Data Visualization
InfoVis Project

Group 12

18 December 2018

Name	Student Number
Nele Albers	4853261
Colm Seale	4772474
Mihai Bogdan Voicescu	4951433

1 Introduction

The discussions about the recently adopted UN migration pact, the European migrant crisis, and the lack of skilled labor in countries such as Germany are just some examples that highlight the importance of better understanding international migration in order to able to design effective solutions for all involved countries. The inherently complex nature of the corresponding data of migration flow, and of factors that may be correlated with migration flow, thereby means that a visualization that aids the user in the exploratory analysis of the data is necessary. Consequently, this paper delineates a visualization application designed for people wishing to take part in improving some of the aforementioned problems by discovering geographical and temporal patterns and correlations.

Following the four levels of analysis developed by T. Munzner [1], the visualization design will be justified by first describing the domain situation in Section 2.1, mapping the domain-specific questions to abstract visualization tasks in Section 2.2, outlining the visual encoding and interaction idioms chosen in Section 2.3, and lastly describing the algorithm employed in Section 2.4. Afterwards, we will evaluate our visualization by examining the way in which it helps answering the initially formulated domain-specific questions in Section 3, before summarizing our approach and providing further perspectives in Section 4.

2 Visualization Design

2.1 Domain Situation

In this section, we will outline the target users of our visualization as well as their data, domain of interest, and questions [1].

The target users of our visualization are people interested in exploring international and national patterns in migration flow as well as correlation with factors such as a country's GDP or homicide rate. Such people may be politicians eager to design new migration policies or to evaluate existing ones, journalists wanting to educate the public about migration flow in order to counter existing bias and misinformation, and aid organizations aiming at supporting countries with particularly high levels of migration.

The foundation of our information visualization is the DEMIG C2C data collected by the University of Oxford [2]. It contains bilateral migration flow data for 34 countries and from up to 236 countries for the years 1930 to 2011. More specifically, it provides information with respect to netflow, inflow, and outflow with a gender and coverage breakdown for most of these countries.

This data are of interest to anybody who would like to put the current migration situation in one country into a geographical as well as historical context. With respect to the geographical context, the first important aspect is the ability to get an impression of the total flows for all countries within a certain time range, which allows the evaluation of one country's flow relative to the ones of other countries. For

example, it is relevant to see whether countries that are spatially close to each other have similar levels of migration, and to discern where regions with comparatively high and low migration flows are located. Moreover, one might want to directly look for countries with similar migration flows independently of where they are situated geographically.

Regarding the historical context, an analyst is interested in the change in the migration flow of a specific country over time. This makes it possible to identify periods of low and high migration as well as overall trends. One possible application of this is the evaluation of the effectiveness of policies aimed at altering specific types of flow, such as a law targeted towards increasing the inflow of skilled workers for a specific country. Whenever a noteworthy year or time range has been found, the analyst wants to be able to focus on this period of time in order to get a more detailed impression. Thereby, it is crucial to compare national observations to the global context in order to identify whether a peak in one country's flow is accompanied by high flow levels in neighboring or otherwise related countries, for instance. This would enable conclusions as to whether the observation for the given country is likely to be explained by a certain global trend or rather a national one. Lastly, there are two important breakdowns of the total flows that make it possible to discern the change in flow with respect to certain groups of people. The first one is, similarly to the geographical context, a breakdown of the netflow, outflow, and inflow into partial flows such as the outflow of citizens. Yet, it is also useful to visualize the sources and destinations of the inflow and outflow, respectively, from a certain country, because it allows the analysis of the correlation between events such as a famine or war in one country and the outflow to and inflow from that nation.

However, in order get a better understanding of the forces behind migration and to possibly design solutions to problems related to migration, it is not sufficient to simply look at the flow numbers and their geographical and temporal change. Instead, it is crucial to find correlations between migration flow and other factors related to areas such as economic growth, health, and education. Hence, we combined our initial migration flow dataset with a time series containing indicators drawn from the world development indicators for 217 economies for the years 1960 to 2017 [3]. Based on the resulting dataset, an analyst is interested in discovering correlations between a country's migration flow and values for a specific factor and to compare correlations across different factors. For example, a country's GDP may be positively correlated to its inflow, while there might be no correlation between the country's inflow and homicide rate. Moreover, outlying data for specific years provide a starting point for investigating what kind of events might overrule such general correlations. While an analysis of correlations does not make it possible to pinpoint causes of migration, since an increasing GDP may also be the result and not the cause of increasing inflow, for example, it does pose a valuable foundation based on which more detailed hypotheses can be formulated.

With respect to both employed datasets, it is important to point out their limitations. First of all, data for certain year and country combinations are missing for both the factors and the migration flow. We considered several options of meaningfully filling in missing data by different forms of interpolation, but ultimately decided against it. The reason is that it might cause an analyst to find trends and formulate hypotheses that are in fact not sufficiently supported by the actual data, as some data gaps are rather large. Secondly, for some countries, there are no migration flow data available at all. This is the case because, in contrast to migration stocks data, there are no globally collected data. Hence, annual and comparable migration flow data are mostly limited to countries in the OECD [4]. Thirdly, countries at different times employ varying ways of defining and measuring outflow, inflow, and criteria such as citizenship. Yet, despite these limitations, we think that our combined data can be of great use to anybody wishing to take part in the current debates about migration, as long as he or she keeps the restrictions in mind.

2.2 Task and Data Abstraction

The next level in the framework for visualization design described by T. Munzner in [1] is the task and data abstraction, which is a necessary step because questions from dissimilar domain situations can map to the same abstract visualization tasks. Thus, we will begin with the data abstraction and will, following the approach outlined in [5], conclude with the task abstraction.

2.2.1 Data Abstraction

Our main dataset, the one containing the migration flow data, can be characterized as a multidimensional table. The reason is that each individual migration flow value can be uniquely identified by the combination of a year, a high-level flow type such as inflow, a low-level flow type such as male or female,

a reference country, and a destination or source country in the case of outflow and inflow. Therefore, as time is one of the keys of the table, our dataset is a time-series dataset. However, because countries can be identified by their location on a world map, our dataset also has a spatial component. With respect to the attribute type contained in this multidimensional table, we are dealing with ordered and, more specifically, quantitative data. Thereby, this data can be seen as sequential when looking at outflow and inflow separately. Yet, when studying them together, it becomes obvious that the attribute has an inherently diverging character with outflow and inflow constituting two sequences pointing in opposite directions from the point of zero flow. Lastly, our dataset's availability can be characterized as static, because all data are available at once.

