Visualizing Students Migration in Elementary and Secondary Schools in São Paulo/Brazil

Carolina Román Amigo and Dylan Dong

Abstract— Understanding student's migration may benefit both public and private schools as it can help them to identify and understand eventual losses of students. In Brazil, the educational census provides freely available data to the level of granularity of each individual student, what can be used to trace migration flows from year to year based on grade progression. In this paper, we propose an overview panel and a dashboard to help government analysts and private school's directors to identify major student's migration. These visualizations are aimed to help users in answering questions such as: Are there an expressive student migration from public to private schools? Which schools are gaining/losing more students? To which schools are students going to/coming from? With the help of a walk through a scenario, we demonstrate that the implemented panels can answer these questions as planned. However, more elaborated overview visualizations can be developed to explore further the potential of this dataset, for example regarding the identification of migration patterns among years.

Index Terms—Visualization, Migration Flow, Education, Schools.

INTRODUCTION AND DOMAIN

This project domain is elementary and secondary education in the state of São Paulo, Brazil. As in Canada, elementary and secondary education in Brazil consists of twelve years of education for children aged from 6 to 18 years (grades 1 to 12). There are both public and private schools available, the private ones outperforming public ones regarding education quality [18, 5]. This is well known by parents and affects their school choices; it is possible to establish a link among public education quality in Brazil and number of enrolments in private schools [5]. When quality of public education increases, there is a decrease in the number of enrolments in private schools.

In the state of São Paulo, 35.5% of the schools were private in 2014. These schools find a large number of families willing to pay for their children's education, as the state has the highest income per capita of the country. As any business, private schools share this market and compete with each other for students to survive. Factors such as home-school distance, increase of tuition fees, and periods of economic growth or decline may influence the number of enrolments. Another high impact factor is the number of students from a given school who were able to get to the top three universities of the state, considering that the only admission criteria is a high score in a standardized test. Schools that specialize in training students for that test are most likely to have a large number of new students enrolments at high school, since these are the years that precede college.

As outlined above, students may migrate from school to school for several reasons. Understanding student's migration is useful for both government and private schools, because it may help them identify issues and potential areas of improvement. For example, a steady increase of student's migration from public to private schools may be a warning to the government that the perceived quality of public education is decreasing. For a private school, if they are consistently losing students around the 9th grade for another private school, this may be a warning that they are not investing enough effort in preparing students for the college standardized admission test.

Recently, the Brazilian Education Department started to make educational census data freely available on their website. The data is at the level of granularity of each individual student, what can be used to trace migration flows from year to year based on grade progression. Based on that dataset, we are proposing a visualization composed by two panels: an "Overview" panel, providing rankings of student's migration per each school and grade; and a "School View" panel, providing details on migration of students from or to a specific school. These visualizations are aimed to help schools and government analysts to answer the following questions:

Schools
- Is there any particular grade in which migration is more intense?
- To which schools are their students going to/coming from?
- Is it possible to identify a pattern in the geographic location of the main competitor schools?

Government analysts
- Is there an expressive student migration from public to private schools (or vice versa)?
- Which schools are gaining/losing more students?
- Which grade is gaining/losing more students across schools?

With the help of a walk through a scenario, we demonstrate that the implemented panels have the potential to answer all the questions listed above as planned. However, more elaborated overview visualizations can be developed to explore further the potential of this dataset, for example regarding the identification of migration patterns among years. We indicate some possibilities at the future work section.

This paper is organized as follows: a related work section describing the most relevant visualization solutions employed for representing migration flows in the literature; data and task abstraction section explaining how we abstracted the domain specific data and tasks; a solutions section explaining the visual encodings and design decisions in the visualizations; an implementation section providing details of the coding process; a results section providing a walk through the prototype with the help of a scenario; a discussion section highlighting strengths and weaknesses; a lessons learned section with what we learned through the process; and finally a conclusions section summarizing the project and its results.

Carolina Román Amigo is with iSchool at University of British Columbia. E-mail: carolamigo@alumni.ubc.ca.
Dylan Dong is with Computer Science Department at University of British Columbia. Email wdong@cs.ubc.ca.

