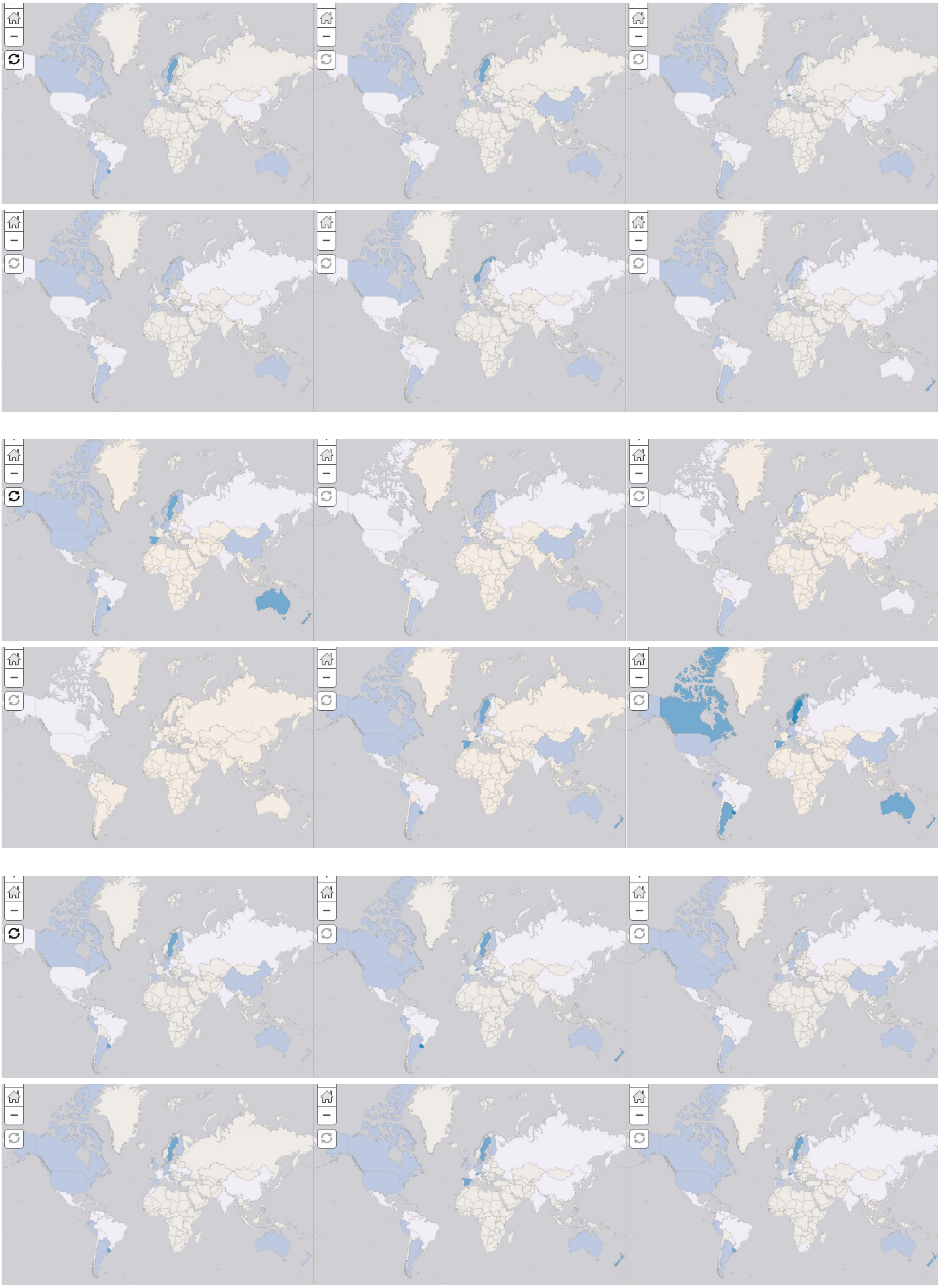
CL outbound LOD 2. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



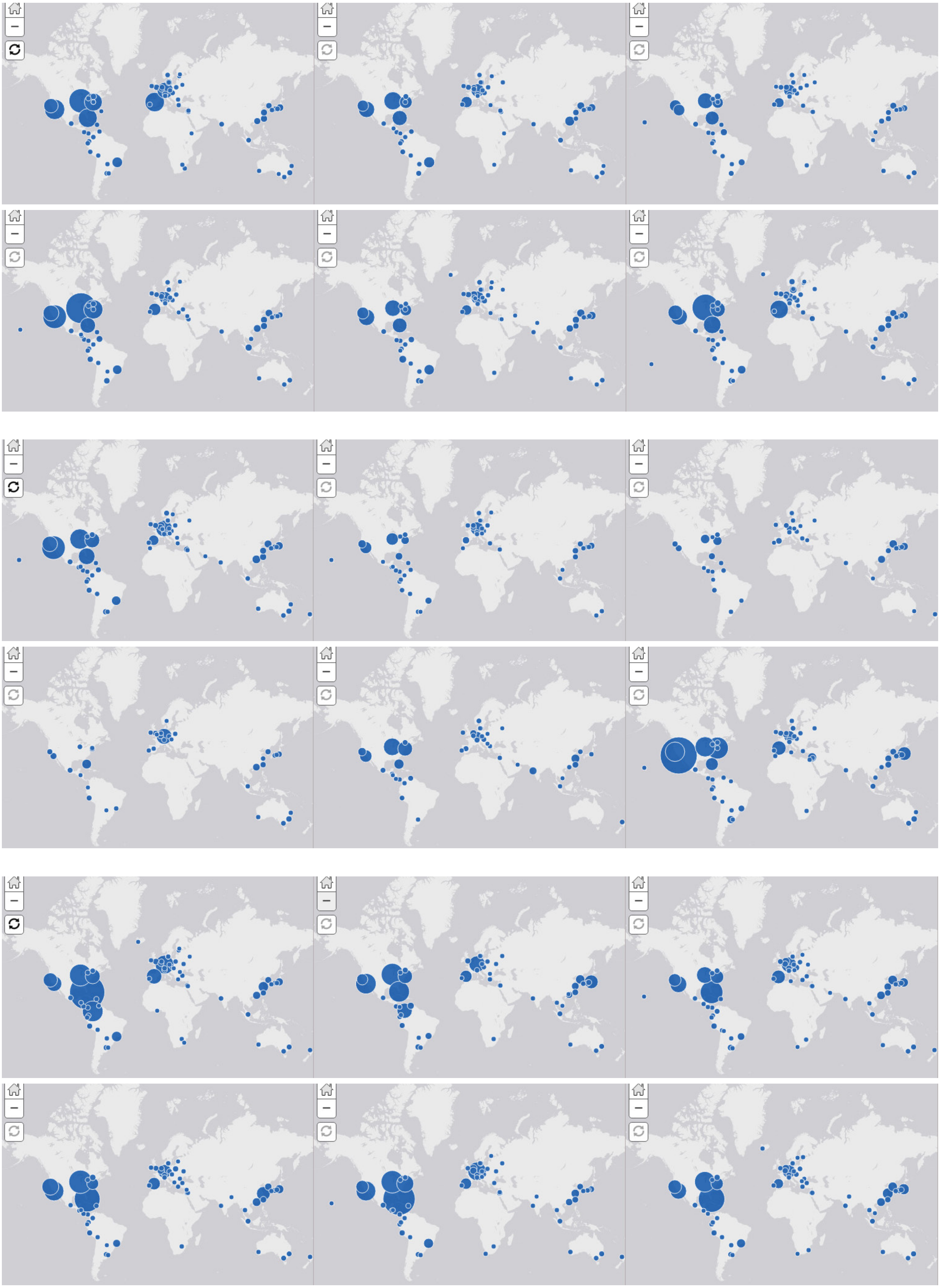
CL inbound LOD 1. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



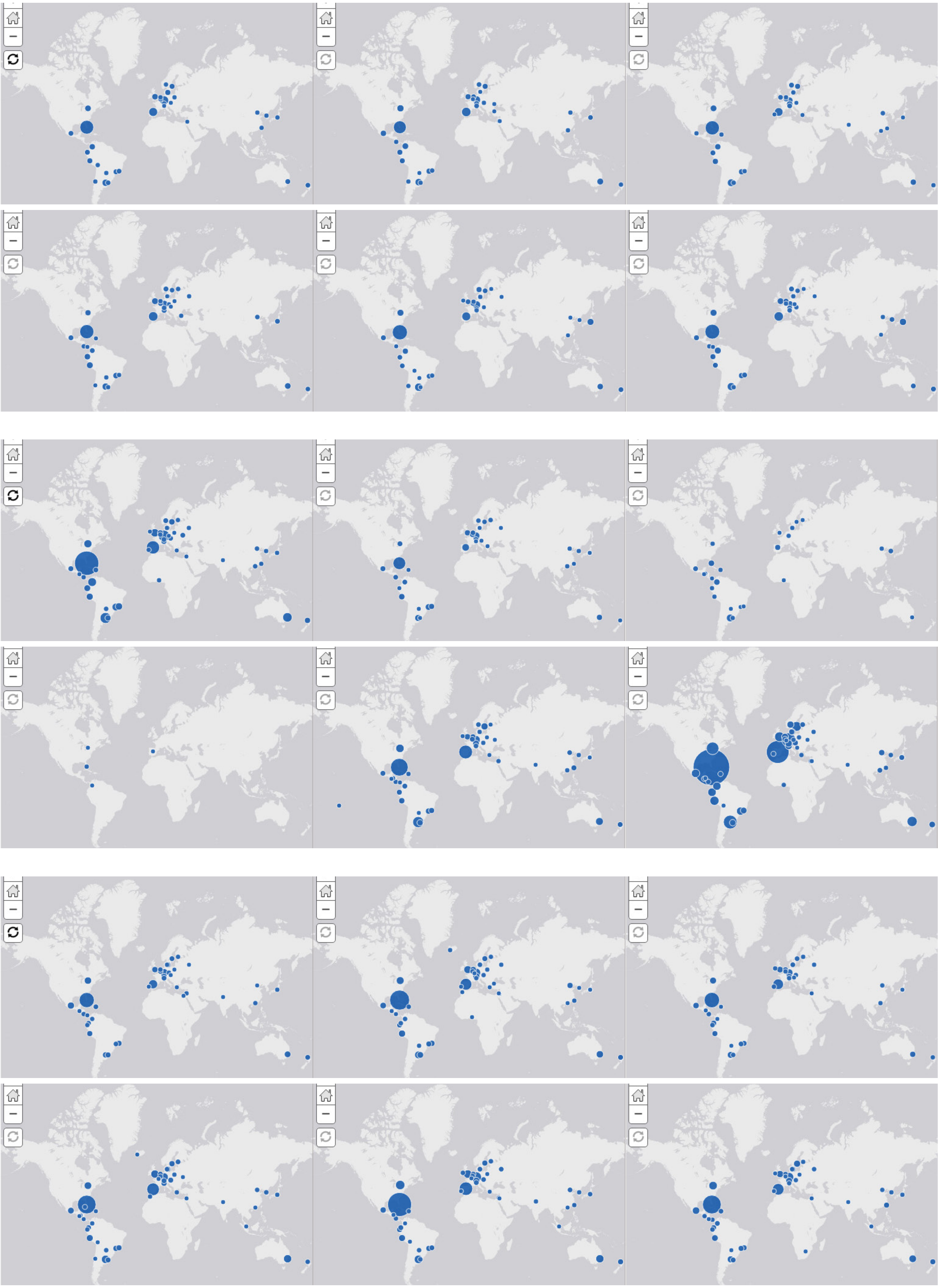
CL outbound LOD 1. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