The additional dataset we introduced in order to allow for the analysis of correlations between single factors and migration flow is also a multidimensional table. This time, a cell is uniquely identified by the year, country, and factor name. Similarly to the migration flow dataset, this one, too, is a time-series with a spatial component. Yet, this dataset does not have a single diverging quantitative attribute, but several sequential quantitative ones. Finally, this dataset is also static.

2.2.2 Task Abstraction

In order to map the detailed questions described in Section 2.1 to abstract ones, we will follow the approach developed in [5] by considering the five dimensions of the design space, which are the goal, means, characteristics, target, and cardinality of the task.

Goal. The first aspect is the goal of the task. In our case, this is mainly an exploratory analysis. Starting from the flows for all countries, the analyst is interested in finding global patterns such as the spatial distribution of regions with high or low flow values as well as temporal trends and correlations for specific countries. Based on this initial exploration, the user can then develop hypotheses regarding questions such as the effectiveness of policies or the relationship between a country's GDP and its inflow.

Means. In order to reach the goals of the tasks, reorganization, relation, and navigation are employed. With respect to the former, data can be filtered based on the time point and country as well as the migration flow value itself. This is achieved by selecting a single country of interest in order to study its migration flow and correlations with certain factors over time, and by choosing a time range for which to depict all countries' flow values and the selected country's correlations. Additionally, a range of flow values can be chosen in order to discover countries with flows that have specific levels. Regarding relation, a country's flow value for a certain time point can always be evaluated against the backdrop of the country's flow over time and of the flows of other countries for the chosen time range. Lastly, navigation allows the analyst to zoom in on specific time ranges and geographical regions in order to get a more detailed view.

Characteristics. In this step, we describe what characteristics of the data the tasks are supposed to reveal. At a low level, simple data values, in our case the migration flow of a certain country for a specific time point, should be visualized. At a higher level, trends over time and geographical trends should be depicted. Thereby, outlying data should also be easy to spot. Moreover, the user should be aided in finding correlations between factors and migration flow, as well as in discerning outliers and clusters from such correlations.

Target. The fourth step of the task abstraction is concerned with the part of the data the tasks are carried out on. Thereby, data objects are linked with respect to both their attributes and each other. The first is achieved by linking data objects over time when showing the temporal change in a country's migration flow and flow composition, and the correlation between migration flow and other factors. Yet, data objects are also linked to geographical areas, because each flow data point is associated with a country, which in turn can be seen as an area on a map. The resulting spatial relations allow finding geographical trends with respect to migration flow. Lastly, data objects are linked with each other, too, because a user may be interested in finding similar migration flow values across different countries as well as in discovering correlations between migration flow and other factors such as a country's GDP.

Cardinality. Next, we analyze how many instances of the chosen target are considered by a task. As our combined dataset contains information with respect to multiple factors and different types of flows, not all information can be displayed at once. Therefore, initially one type of flow and a limited number of factors should be selected as the basis of the visualization. Yet, these selections can be altered by the user. With respect to the time dimension of our data, the task of finding temporal trends initially considers all instances. Yet, zooming in on fewer time instances up to a single one is desirable. The same is true for the task of discovering correlations. The task of analyzing geographical trends, however, is mainly based on an aggregation of the values of individual time instances to a single overall value.

Regarding the spatial dimension of our dataset, the tasks of finding temporal trends and correlations consider a single country at a time. Yet, the task of discovering geographical trends initially is based on all instances, and the user can zoom in on several or even a single country.

2.3 Visual Encoding and Interaction Idiom

The third step in the four levels of visualization design is the development of what users see and how they interact with what they see in the visualization [1]. First, we will consider the choices made for each individual view's visual encoding and interaction idiom, and afterwards we will describe how the views are linked. An overview of the components of our visualization can be found in Figure 1.

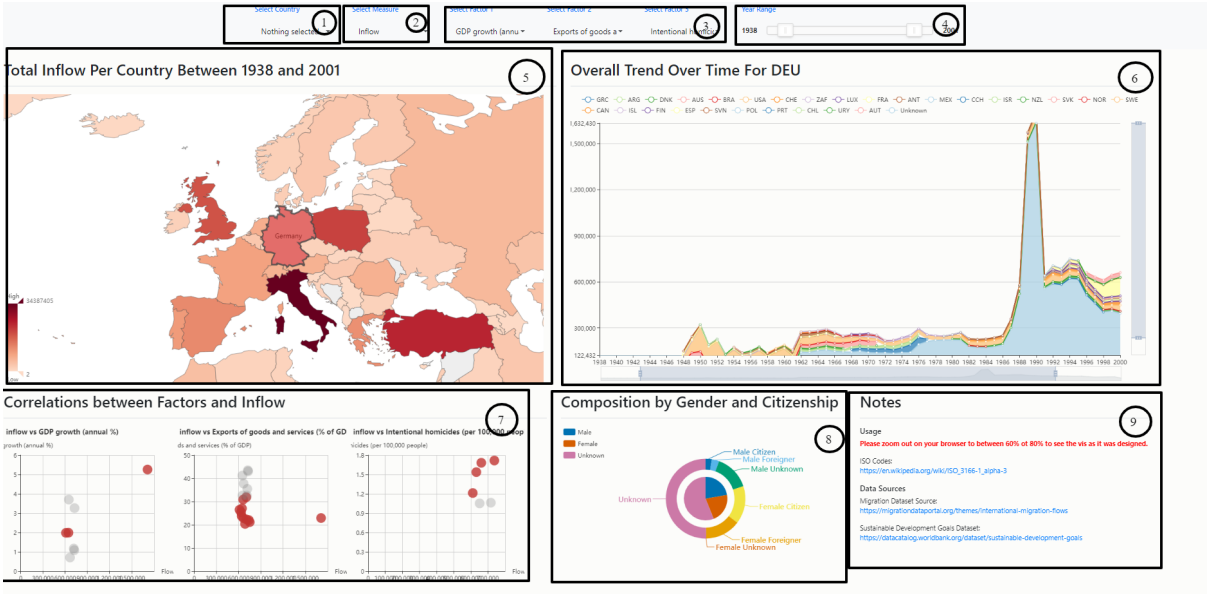

Figure 1: Overview. 1) Country selection. 2) Flow type selection. 3) Factor Selection. 4) Time range specification 5) World map 6) Stacked area chart. 7) Scatterplots. 8) Pie-in-a-doughnut chart. 9) Guide.