1 Related work

Origin-destination datasets e.g. flows of people, animals, traffic, knowledge, disease, etc. typically have a complicated structure and a very large scale. The challenge of representing this kind of information effectively has for long been a concern in the literature: the first example of geographic flow visualization was produced by Ravenstein [18] in 1885. He drew the flow of people around Great Britain and Ireland by means of a series of single headed arrows, crossing county boundaries and typically flowing towards major urban centres, a classical way of representing migration still largely used nowadays. Seventy four years later, the Chicago Area Transportation Study produced the first computer based flow mapping example [3]. Since then, computational advances have been making larger migrations datasets more accessible, and diverse visualization idioms have been proposed.

Boyanin et al. proposes Flowstrates to help users perform spatial visual queries and analyse changes over time [1]. They display origins and destinations of flows in two separate maps, and flow magnitudes changes over time are represented in a separate heatmap view in the middle (Figure 1). Querying, filtering, ordering and grouping techniques are used to help interactive exploration. The idiom is useful for both providing an overview of the dataset and focus in a specific location or period of time. It has a good scalability regarding the number of years that can be represented and minimizes cluttering by representing intensity of flow in a separate view instead of using the stroke width to encode that information.

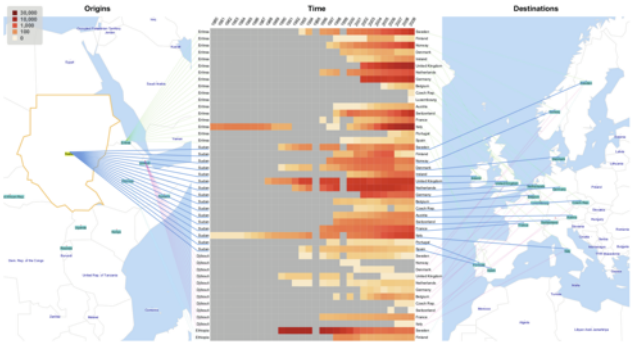

Figure 1 - Flowstrates example showing refugees migration among East Africa and Western Europe from 1980 and 2009 [1].

In order to prevent losing details and introducing arbitrary artefacts in the visual representation, Wood et al. propose a method which maps the origin-destination vector as cells, in contrast to lines used by other methods [25]. They project geographic data on a set of spatially ordered small multiples by constructing a gridded two-level spatial treemap. This idiom is better for providing an overview of the dataset in cases where the vector between a pair of locations has greater importance than the geometric path among them.

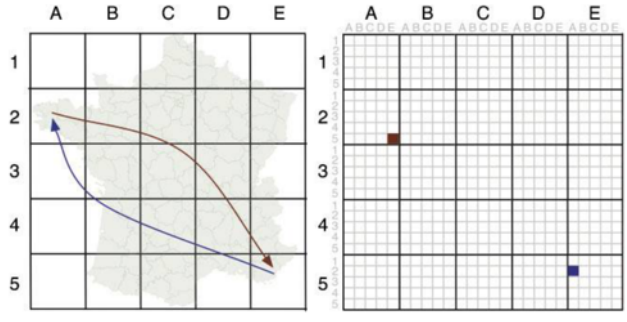

Figure 2 - Geographic space partitioned into a regular grid (left) and an origin-destination map space (right) [25].

Rae uses flow density maps to visualize a large migration matrix from the UK's 2001 census [17], using a GIS application. In Figure 3, they use a coloured scale (varying the hue) to show line density on the map (more lines, more migration paths). In an overlay, they use a combination of lines and marks to show flow intensity (varying stroke width) and marks size to encode total migrants number. This view is good to understanding general patterns of movement, to show specific linkages between places and to spot where the highest levels of mobility exist.