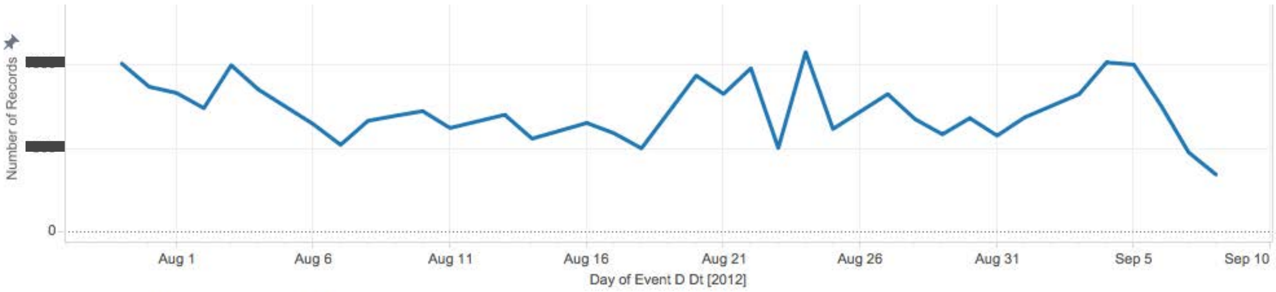
CL inbound LOD 2. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



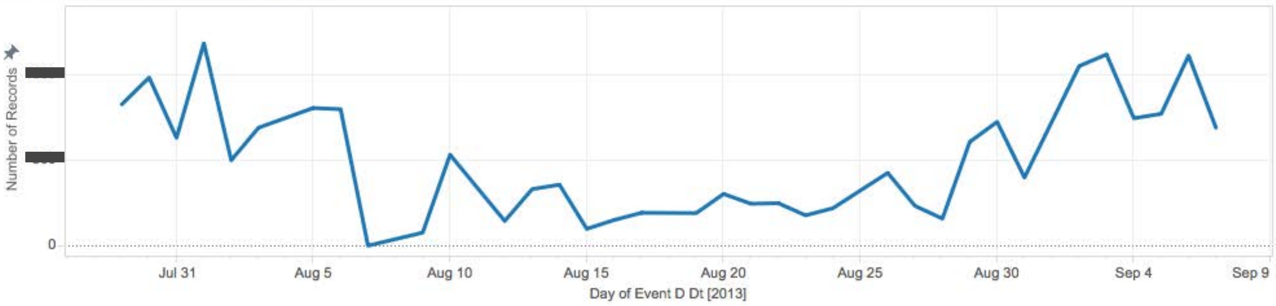
CL outbound LOD 2. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



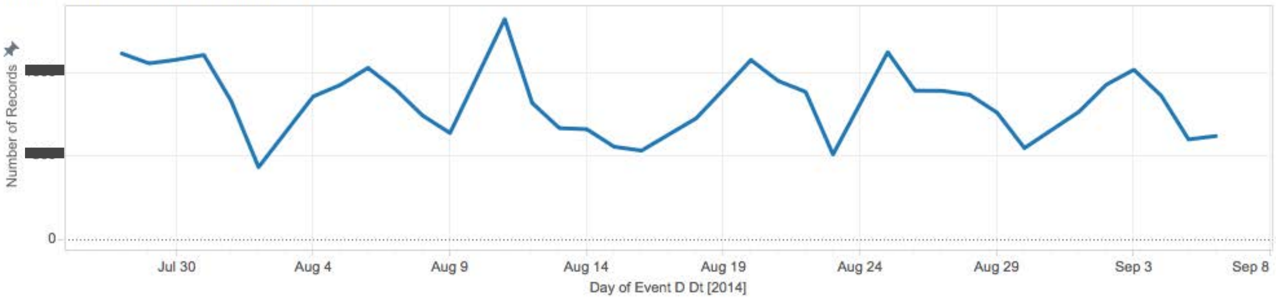
CL inbound LOD 1. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)



Items sent to Chile, Event D, 2013

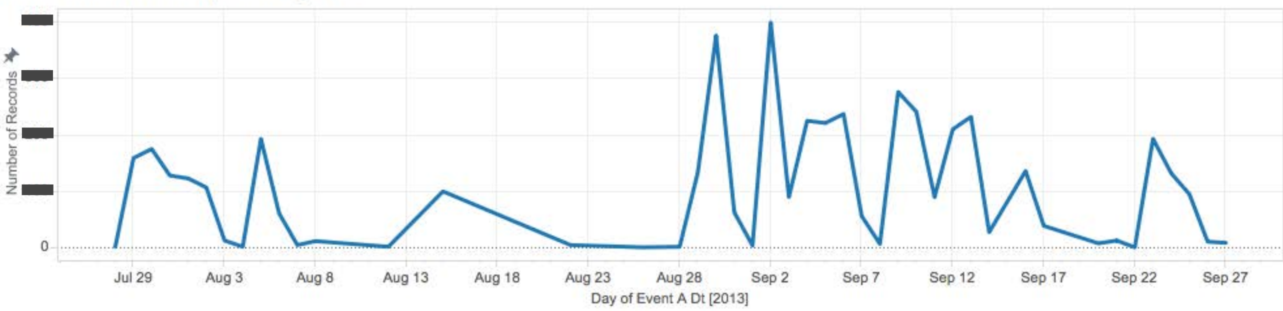


Items sent to Chile, Event D, 2014

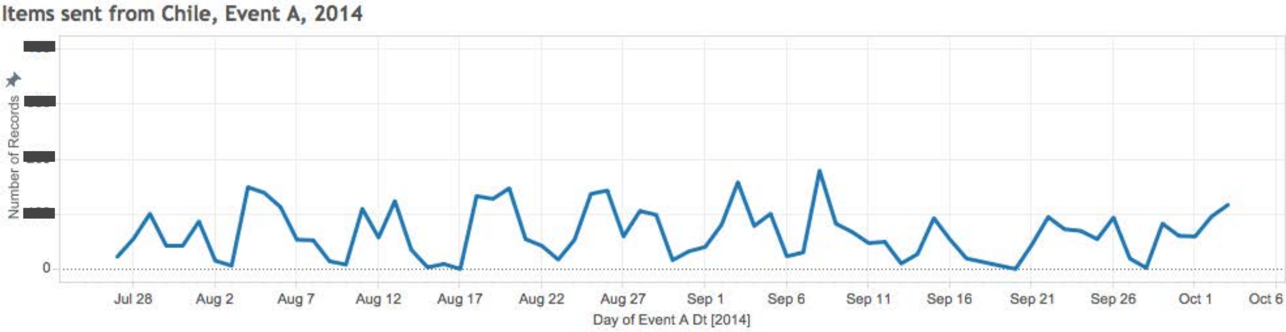


CL outbound LOD 1. CW31-36 2012 (top), 2013 (middle), 2014 (bottom)