2.3.1 World Map

Unless the user already has a specific country in mind whose temporal trends he or she would like to explore, the entry point of our visualization is a world map, which shows the total migration flow per country for a certain time range. It initially serves as an overview and is useful for finding geographical patterns with respect to migration flow as well as to place one country's flow value into a global context. *Visual Encoding.* As this visualization component links migration data objects to their corresponding country, both attributes need to be visually encoded together. Therefore, area was chosen as a mark, because it allows the shape and position to be used as identity channels for the country. This is effective, because countries can be easily identified by the combination of their shape and their geographical position. Moreover, spatial region has the largest amount of salience of all channels for categorical data, which is further increased by additionally employing the second channel. For the diverging quantitative data of migration flow, color saturation was selected as magnitude channel. The reason is that while color saturation is not the most effective magnitude channel, it is the highest ranked one among the ones that do not interfere with the encoding chosen for the countries. Since migration flow is diverging quantitative data, we selected a diverging color scheme [6], which is shown in Table 1. Thereby, we did not actually use bins, but a continuous color scale in the visualization. Also, we made sure that the scheme was colorblind-safe in order to enable most users to distinguish the colors used. For countries with no available data for the chosen time range, we decided to use a light gray color, because it is both unremarkable and not included in the diverging color scheme. Finally, the highlighting of selected

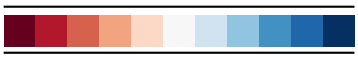

Table 1: Diverging color scheme. Red represents higher net inflow, and blue encodes outflow. When the user views inflow or outflow, the map will use that half of the scheme.

countries was chosen to be visualized by a thicker outline of the specific country's shape. While this theoretically could have been further enhanced by means of a different outline color, there is no color that can be easily distinguished from all colors selected for the diverging color scheme. Altering the color of the country's area itself is not an option, because the fill color encodes the magnitude of the migration flow. Both the highlighting and the visualization of countries with no reported data is depicted in Figure 2.

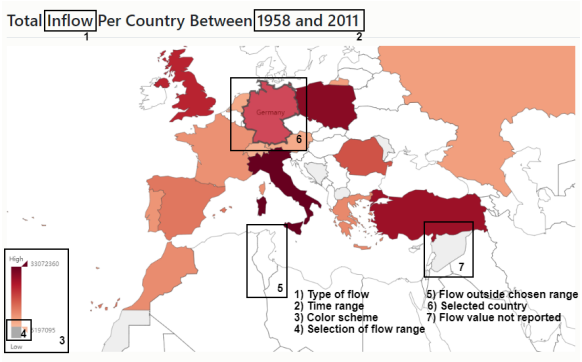

Figure 2: Visual encoding and interaction idioms chosen for the world map.

Interaction. There are several ways for the user to interact with the world map. The first one is through the navigation idiom. Specifically, users can zoom in on regions of interest in order to reduce the number of countries depicted. This also helps coping with one of the weaknesses of the visual encoding chosen, as users can only distinguish a limited number of shapes and positions at a time. The second interaction idiom is selection, whereby users can select both single countries and migration flow ranges. The former is accomplished by hovering over a country's shape, results in the country's exact migration flow value as well as its name being depicted, and serves the purpose of allowing a user to look at a single country's migration flow value. The latter is achieved by dragging the boundaries of the color scale and has two goals. The first one is that it makes it possible for users to easily find countries with similar migration flow values, and the second one is that it limits the number of different colors depicted. This is important because in non-contiguous small regions, only about twelve bins of color can be distinguished from each other by humans, which is why it is difficult to exactly compare the migration flow values of many countries at once.

2.3.2 Stacked Area Chart

The second important task our visualization is based on is the analysis of historical trends in migration flow for a specific country. Simply changing the world map over time would not be effective, because humans have limited attention and memory and would hence not be able to remember multiple different flow values for a country. Since individual filled area charts are well suited for global tasks such as the finding of temporal trends and because we would also like to visualize the change in the flow composition over time, we designed a stacked area chart. It depicts the change over time for a particular flow and its composition with respect to the flow source or destination countries.

Visual Encoding. This view links the magnitude of flow values to years and source or destination countries. The year and the flow value are encoded via the position channel, which is very effective since both attributes are ordered data. Thereby, the marks are both points and areas. Individual points indicate the flow value with respect to one country and one year, and areas are used to visualize the flow corresponding to one country over the entire time range. In order to encode the categorical attribute of country, we decided to employ the identity channel of color hue, because it is the highest ranked identity channel that does not interfere with the encoding for the year and flow value. However, this alone is not sufficient, because humans can only distinguish about twelve bins of color in discontinuous regions. Even though the regions in the stacked area chart are continuous, which allows for a slightly higher number of different colors, there are still more countries than humanly distinguishable colors. The colors we picked are portrayed in Table 2 [6]. It will be described below how we chose our interaction idiom in a way that allowed it to make up for these limitations of the visual encoding. With respect to highlighting, selected individual data points are emphasized by enlarging their area. Choosing a different outline or fill color would not have been effective, because there are already very many colors in the chart. Moreover, an encoding via motion would strongly draw the user's attention to the selected point and make it difficult to at the same time keep the temporal context in mind. Lastly, animated transitions are employed whenever a subflow is included again by the user, which makes it possible for the user to more easily notice changes.

Table 2: Categorical colorblind-safe color scheme.

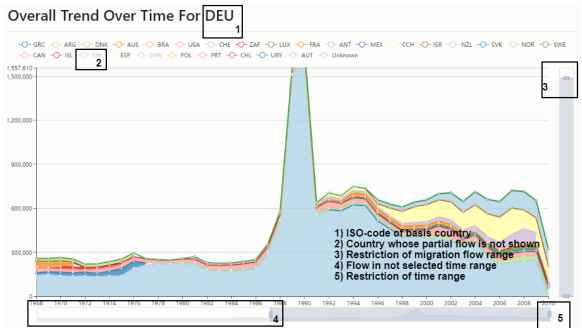

Figure 3: Visual encoding and interaction idioms chosen for the stacked area chart.