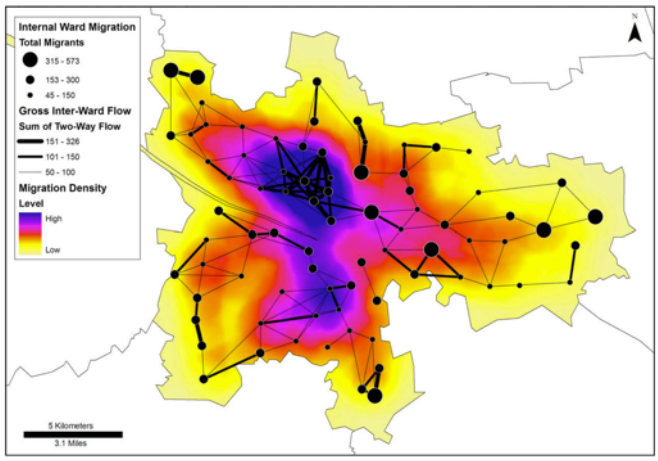

Figure 3 - Intra-city migration in Glasgow [17].

Similarly to Rae [17], Gilbert et al. use statistical summaries of spatial association to visualize the movements of animals infected by bovine tuberculosis on the flow density maps [7]. They explore the association between bovine tuberculosis occurrence and the predictors by conducting a stepwise multiple logistic regression analysis of 2002 and 2003 bovine tuberculosis distribution data.

Verbeek et al. propose a method based on spiral trees, a type of Steiner tree which uses logarithmic spirals, to visualize flow maps [2]. They integrate edge-bundling to their algorithm and compute crossing-free, merge smoothly, and naturally cluster flows. The high-quality flows are produced by minimizing a global cost function which consists of obstacle cost, smoothing cost, angle restriction cost, balancing cost and straightening cost. An example is depicted in Figure 4.

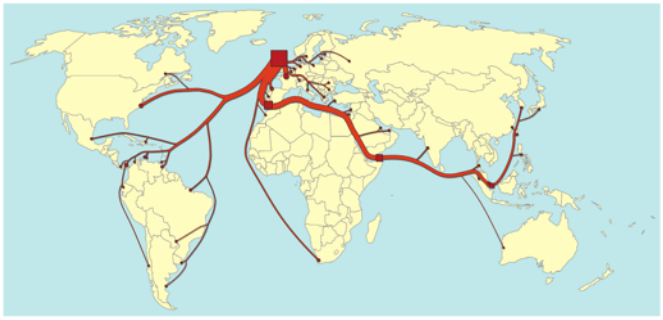

Figure 4 - Top 50 whisky exports from Scotland in 2009 by volume, example of flow mapping using spiral trees [2].

Phan et al. present a method to draw flow maps based on hierarchical clustering [16]. Their system consists of two phases: layout phase and rendering phase. They use distortion to ensure the nodes are well spaced but still preserve their relative positions to the neighbours. The edges are merged based on their destinations using hierarchical clustering. They use the spatial information given by the hierarchical clustering to do edge routing to avoid edge crossings. An example is depicted in Figure 5, compared to other two types of flow maps. Edge-bundling algorithms based on hierarchical information

[9], geometry information [4], force-directed algorithm [10] and quadtree structure [12] are also used to visualize origin-destination data because they can reduce visual clutter by merging edges.

Figure 5 - (a) Minard's 1864 flow map of wine exports from France [21] (b) Tobler's computer generated flow map of migration from California from 1995 - 2000. [19; 20] (c) A flow map proposed by Phan et al. [16] that shows the same migration data.

In addition to the above mentioned single-view methods, Guo uses multi-view displays to visualize migration flows [8]. The methodological framework consists of methods for hierarchical regionalization, flow mapping, multivariate clustering and visualization. The multi-view displays use a self-organizing map, parallel coordinate plot, and a flow map to present flow structure, multivariate information, and spatial patterns at the same time.

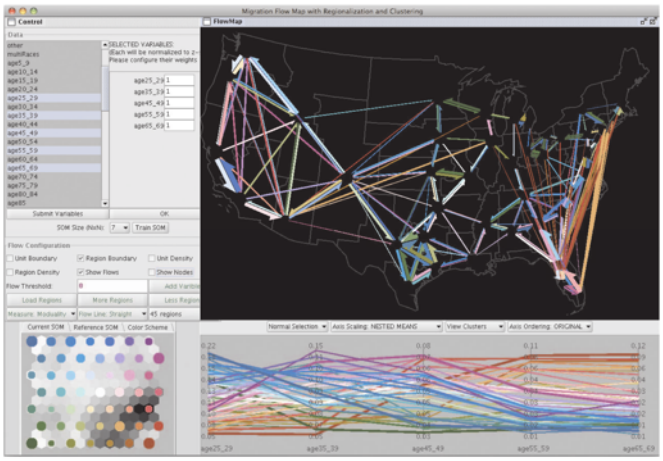

Figure 6 - Multivariate flow mapping using self-organizing map (bottom left), parallel coordinate plot (bottom right) and a flow map (top right) [8].