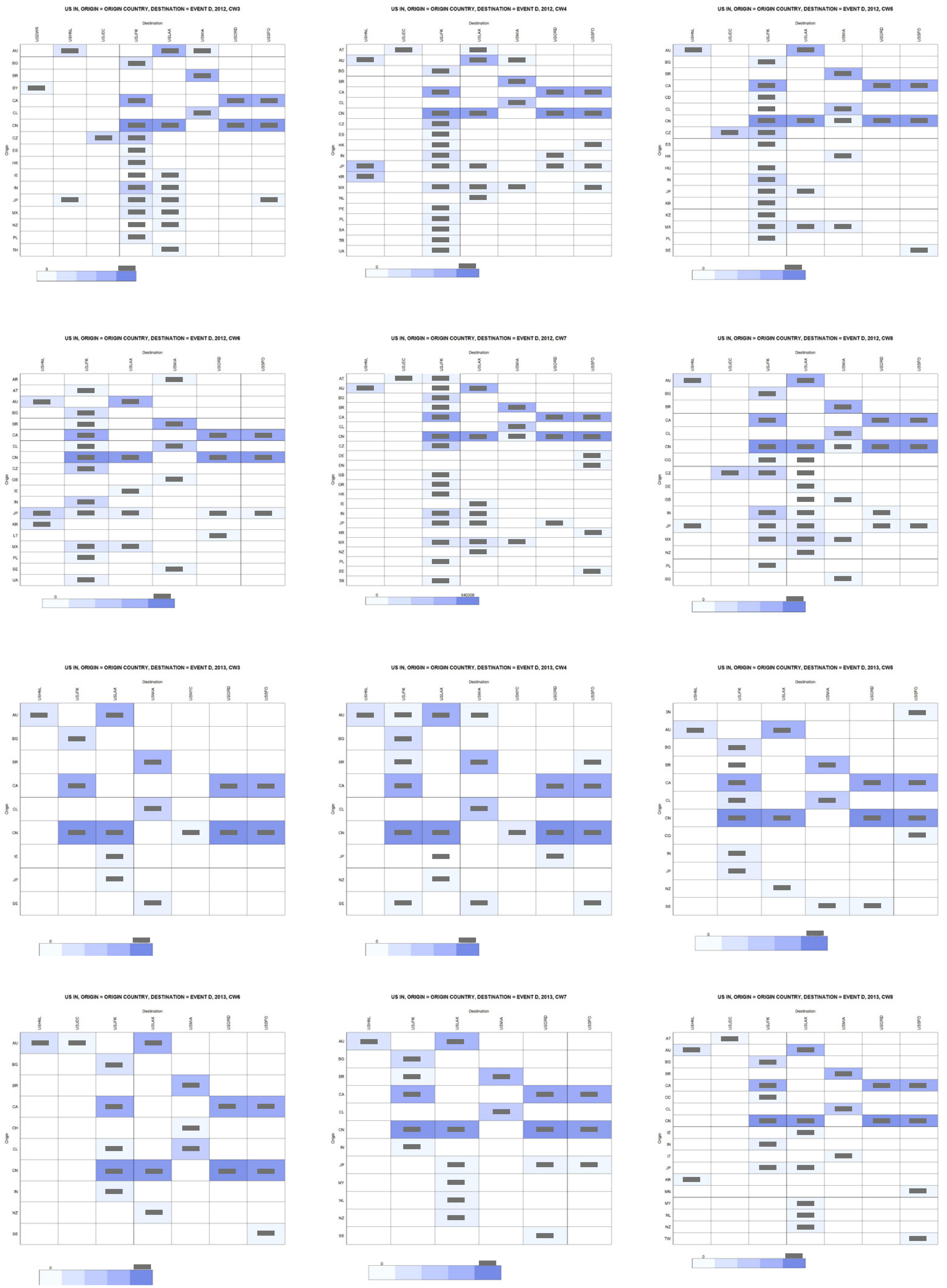
Items sent from Chile, Event A, 2013



Items sent from Chile, Event A, 2014

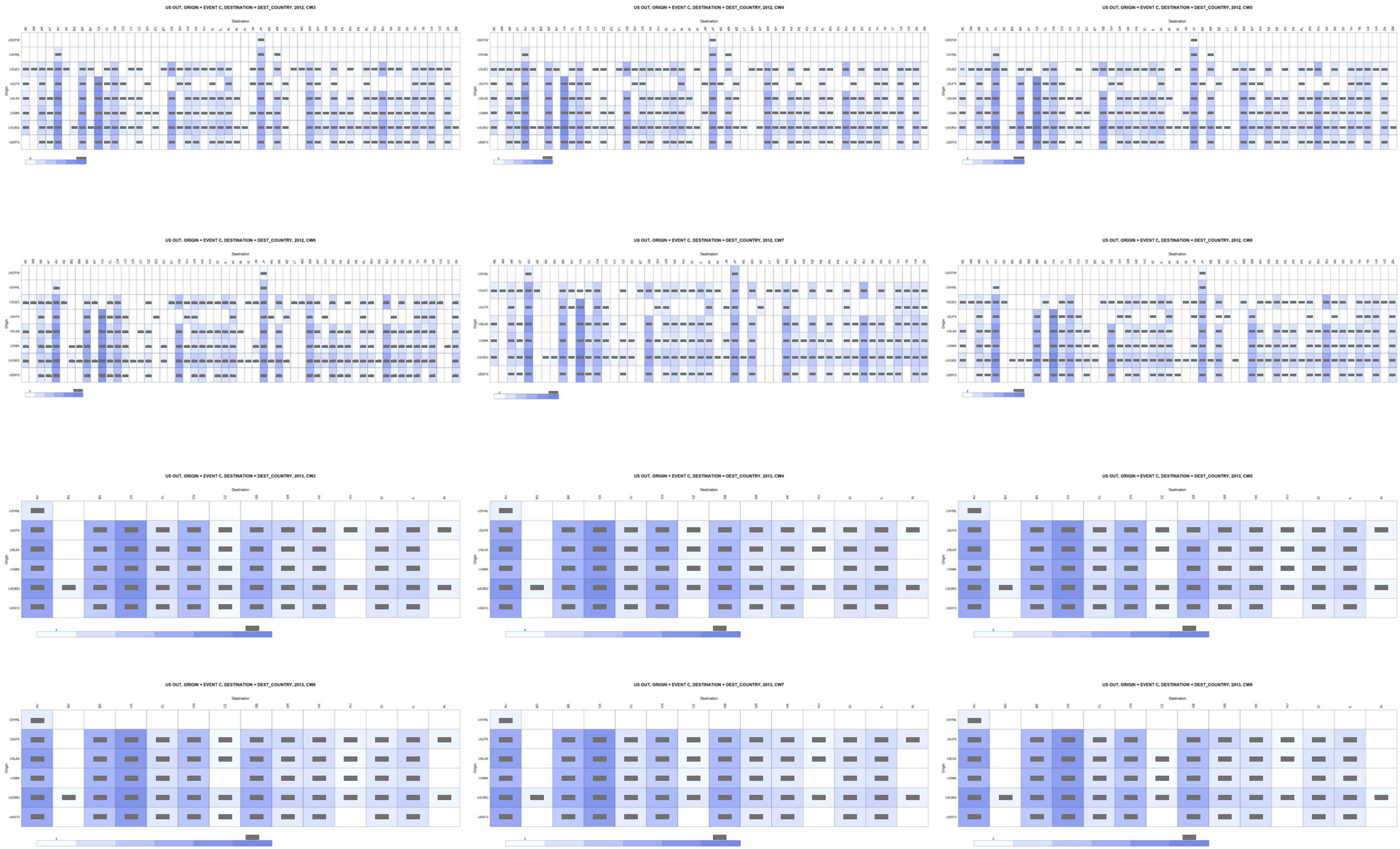


US inbound LOD 2. CW3-8 2012 (top) and 2013 (bottom)

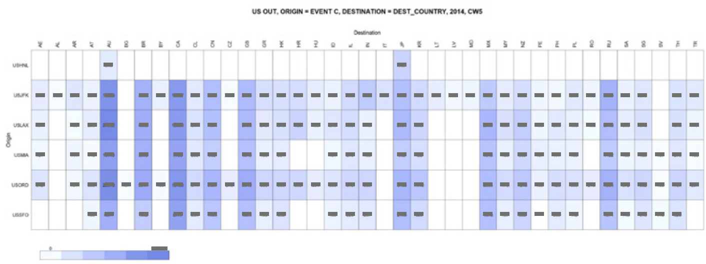


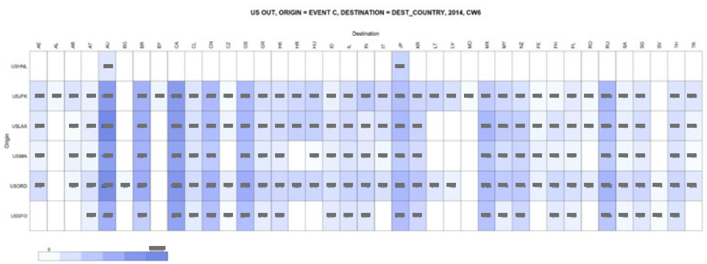
US inbound LOD 2. CW3-8 2014 (top) and 2015 (bottom)

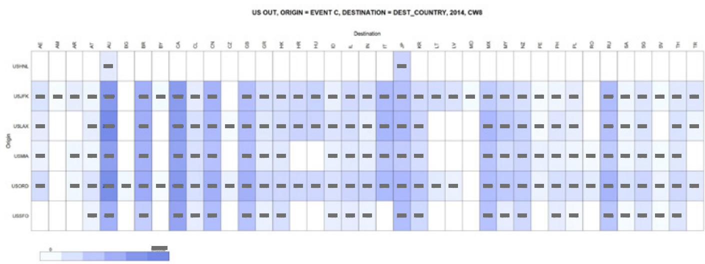
US outbound LOD 2. CW3-8 2012 (top) and 2013 (bottom)

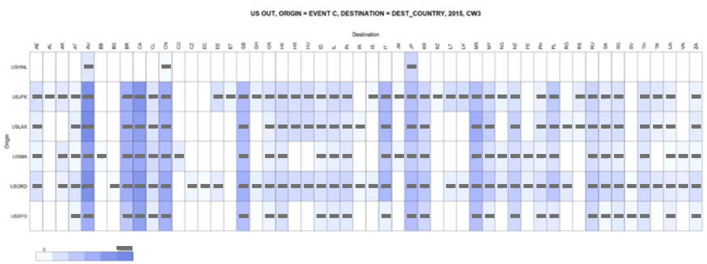


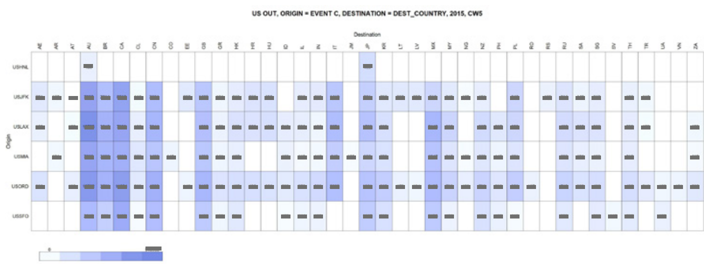
US outbound LOD 2. CW3-8 2014 (top) and 2015 (bottom)