Interaction. The user can interact with the stacked area chart through different types of selection. First, both the total flow value range and the time range can be restricted. This allows the user to get a more detailed impression of the change in total flow and flow composition. With respect to the time range, it is, however, crucial to keep the entire temporal context in mind when evaluating a specific year's migration flow value. Therefore, we designed a small overview line graph of the total migration flow over the entire available time range with a shaded region indicating which time range is the basis of the stacked area chart. A second type of selection is that the subflows with respect to certain countries can be deselected. This makes it possible to look at just one subflow's change over time as well as to get a better impression of subflows whose areas are relatively small compared to others. Moreover, this selection and the resulting reduction in reference countries also lowers the number of colors shown, which enables the handling of the drawbacks of the identity channel selected for the country attribute. Another interaction idiom that makes up for the large number of colors is that hovering over a country's ISO-code in the caption highlights all corresponding individual data points. Lastly, hovering over a single data point produces a text box with the country's ISO-code, the year, and the migration flow value. Some interaction idioms are depicted in Figure 3.

2.3.3 Pie-in-a-Doughnut Chart

The stacked area chart delineated in Section 2.3.2 only depicts the change over time with respect to one type of flow composition. Regarding the other two flow compositions, gender and citizenship, we had to evaluate whether it would be more useful to the user to see their individual temporal change, or to analyze how the two relate to each other as parts of a whole. We ultimately decided that the latter was more beneficial, as it enables the user to look at what proportions of the flow are represented by specific groups such as male citizens. The resulting pie-in-a-doughnut chart makes it possible, for example, to determine exactly what groups of people come into or leave a country during a certain time range relatively to the entire flow. However, we are aware that this choice renders a temporal comparison more difficult, because the user is not capable of remembering the changes in the proportions of each category as a new time range is chosen.

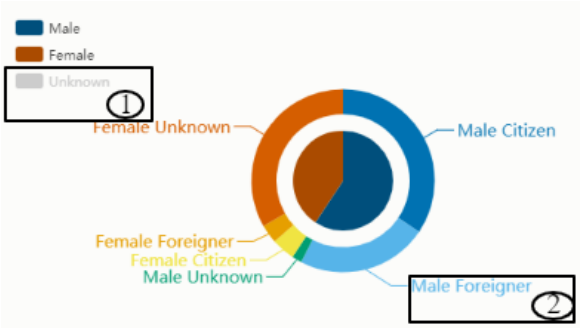

Figure 4: Visual encoding and interaction idiom of the pie-in-a-doughnut chart. 1) Deselected gender level. 2) Gender-coverage subgroup.

Visual Encoding. Despite the fact that area is not the most effective channel for quantitative data, it is useful when one wants to visualize how parts relate to a whole in a pie chart. Therefore, both the marks and the channel selected for the total flow are area. Since the parts, the types of flow, constitute a qualitative attribute, we chose color hue as the channel for this attribute. The reason is that even though spatial region is more salient than color for categorical data, it cannot be meaningfully combined with area while still creating an impression of how the parts relate to the total flow. Moreover, since at most nine different parts are shown, the maximum number of color hues that are distinguishable for humans is not reached. Thereby, we also made sure to pick a colorblind-safe color scheme [7]. Lastly, the highlighting of a selected item is achieved by making the segment extend further beyond the boundary of the respective circle. This was chosen, because a noticeable outline color that can be easily distinguished from all colors employed in the circle does not exist, and because a thicker segment outline would temporarily cover the color of small segments.

Interaction. The first way the user can interact with the pie-in-a-doughnut chart is through the

selection idiom. Specifically, hovering over one segment gives both the category name, total flow value, and percentage. Percentage values are thereby helpful as exact proportional values are impossible to be read from the size of a segment alone, and total flow values are necessary if the user wants access to exact numbers without having to calculate them based on the proportions and the total flow for the country shown in the map. Secondly, the user can interact with the pie-in-a-doughnut chart through the navigation design choice of slicing. By deselecting one of the gender categories, the specific gender segment and all corresponding coverage segments are excluded from the chart. This is useful, for instance, if one wants to focus only on flows with known gender and see the respective segments in more detail. This interaction idiom is visualized in Figure 4.

2.3.4 Scatterplots

The third important task is the exploration of correlations between factors and migration flow. Since a user is also interested in comparing the correlations of different factors and because human short-term memory is limited, we decided to create three scatterplots. These depict data points for one country for one type of flow over the entire available time range.

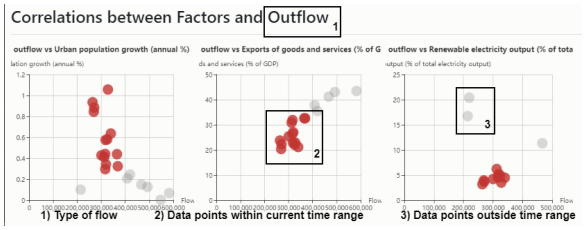

Figure 5: Visual encoding idioms chosen for the scatterplots.

Visual Encoding. This view combines two ordered attributes, which are both encoded using point marks and the position channel. This is effective, because position is the highest ranked channel for quantitative data, and because it makes it possible to identify correlations or the lack thereof. As the user may be interested in comparing the correlation within the current time range to the overall correlation, data points within the selected time range are highlighted via the color hue red, while those outside the time range are colored in light gray. The color red was chosen since it easily sticks out. While motion would be even more noticeable, it should not be used for prolonged periods of time. Finally, individual points are highlighted by enlarging their area as well as thickening their outline. The latter thereby provides a link to the world map, where countries are highlighted in the same way. Moreover, the area enlargement is safe since area is not used to encode anything else in these scatterplots. Some visual encoding idioms chosen are depicted in Figure 5.

Interaction. The user can interact with the scatterplots via the interaction idiom of selection. More specifically, hovering over an individual data point opens a small text field that contains the country's ISO-code, the year, flow type, migration flow value, factor name, and the factor value. This detail view makes it possible for the user to obtain exact values and to get a better idea of the temporal trends in the scatterplots. Yet, the view is only temporary in order not to distract from the correlations shown.

2.3.5 View linkage

All four previously described views are linked to one another as well as to the menu. Clicking on a country in the map or choosing it from the dropdown in the menu selects the country for all other views. Moreover, restricting the time range in the stacked area chart or in the menu alters the time range for the world map, scatterplots, and the pie-in-a-doughnut chart. Additionally, factors as well as the flow type can be selected in the menu. One last aspect is that links to an explanation of country ISO-codes and to the dataset sources are provided as a help to the user in the right corner of the visualization.

2.4 Algorithm

The final one of the four levels of visualization design is the development of the steps that allow the visual encoding and interaction idioms selected in the previous section to be efficiently handled by a computer [1]. In order to visualize the data, we needed to build the appropriate data pipeline, which takes the data from its raw form and processes it in such a way that it can be rendered properly. The pipeline consists of the three steps shown in Figure 6. The first stage is the *data pre-processing* step,

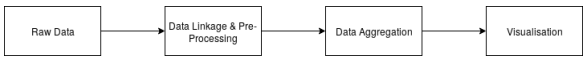

Figure 6: Data Pipeline.

which takes the raw datasets, cleans them and links the different datasets together. The second phase is the *data aggregation* step, where we sum the data across different categorical variables such as year, country, and gender. Finally, we can provide this aggregated data to our visualizations for rendering. In the following sections, we will describe these steps in detail and we will also briefly delineate how we dealt with the issue of missing data in our datasets.

2.4.1 Data Pre-processing

In order to provide the necessary data in form that can be processed by the front-end, aggregations had to be made to the basic CSVs. Because the amount of data was considerable, we chose to use a storage engine, namely Elasticsearch. We picked this engine because of its search capabilities, powerful aggregations[8], and script running features, which will prove valuable if further unplanned processing is needed. As we wanted to facilitate the adding of data, we created an extractor engine that takes the path to the CSV and certain meta-data and adds the data in the Elasticsearch indices. Since the pre-processing, even with the worker pool and bulk queries model, takes a lot of time, we decided run the setup script on a powerful machine and use data dumps to set up new instances. The result was a single document for each country and year combination that contains all the inflow/outflow/netflow totals and segmented data on each level for that combination, as well as the association and reporting country information. An example of such a document can be found in Appendix D. These documents are provided to the front-end via ReST services built on top of the corresponding Elasticsearch queries. Lastly, in order to further increase the speed, we employ caching headers.

2.4.2 Data Aggregations for Visualizations

Once the data had been pre-processed and linked, we had to aggregate the data for display in the various components of our visualization. In all cases, we simply totalled data across the different dimensions including, among others, by year, country, gender, citizenship, and factor. Many views required data to be aggregated on combinations of these different dimensions. The only exceptional consideration in this step was the handling of missing data, which we will outline in the following section.

2.4.3 Dealing with Missing Data

A major issue with the DEMIG C2C Dataset is the fact that the level of detail and accuracy in reporting varies from country to country and over time. The most substantial manifestation of this variation in the data recording is the presence of many missing values. We will briefly describe our methods for handling this issue in regards to both the data aggregation stage and the visualization stage of our algorithms.

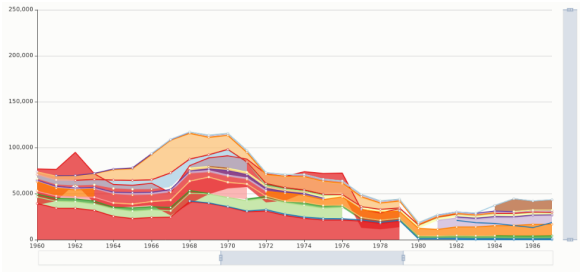

Figure 7: Stacked area chart features visible "holes" to represent missing data.

Data Aggregation. When performing data aggregation, we treat any missing values as zero, and then simply sum the data per category. For performing summations across different categories, this approach makes sense. Yet, there is a more important consideration in that we could not use aggregations such as averages, as filling missing values with zeros would skew the data. Likewise, if we tried to average only over the reported values, the aggregations would not be consistent between countries and thus, again, the visualizations would have skewed results. Therefore, we think the most accurate representation of the data and migration flow is to avoid aggregations that would be sensitive to the insertion of zero values. Moreover, the interpolation of the missing values would not be desirable either, as the data are highly variable and this could possibly lead to artifacts in our visualizations and inconsistencies in the data, especially when we are attempting to correlate real data. For our given visualization tasks, we think that a user equipped with the understanding that there are missing values can more accurately interpret the visualizations presented to them.

Visualization. For the map, pie-in-a-doughnut chart, and scatterplots, we simply do not show any points or colored area whenever data are missing. This is a relatively simple and clear solution. However, for the stacked area chart, we had a number of considerations. The first was to remove the stacking. This approach clearly highlights where data are non-existent, yet, one can no longer clearly distinguish

totals and proportions, while also obfuscating certain areas behind others. Another idea was to represent missing-values as zeros in the chart, yet, this adds artifacts to the visualization due to the stacked nature of the chart. In the end, our solution was to allow for "holes" in the stacks, which demonstrates visually where certain data are missing. You see an example in Figure 7.

3 Results

In order to evaluate the effectiveness of our visualization for the tasks formulated, we discuss one interesting finding we made for each of the main tasks, which are geographical and temporal patterns as well as correlations. Following the information seeking mantra [5], we began by looking at the total inflow values per country from 1930 until 2011 depicted in the world map. As shown in Figure 8 in Appendix B, overall, European countries have the largest total flow values. However, this is also due to the fact that very few African countries reported their migration flows, for example. Given the prevalence of news about migration to Germany, Turkey, and Italy, we did not expect Poland to have a larger total inflow than Germany does. We also noticed that Argentina has a large total inflow value compared to the rest of South America. Moreover, it surprised us that its total inflow is larger than the one of the USA and that Mexico has the largest total inflow among all countries in North and South America together.