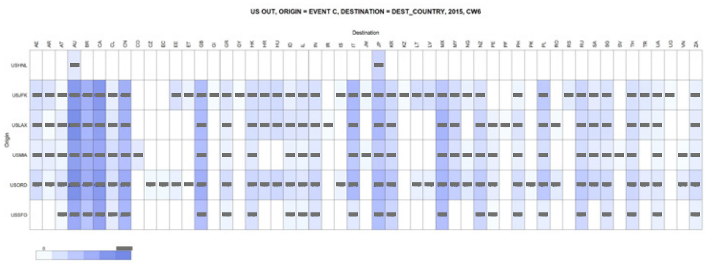


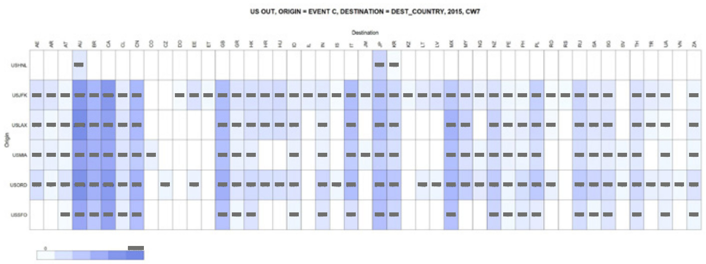


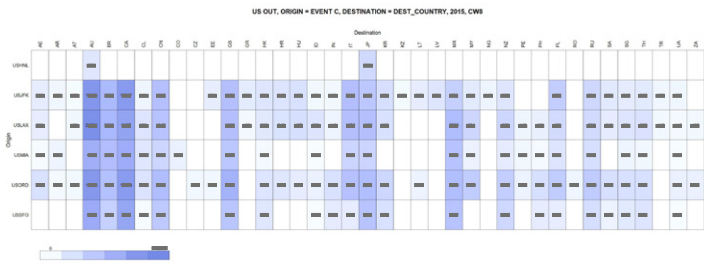




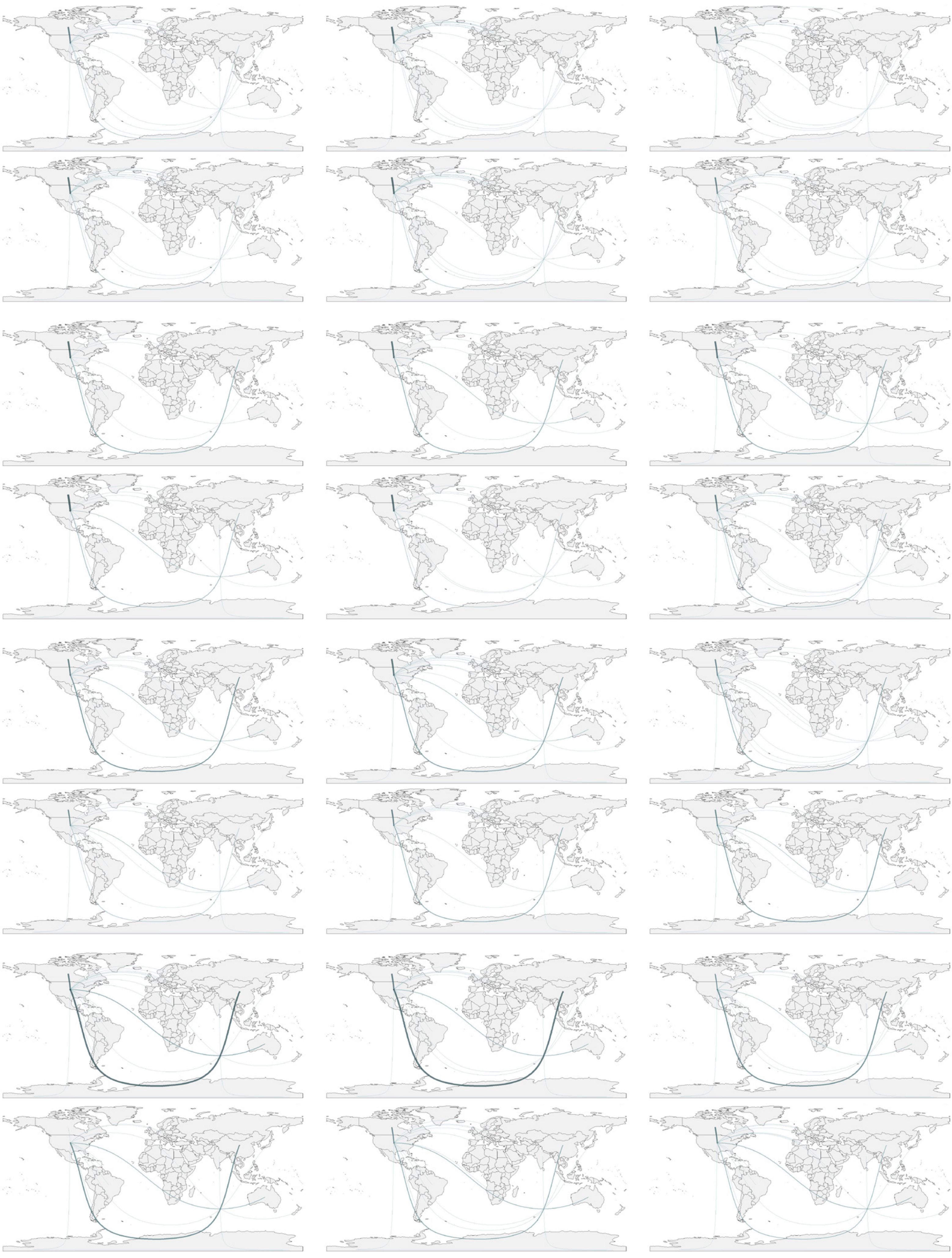




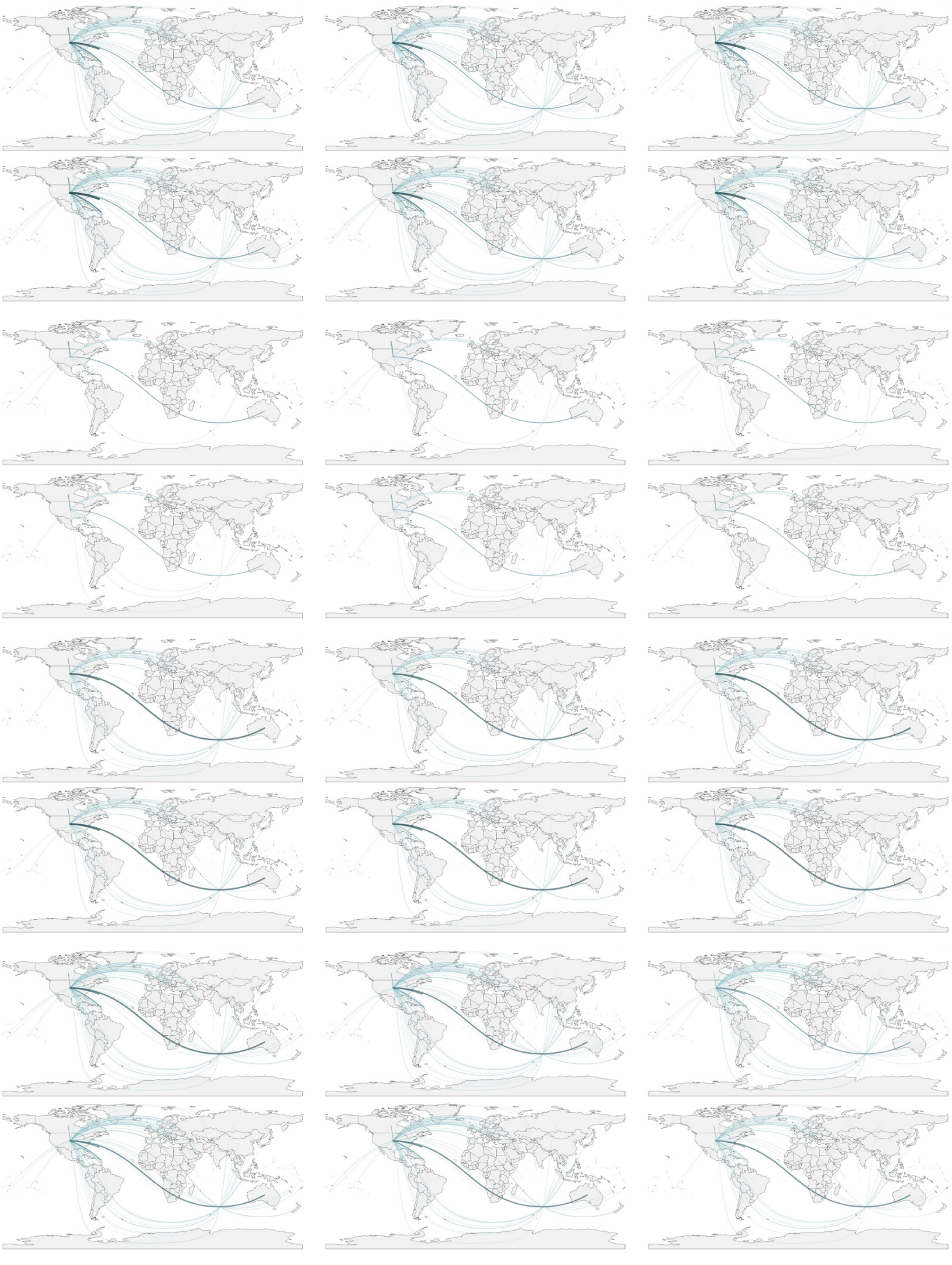




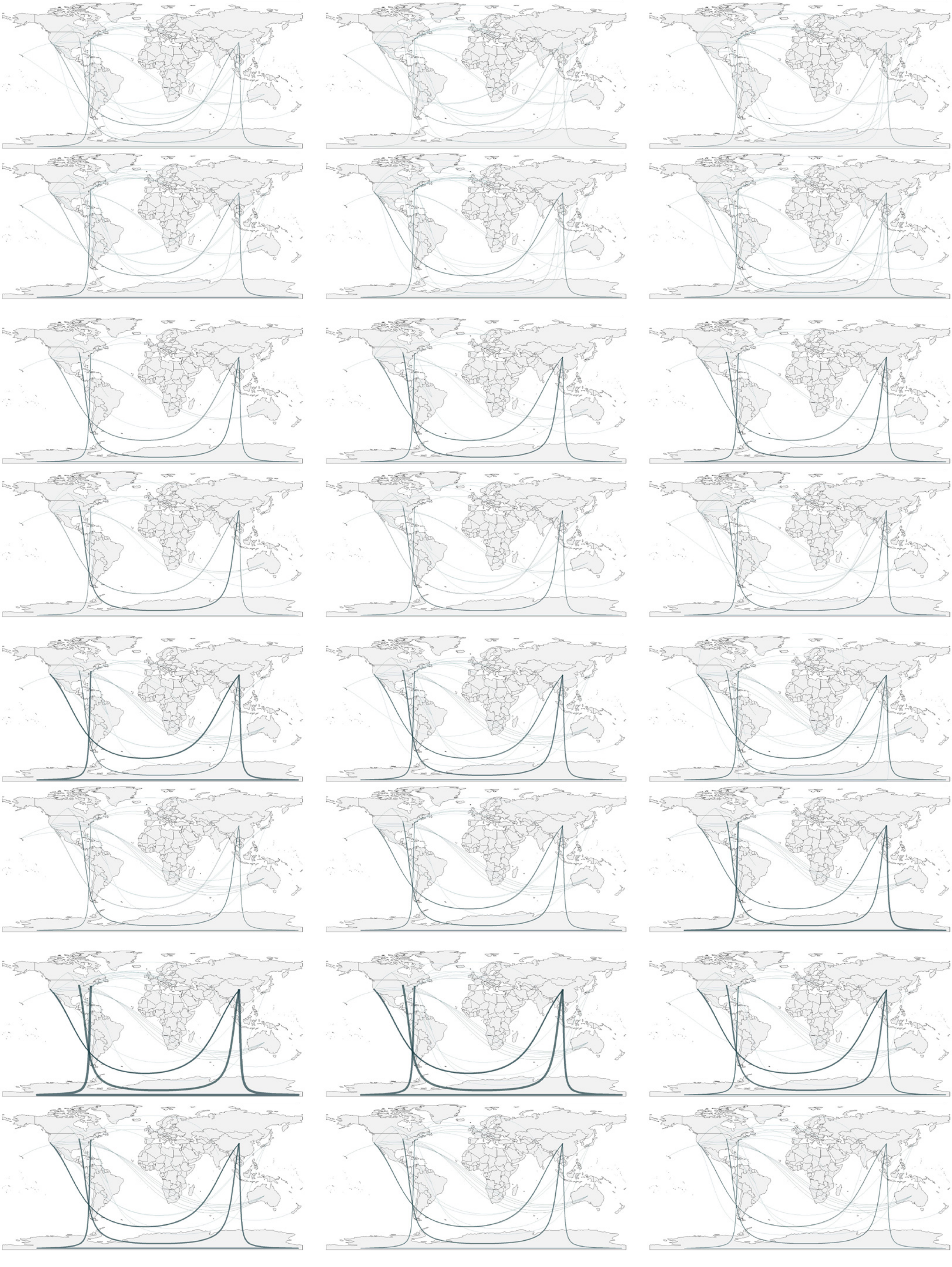
US inbound LOD 1. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)



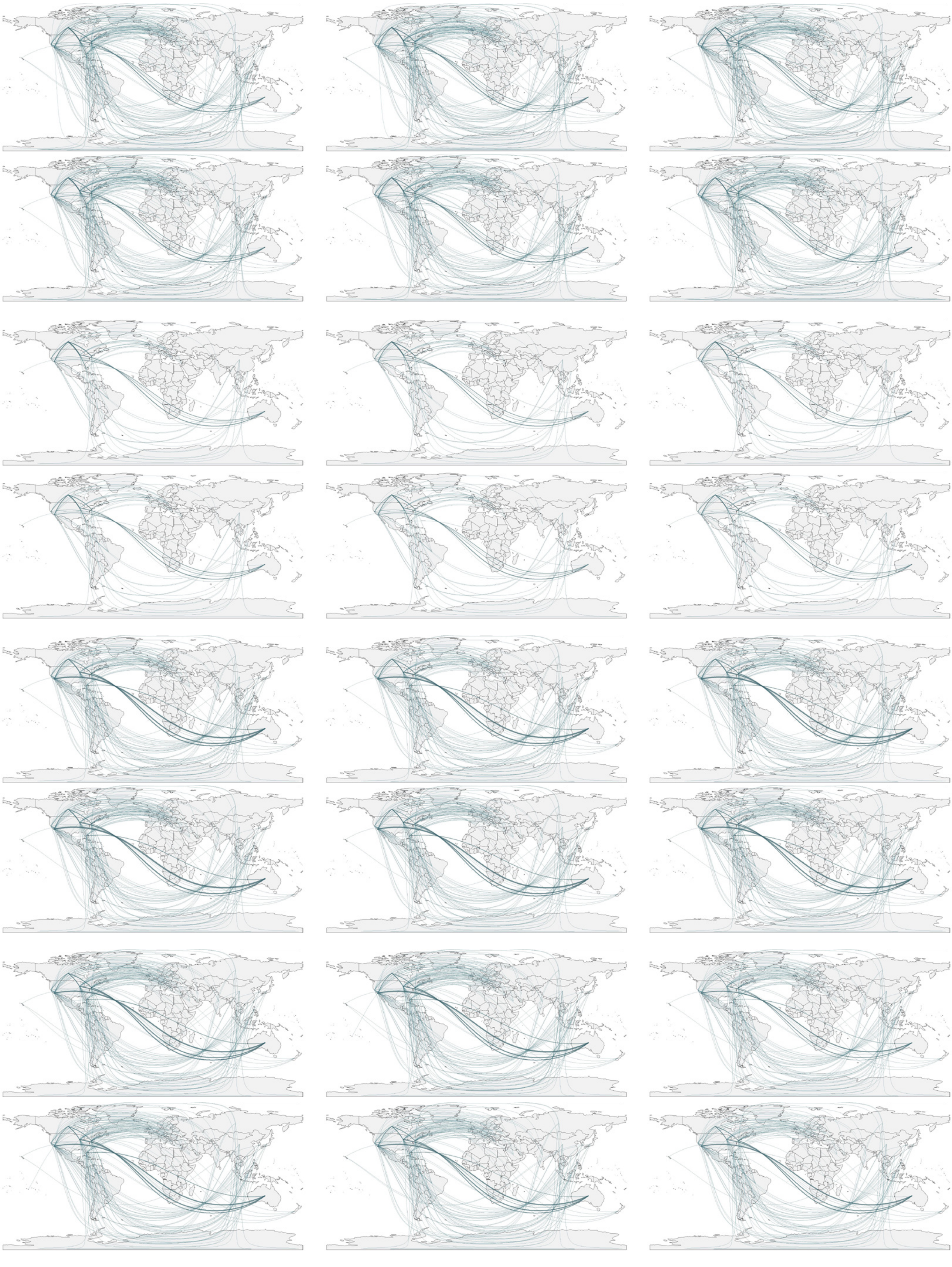
US outbound LOD 1. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)