With respect to temporal patterns, we decided to focus on Poland, because its inflow values were unexpected by us. Looking at the stacked area chart depicted in Figure 9, we found that there are two peaks in inflow, one around the year 1989 and another smaller one around the year 2007. Filtering out the unknown source countries, it was interesting to see that people coming from the USA pose more than half of the incoming migrants whose origin country was recorded for the years 1974 until today with the exception of the years 2007 and 2008. Noteworthy is also a peak of migrants from Denmark and Spain around the year 2007, as almost no migrants from Denmark or Spain are reported for earlier years. Zooming in on the time range from 2003 to 2011 to get a more detailed view as visualized in Figure 10, it is noteworthy that even though Poland's total inflow is still relatively large within Europe, it is exceeded by the one of Romania for this period of time as seen from the world map. It would be interesting to try to relate these observations to real-world events such as political or other conflicts. Furthermore, from the pie-in-a-doughnut chart in Figure 11, one can conclude that the high total inflow to Poland for this time period is mainly constituted by males and more specifically male citizens, if one disregards the migrants whose citizenship and gender were not recorded.

Lastly, we analyzed possible correlations between Poland's inflow and several factors. Thereby, we noticed that there is a positive correlation between Poland's agricultural production and inflow, with the year 1990 being an outlier with a comparatively large inflow and rather low agricultural production. On the other hand, there is no correlation between Poland's urban population percentage and inflow, but a slight positive correlation between Poland's urban population growth and inflow, with the year 1990 being an outlier again. Yet, there is not data for the latter for the years 2003 to 2011, which makes conclusions about that specific time range difficult. These results are displayed in Figure 12.

4 Conclusion and Further Perspectives

Overall, our visualization of migration flow data and possibly correlated factors is very effective in allowing the user to accomplish the tasks formulated in Section 2.1, and succeeds in capturing the workflow that results from the information seeking mantra. The three main tasks of finding geographical and temporal patterns as well as correlations are facilitated by means of a world map, a stacked area chart, three scatterplots, and a pie-in-a-doughnut chart that are mutually linked. Yet, there are a few further perspectives that could make certain tasks even easier and smoother. One of them is an option for the world map to depict total partial flows per country such as female inflow. This would facilitate a comparison of partial flows across countries and allow for the discovery of spatial patterns with respect to subflows. Moreover, while not all countries' temporal flow changes can be compared at the same time, a direct comparison of two countries' changes in flow over time might be beneficial in order to more easily evaluate one country's trend over time. Lastly, while the pie-in-a-doughnut chart provides information with respect to how flow subcategories relate to the total flow for a certain time period, a user's ability to find trends in these flow compositions over time is limited. Hence, an option for the user to choose one of the three flow composition types for the stacked area chart would be a useful addition. Further known issues are delineated in Appendix C.

References

[1] T. Munzner. *Visualization Analysis and Design.* CRC Press, 2014.

[2] DEMIG. Demig c2c, version 1.2, limited online edition, June 2015. Oxford: International Migration Institute, University of Oxford.

[3] The World Bank Group. Sustainable development goals, 2018. https://datacatalog.worldbank.org/dataset/sustainable-development-goals.

[4] Migration data portal. International migration flows, November 2018. https://migrationdataportal.org/themes/international-migration-flows.

[5] H. J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2013.

[6] Cynthia Brewer, Mark Harrower, and The Pennsylvania State University. Colorbrewer 2.0. http://colorbrewer2.org.

[7] rdrr.io. Colorblind color palette (discrete) and scales, August 2018. https://rdrr.io/cran/ ggthemes/-man/colorblind.html.

[8] Elastic. Aggregations, December 2018. https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations.html.

Appendices

A Hosted Version of Visualization Application

A hosted version of our visualization application can be found here. Please zoom out on your browser to between 60% and 80% in order to see the application the way it was designed.

B Visualization of Results

This section of the appendix contains visualizations of the results described in Section 3.

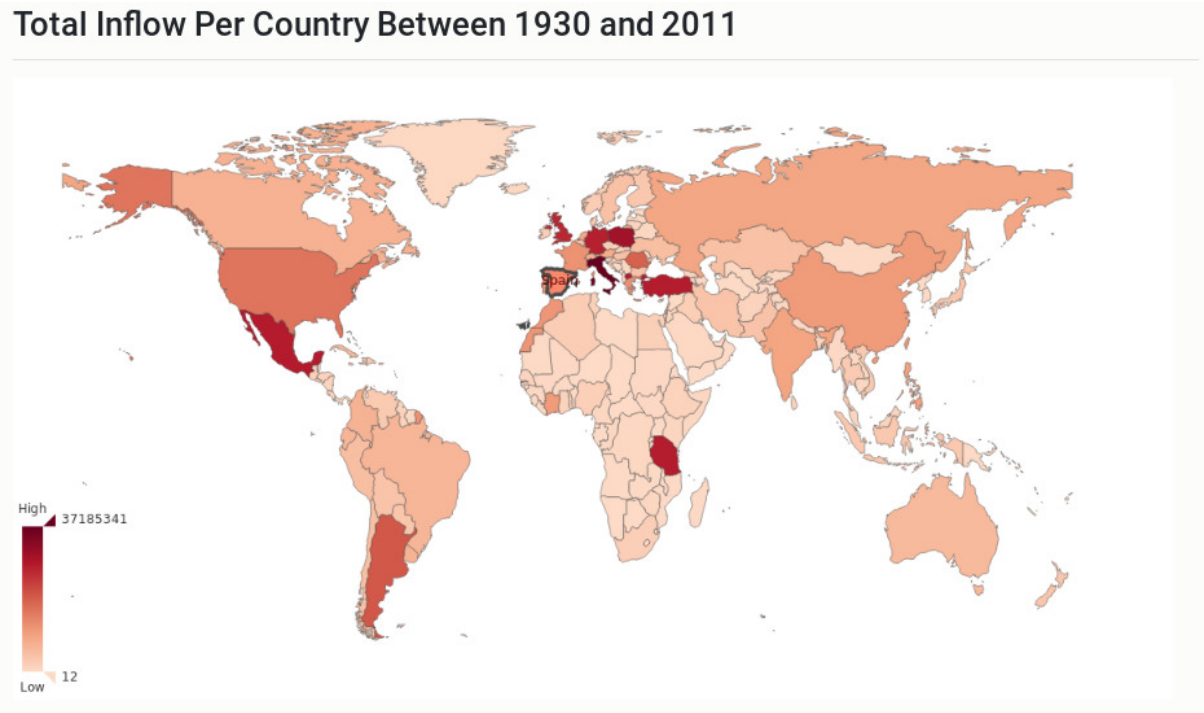

Figure 8: Total inflow per country for 1930 - 2011.

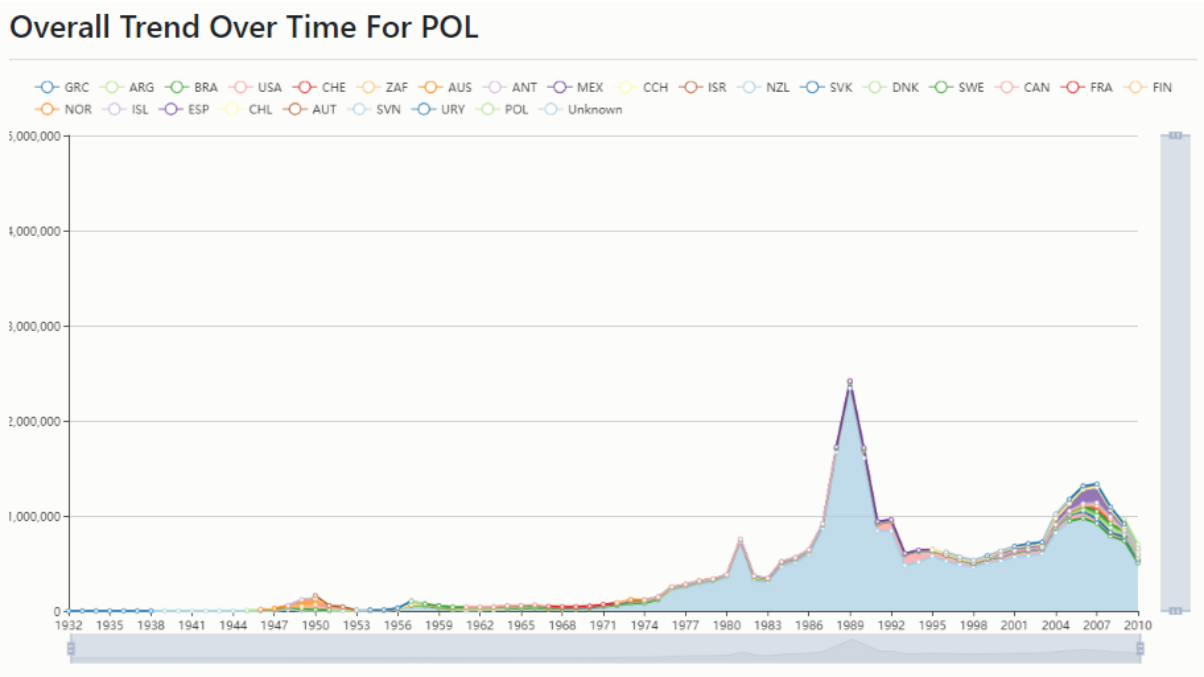

Figure 9: Poland's inflow and inflow decomposition for 1930 - 2011.

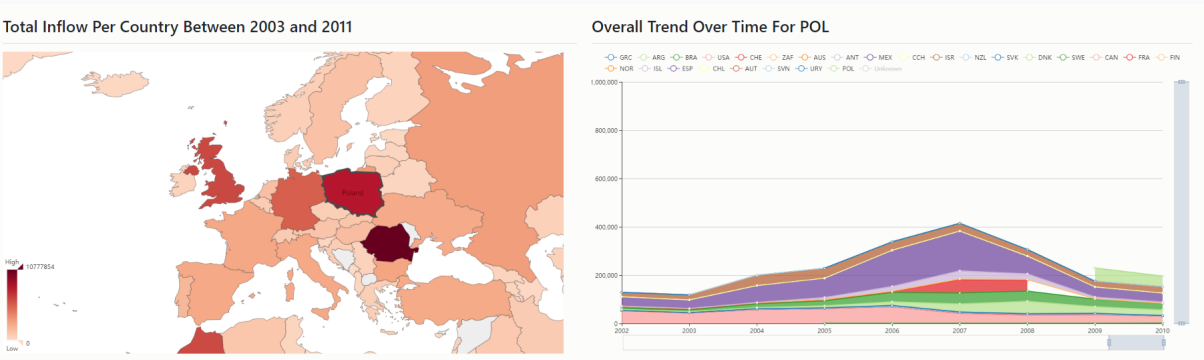

Figure 10: *Left*: Total inflow per European country for 2003 - 2011. *Right*: Inflow composition without unknown source countries for Poland for 2003 - 2011.

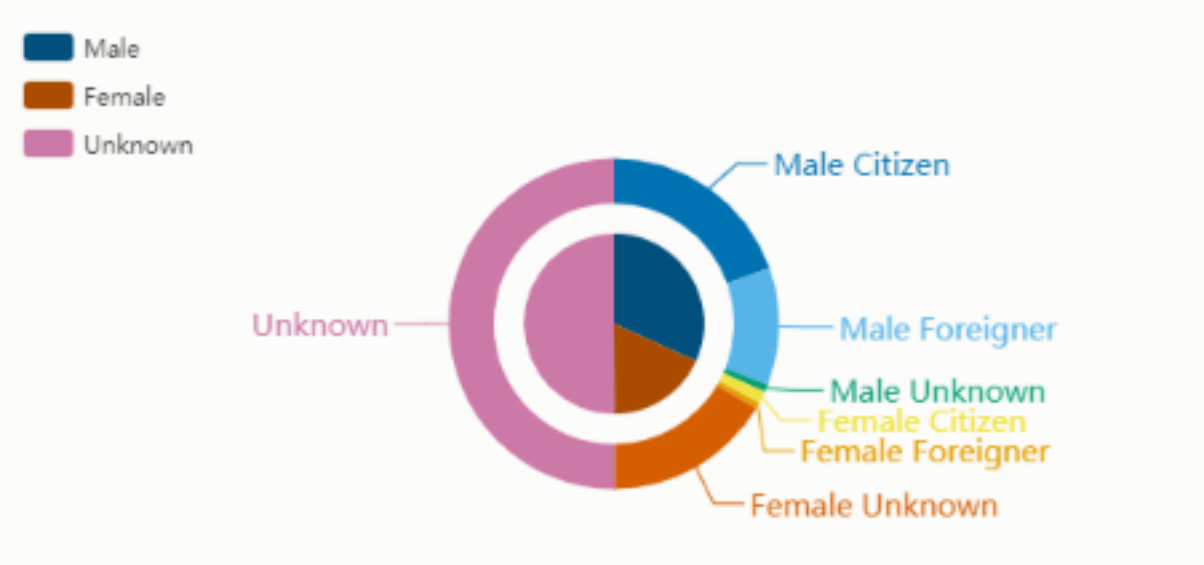

Figure 11: Composition of Poland's total inflow from 2003 - 2011.