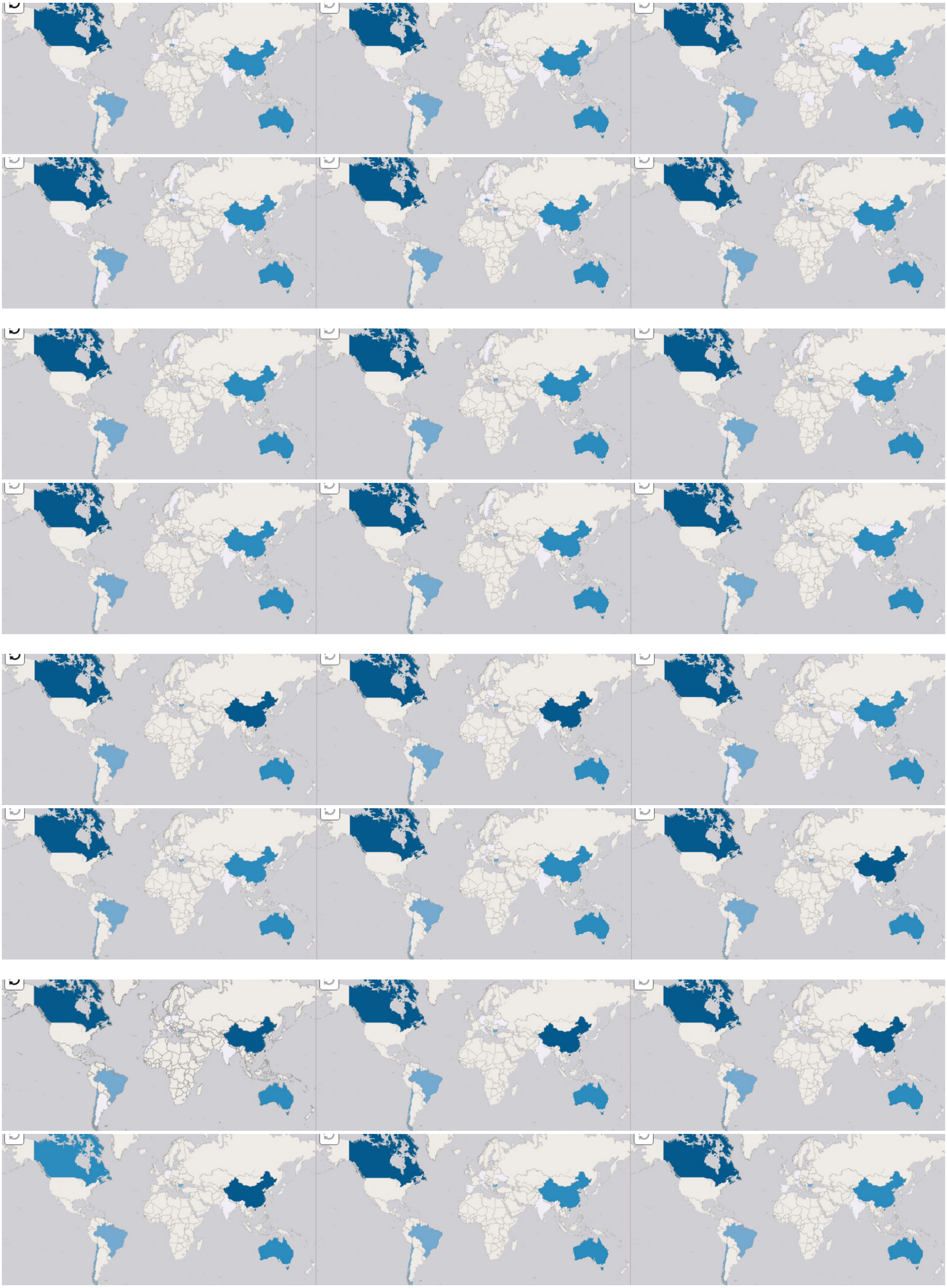
US inbound LOD 2. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)



US outbound LOD 2. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)

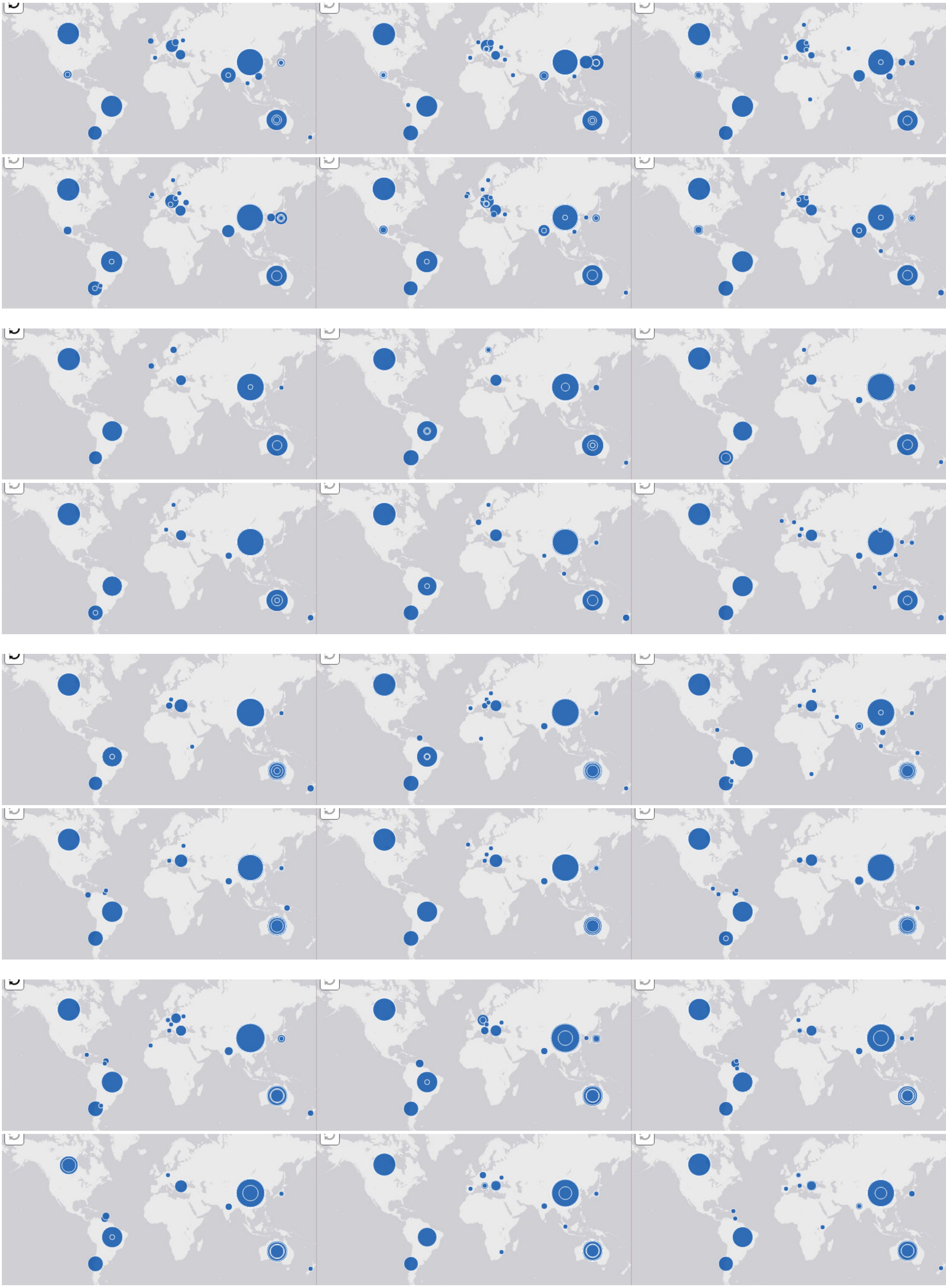


US inbound LOD 1. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)

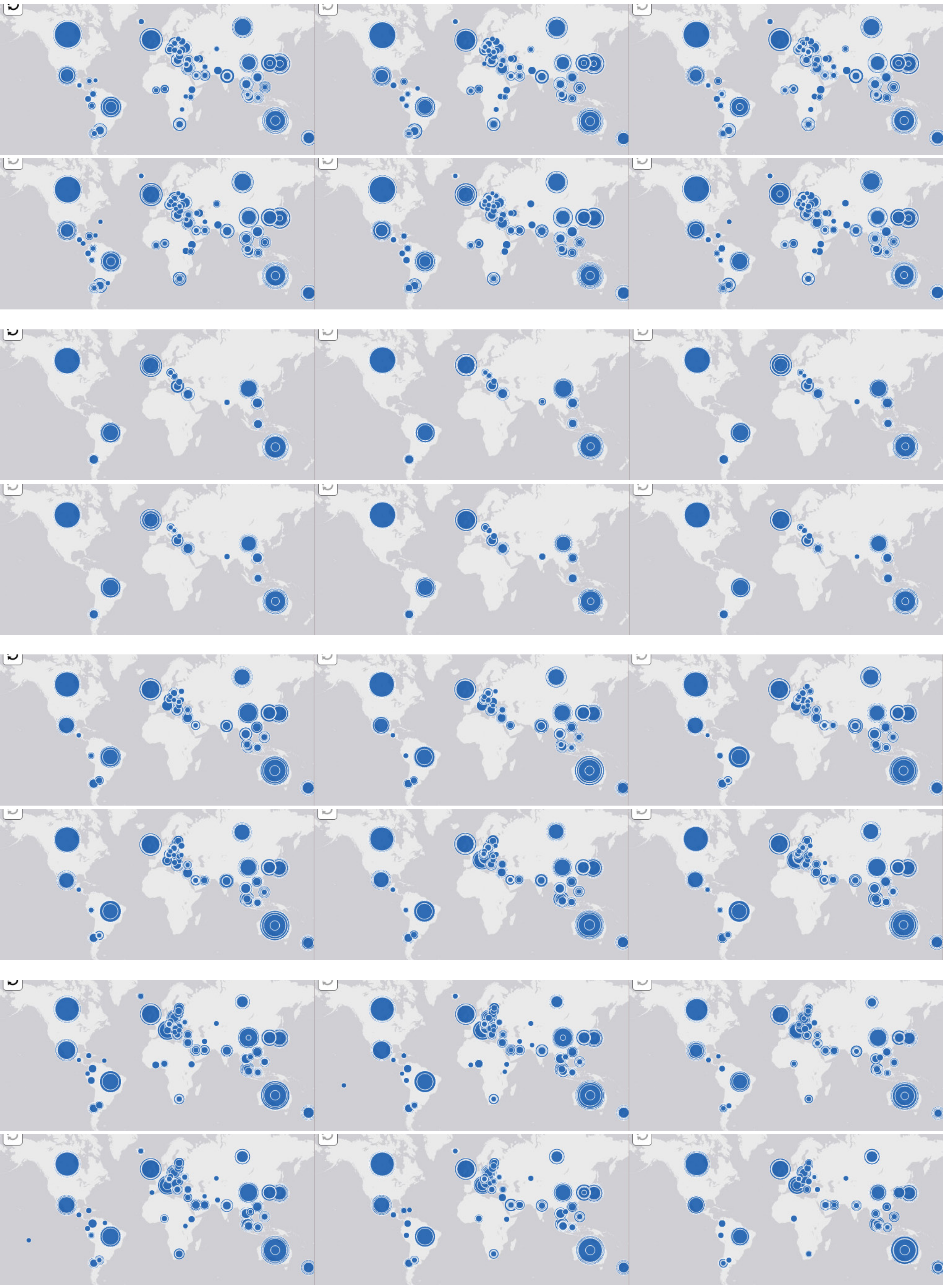


US outbound LOD 1. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)