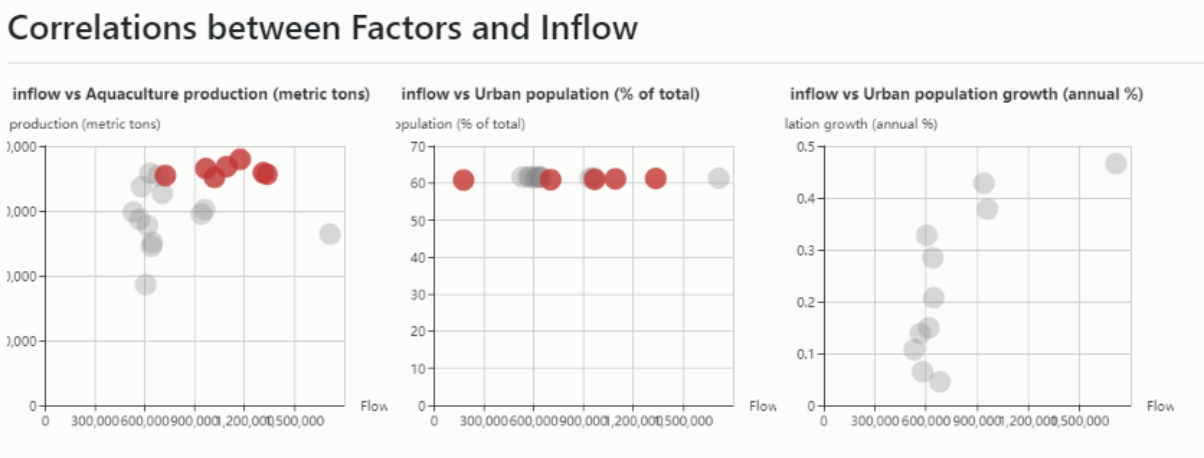

Figure 12: Correlation between Poland's inflow and (1) agricultural production, (2) urban population, and (3) urban population growth. Data points from the years 2003 - 2011 are highlighted.

C Known Visualization Issues

There are two known issues in our current visualization application that need to be pointed out.

Opacity change on map selection. The first one is that we were not able to disable the slight change in opacity of a selected country in the world map. In combination with the white background color, this leads to the country's area being drawn in a lighter color. This is harmful as the color of a country encodes the magnitude of the total migration flow.

Screen Resolution. The second point is that the visualization is designed for a larger screen and with the intention of seeing all components at once. As such, it cannot be viewed as intended on many standard monitors. This issue can easily be dealt with by the user if he or she uses the "zoom" feature on the browser to zoom out to between 80% and 60% depending on the screen resolution.

D Example Elasticsearch Document


```
_index" : "migration_total",
"_type" : "_doc",
"_id" : "AFG1993",
"_source" : {
  "year" : 1993,
  "countryId" : "AFG",
  "countryName" : "Afghanistan",
  "countryCC" : "4",
  "associations" : [
    {
      "name" : "Development assistance",
      "value" : 223800000
    },
    {
      "name" : "Percent labour female",
      "value" : 13.980042241971
    },
    ...
    {
      "name" : "Urban population growth (annual %)",
      "value" : 8.08132666728678
    }
  ],
```

```
  "inflow" : {
    "foreigners" : {
      "total" : 32317,
      "male" : 9012,
      "female" : 7060
    },
    "total" : 44145,
    "male" : 12325,
    "female" : 9575,
    "both" : {
      "total" : 11707,
      "male" : 3301,
      "female" : 2506
    },
    ...
  },
  "reportingCountry" : {
    "ISL" : {
      "countryName" : "Iceland",
      "countryId" : "ISL",
      "countryCC" : 352,
      "inflow" : {
        "foreigners" : {
          "total" : 0,
          "male" : 0,
          "female" : 0
        },
        "total" : 0,
        ...
      },
      "outflow" : {
        "total" : 0,
        "both" : {
          "total" : 0,
          "male" : 0,
          "female" : 0
        },
        ...
      },
      "netflow" : {
        "total" : 0,
        "male" : 0,
        "both" : {
          "total" : 0,
          "male" : 0,
          "female" : 0
        },
        "female" : 0,
        "citizens" : {
          "total" : 0,
          "male" : 0,
          "female" : 0
        },
        "foreigners" : {
          "total" : 0,
          "male" : 0,
          "female" : 0
        }
```

```
      }
    },
    "NZL" : {
    ...
    }
    ...
  },
  "netflow" : {
    "total" : 28790,
    "both" : {
      "total" : 9673,
      "male" : 2606,
      "female" : 2186
    },
    "male" : 7755,
    "female" : 6551,
    "citizens" : {
     ...
    },
    "foreigners" : {
      "total" : 19093,
      "male" : 5141,
      "female" : 4359
    }
  },
  "outflow" : {
    ...
  }
}
```

E Individual Reports

E.1 Mutual Work

During this project, all of us contributed to the initial search of interesting datasets as well as to the final selection of the migration flow dataset and the subset of world development indicators chosen for the factors.

E.2 Nele Albers

My part in this project was primarily the analysis of the domain situation, the task and data abstraction, as well as the design of the interaction and visual encoding idiom. Moreover, I wrote the entire report except for Section 2.4 and evaluated and tested the visualization.

E.3 Colm Seale

The largest component of my work was the coding of the entire front-end portion of our application, including all visualizations, controls, and the interactions between them. This required me to also perform testing, validation, and aggregation of the pre-processed data for visualization. Furthermore, I helped to finalize the design, idioms, and interactions required for the visualization of the data. Finally, I composed Section 2.4 of the report, except for Section 2.4.1, and wrote and recorded the screencast.

E.4 Mihai Bogdan Voicescu

My main contribution to this project was the building of the Elasticsearch indices and the back-end services. I also took care of the used scripts, process managers and the deployment on the cloud. Lastly, I wrote Section 2.4.1 of the report and edited the screencast.