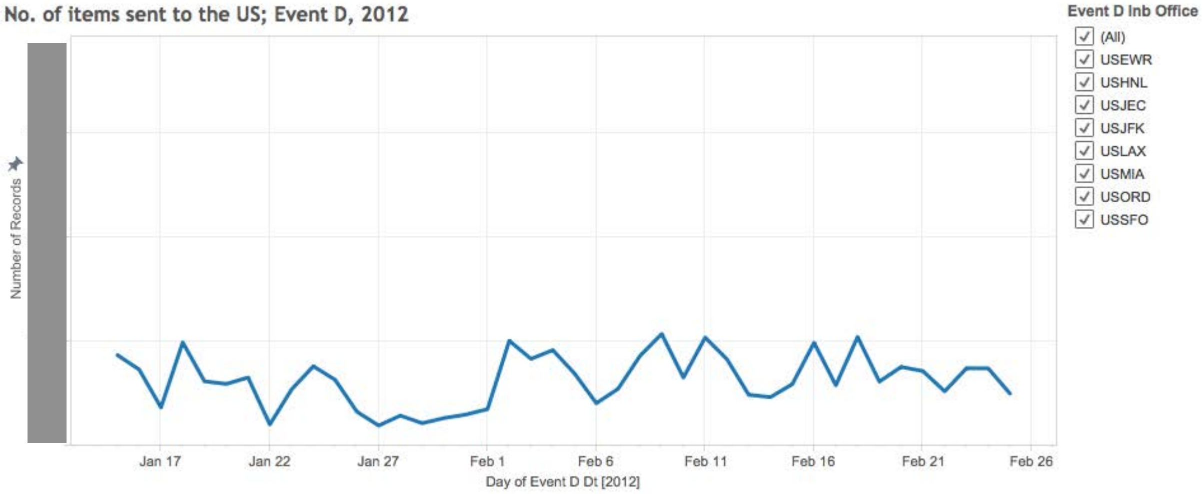
US inbound LOD 2. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)

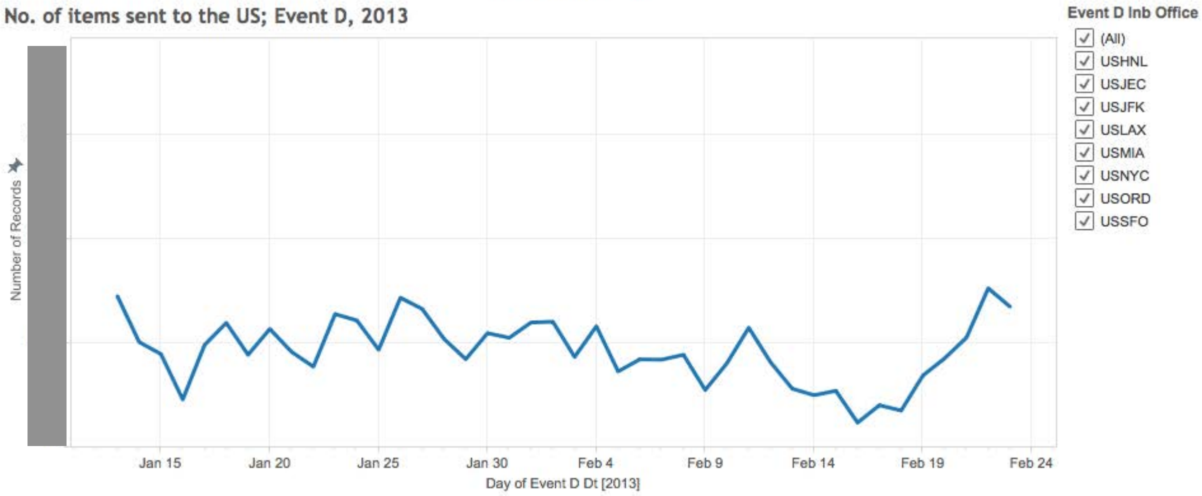


US outbound LOD 2. CW3-8 2012 (top), 2013 (second), 2014 (third), 2015 (bottom)

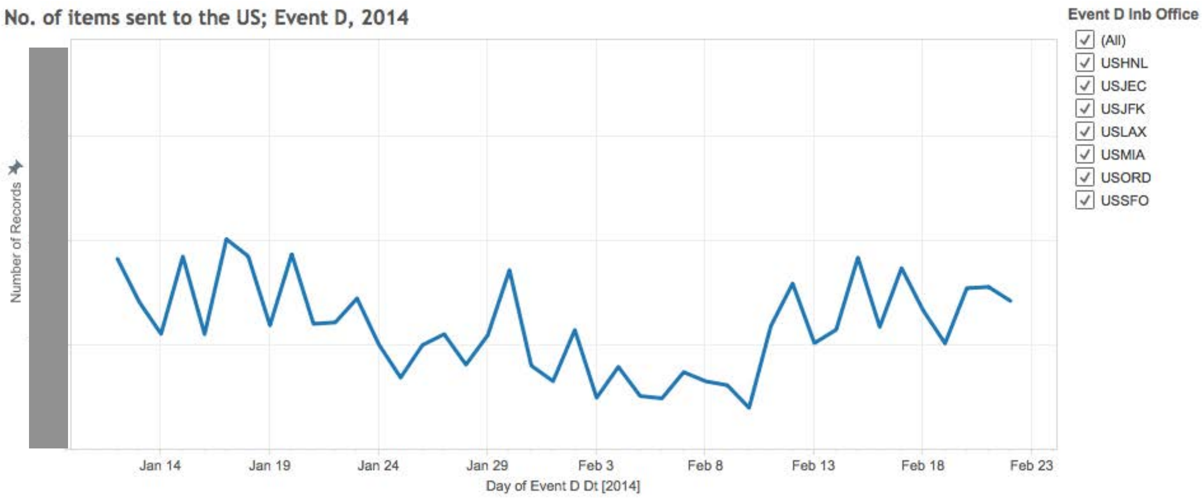


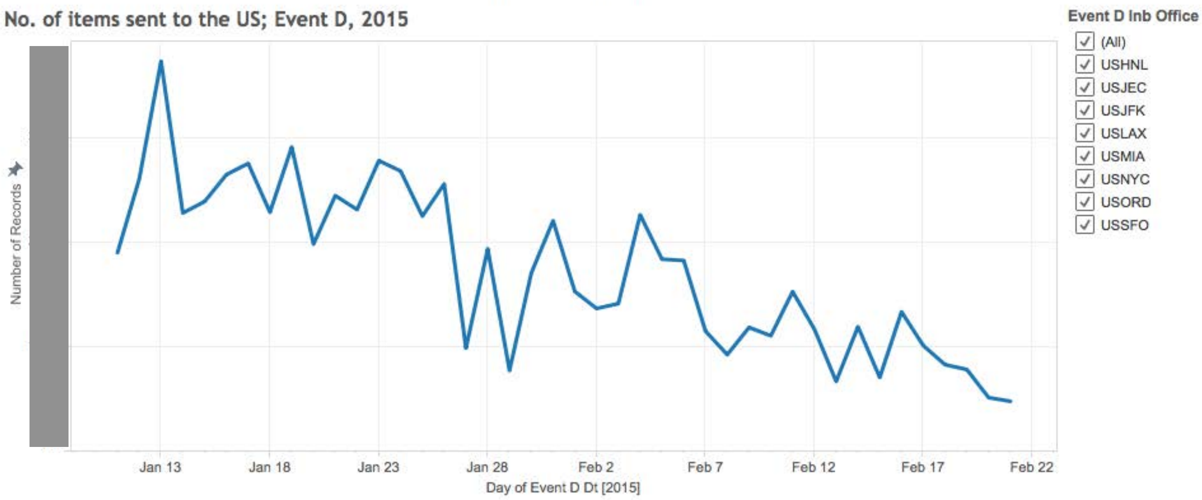
US inbound LOD 1. CW3-8 2012 (top), 2013 (botttom)



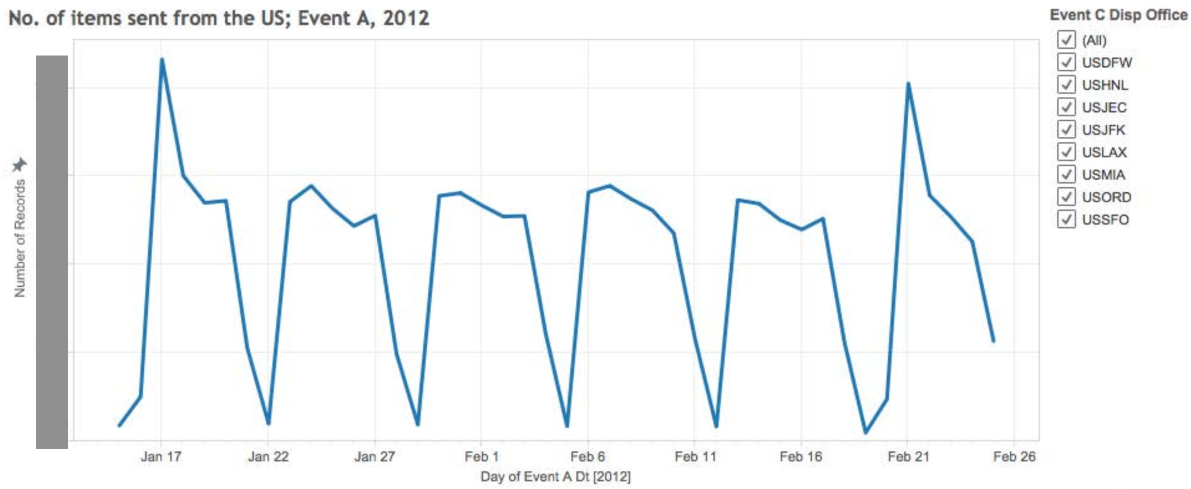


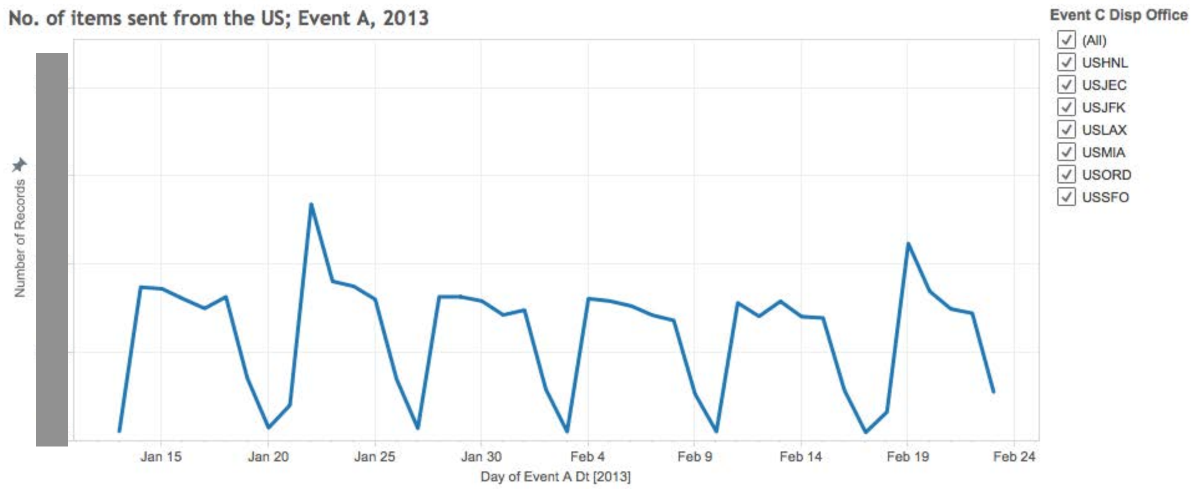
US inbound LOD 1. CW3-8 2014 (top), 2015 (botttom)



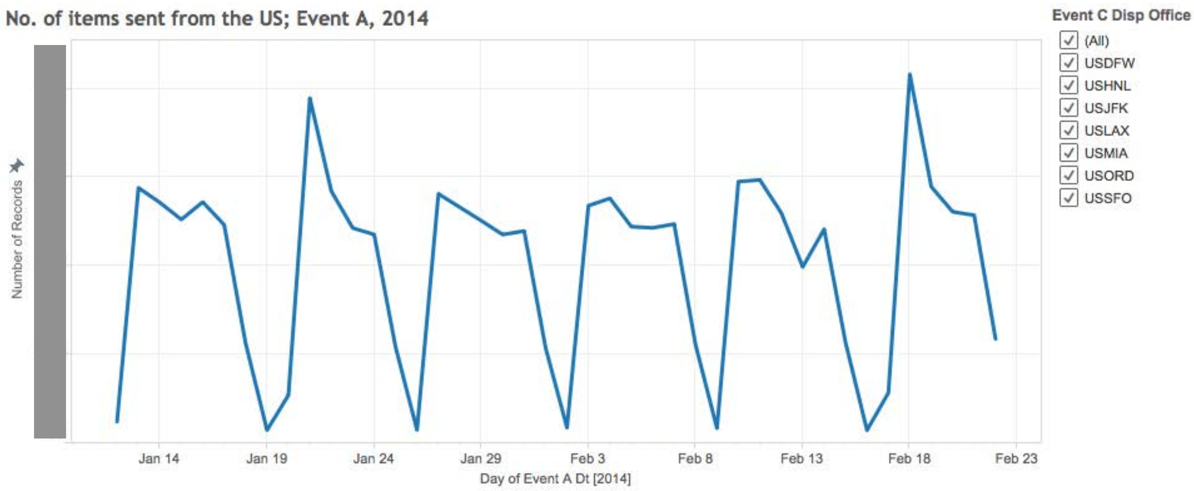


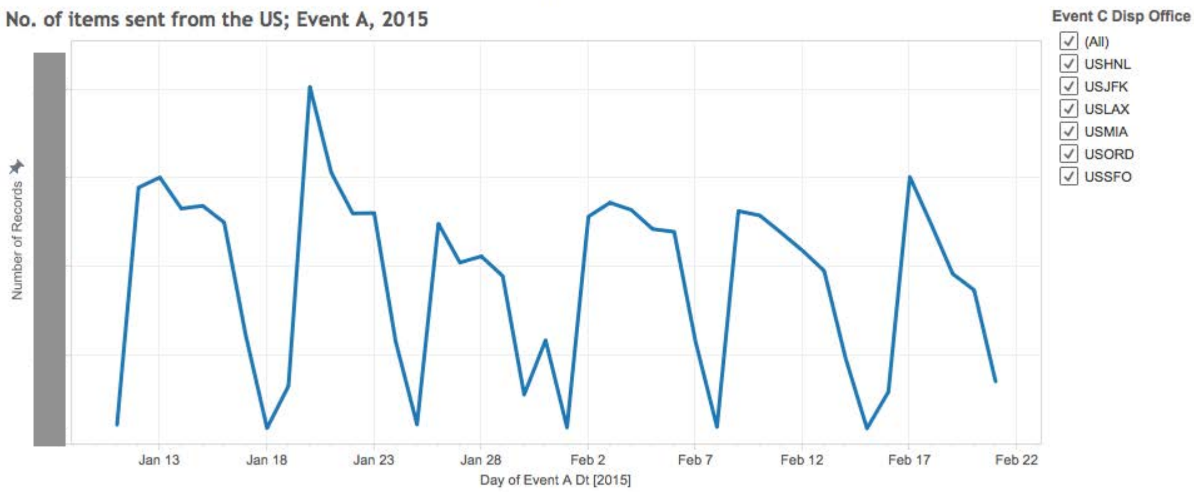
US outbound LOD 1. CW3-8 2012 (top), 2013 (botttom)



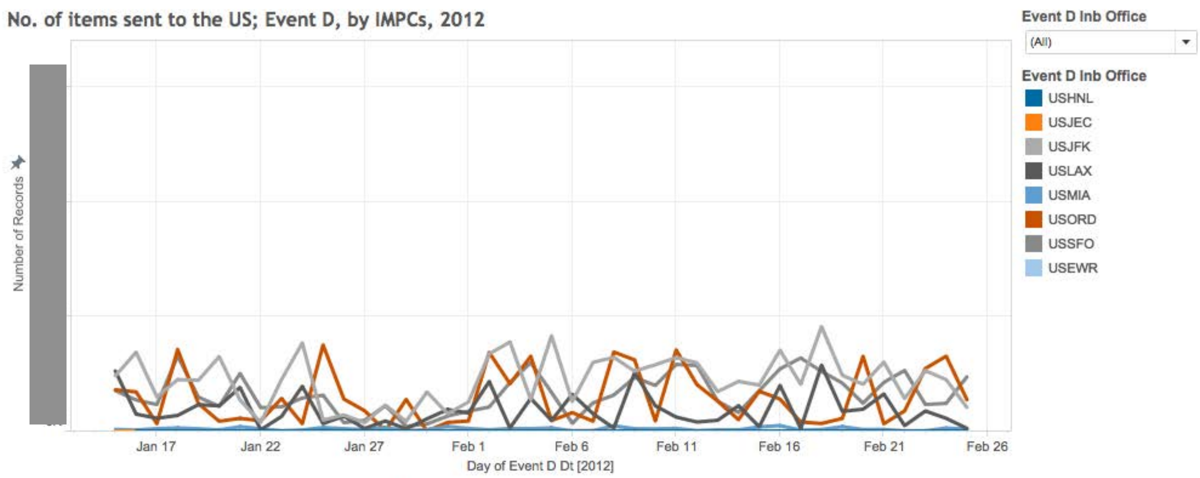


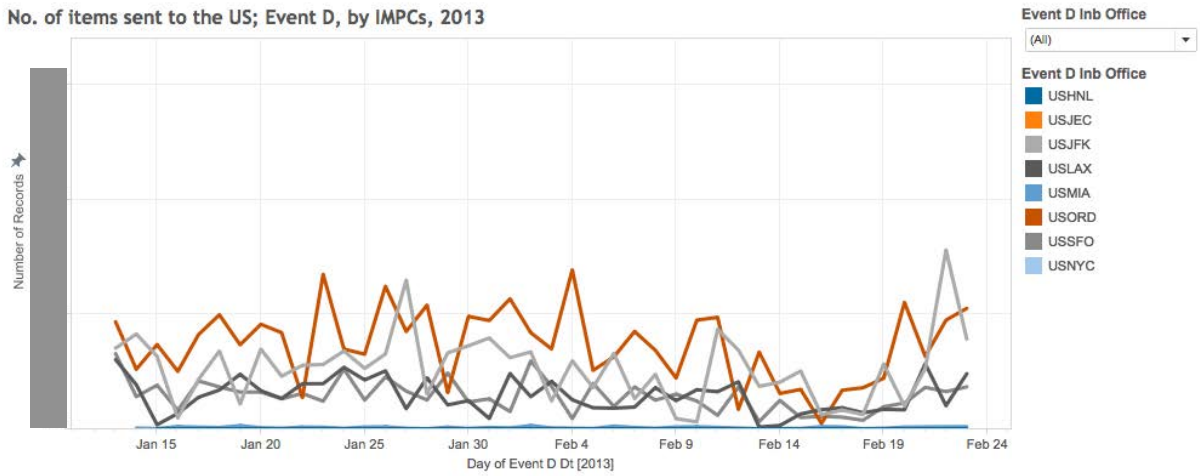
US outbound LOD 1. CW3-8 2014 (top), 2015 (botttom)



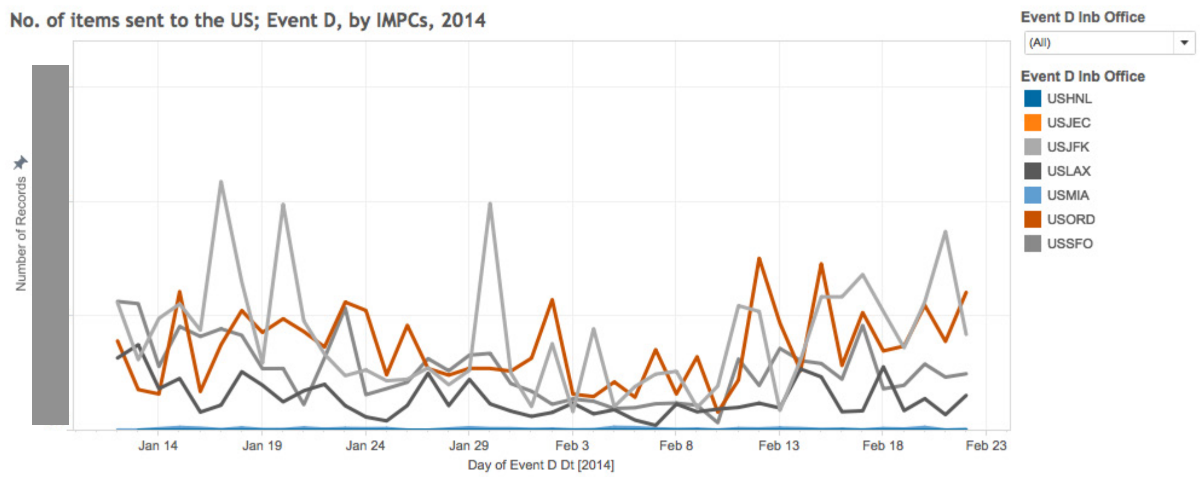


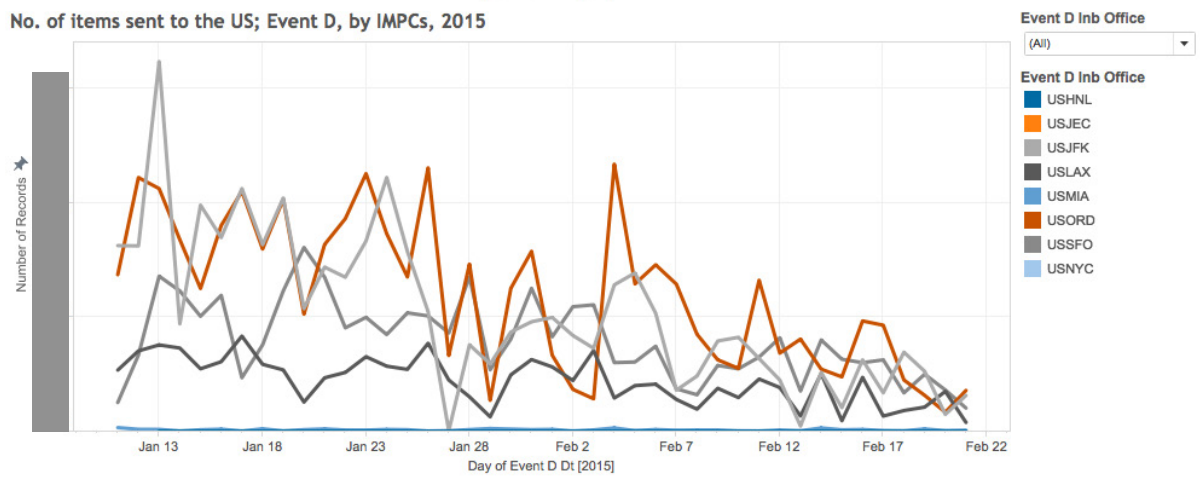
US inbound LOD 2. CW3-8 2012 (top), 2013 (botttom)



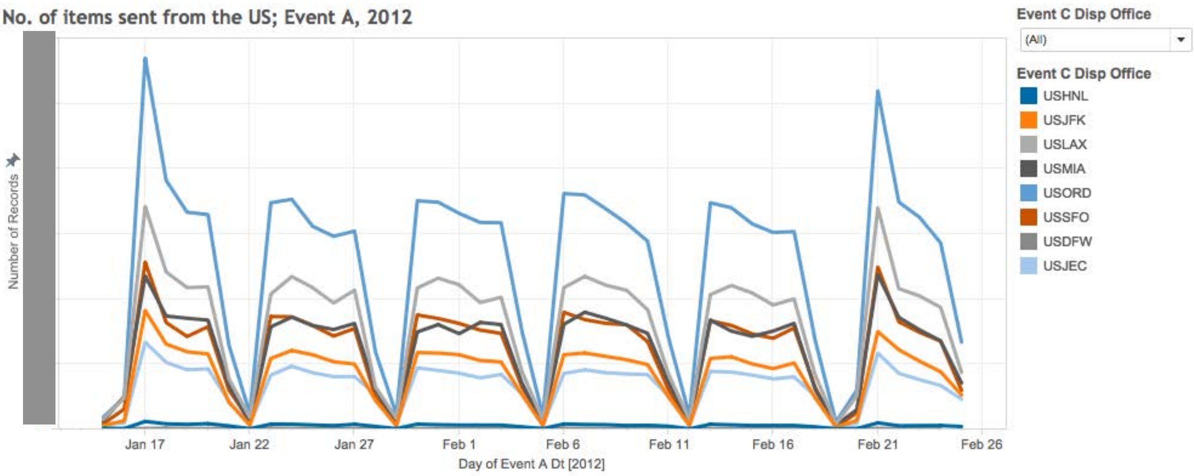


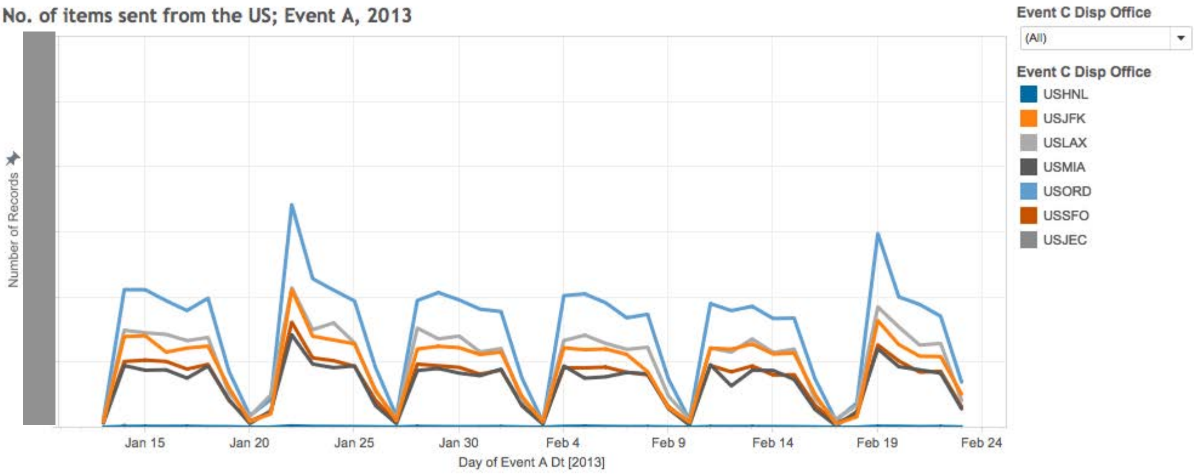
US inbound LOD 2. CW3-8 2014 (top), 2015 (botttom)



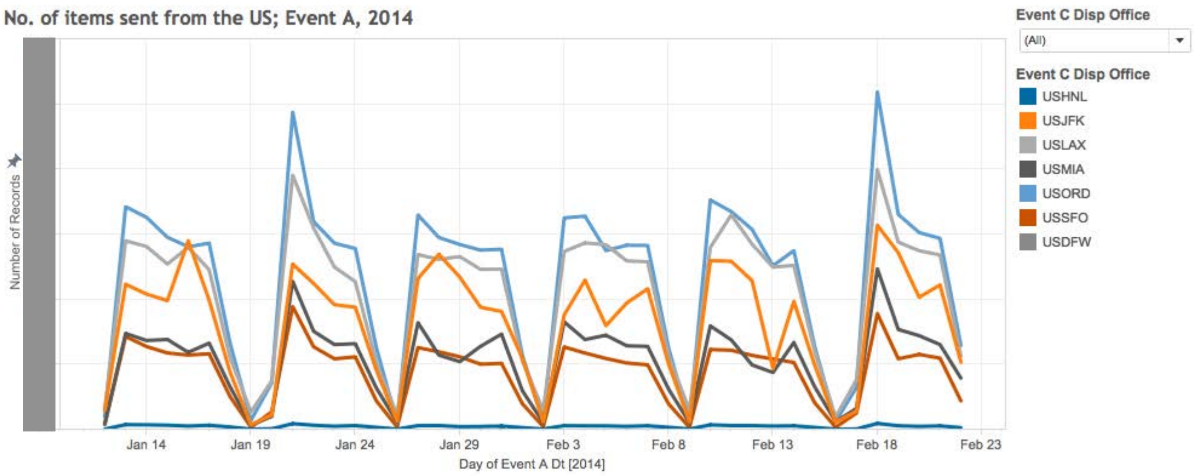


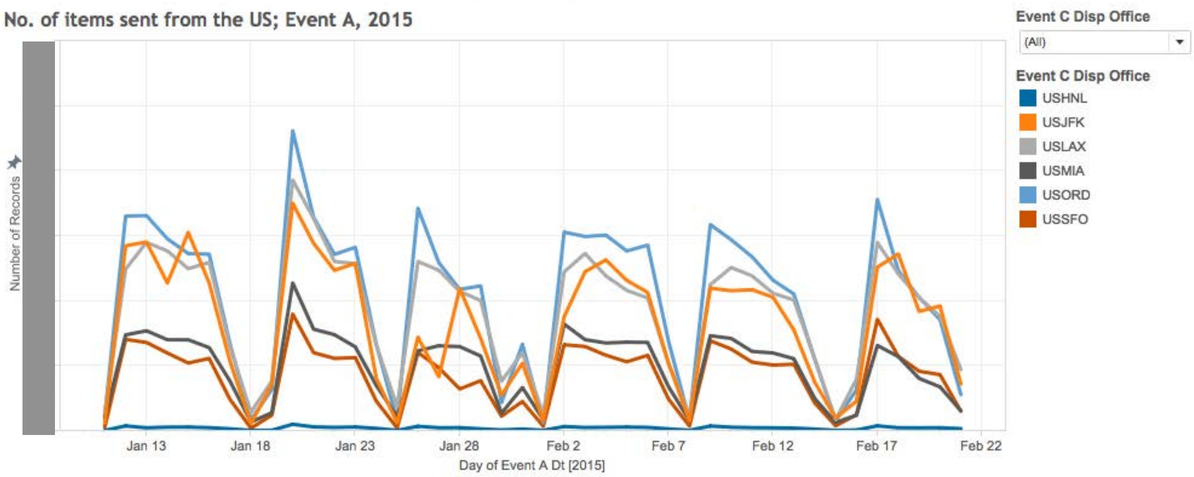
US outbound LOD 2. CW3-8 2012 (top), 2013 (botttom)





US outbound LOD 2. CW3-8 2014 (top), 2015 (botttom)





Appendix G

R Scripts

This Appendix provides a brief overview of the R scripts used for creating the visualisations in this thesis.

readAndDropCols.R
Reads in the .csv files, drops columns that are not needed, cleans the data of erroneous entries, makes subsets (years and CWs), saves clean files and writes .csv files for ArcGIS and Tableau.

population.R
Prepares data frames with population numbers of the needed years from population data obtained from the World Bank.

lookupWorldmap.R
Makes a look up to check for missing coordinates in the postal data for enabling plotting flow maps.

subsetsUSA.R
Creates clean subsets of US data.

CLNormalizePop.R
Merges CL data set with population data to normalise the number of items to population of the respective countries for LOD 1.

USNormalizePop.R
Merges US data set with population data to normalise the number of items to population of the respective countries for LOD 1.

ChileFlowmap.R
Used for creating flow maps for Chile (inbound and outbound, all CWs, LOD 1 and 2). Calls function from functions.R.

USFlowmap.R
Used for creating flow maps for the US (inbound and outbound, all CWs, LOD 1 and 2). Calls function from functions.R.

USODM.R
Used for creating ODMs for the US (inbound and outbound, all CWs, LOD 2). Calls function from functions.R.

difftime.Tran1.EVTD.R
Used for computing the time difference between the planned and actual time of arrival (or departure) of items. Calls function from functions.R.

timeSeries.R
Creates time series plots. Used packages: dygraphs, TSA.

chordDiagram.R
Creates chord diagram plots. Used packages: circlize, migest, dplyr, tidyr.

writeGexf.R
Writes gexf files that can be used in Gephi to create graphs. Used package: rgexf.

UtCTimeConversion.R
Converts local time to UTC time. Used packages: lubridate.

functions.R
Contains all functions used for creating the visualisations used in this thesis.

- UTC.conversion
- f.ts.counts (creates time series visualisation)
- f.ts.trend (creates time series visualisation)
- f.graph.flows (creates static graph visualisation)
- f.graph.flows2 (creates static graph visualisation)
- f.odm.vis.logcol.OUT (creates ODM for US outbound data)
- f.odm.vis.logcol (creates ODM for US inbound data)
- map_world_CL.IN (draws world map, calls function add_lines_CL.IN; used packages: maps, geosphere, scales)
- add_lines_CL.IN (adds lines to the plot, calls function clean.Inter_CL.IN)
- clean.Inter_CL.IN (checkDateLine_CL.IN)
- checkDateLine_CL.IN (checks date line for plotting all lines within the map)
- f.timediff.TRAN1.EVTD (time difference between planned and actual arrival)
- f.timediff.TRAN1.EVTC (time difference between planned and actual dept.)
- f.timediff.AC (time difference between two *Events*)