Parallel coordinates can be a good alternative to flow maps to show migration over time when the geographic location is less important than the link among locations and flow intensity. It has been applied to many multidimensional problems and has been incorporated into many commercial and public-domain systems, such as WinViz [15] and XmdvTool [23]. Fua et al. enhance the parallel coordinates technique by developing a multi-resolutional view of the data via hierarchical clustering [6]. They make use of variable-width opacity bands to represent the information at a node. They also use a proximity-based coloring scheme to guarantee that data and clusters from similar parts of the hierarchical structure are shown in similar colors. Novotny et al. integrate focus+context visualization in the parallel coordinates [14]. After binning the data into different levels of detail, they can visualize context information at several levels of abstraction while leaving enough visual resources for the outliers and for the data items in focus.

Parallel sets offer the possibility to also encode the amount of migration flow among two locations, using the size channel as Kosara et al. [11] shows in Figure 7. While parallel coordinates represent categories by points on continuous axes, parallel sets uses sections of the axis to encode frequency, thus being able to represent frequency and relations in the same view.

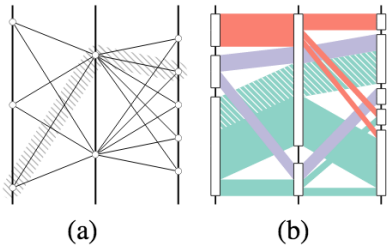

Figure 7 - (a) Parallel coordinates represent categories by points on continuous axes; (b) Parallel sets show the frequencies of categories and relations [11].

2 Data and Task Abstractions

Herein we describe the data we used, selected users and their tasks, and explain the data and the task abstraction process.

2.1 Data

We used data from the Brazilian Educational Census, which is released every year and it is public available at the Brazilian government website. For each student of the country, this census shows in which school they were studying in that year, in which grade they were, and the school type (private or public), among other related information. It is a huge dataset; just for the state of São Paulo we have 10,581500 students and 28,718 schools, as of 2014. The actual number of students we are going to work with is larger as we are going to use data from census since 2012 until 2014, in the order of 10,000,000 rows each.

We used two tables from the census dataset: "Enrolments" table which has enrolment_id as a primary key, and has all student's enrolments for that census year; and "Schools" table, which has school_id as a primary key and lists all the schools in the state, with names, latitude and longitude information, among others. Herein is a table summarizing the fields we used in our project.

Table 1 – Dataset fields selected from the source files

Source table	Field name	Description
school	school_id	primary key, id for each school.
	school_name	the name of the school
	school_city	foreign key, city code of the school
	school_district	district code of the school
	post_code	postal code of the school
	latitude	geographical coordinates of the school
	longitude	
	school_status	status of school (active, inactive)
enrolment	year	enrollment year of a student, from 2012 to 2014
	enrollment_id	primary key, id for each enrollment item
	student_id	id for each student
	education_grade	student's educational grade
	school_id	id for each school
	school_city	city code of the school
	school_type	type of the school, public (federal, state, city) or private

2.2 Users and Tasks

One of this project authors works for a company that is in the preliminary stage of development of tool for visualizing geo-referenced data about primary and secondary schools in Brazil using the educational census. Although migration is just a small part of the tool they are envisioning, they collected data about users needs and requirements that are useful for this project as well. We had access to

a summary this material and used it as a starting point to define stakeholders and their tasks.

Two different types of stakeholders, schools and government, have interest in understanding migration of students, but in different granularity levels; while for the government is interesting to get a ranking of migration among schools, schools are mainly interested to understand the specific migration flow of students from and to them. They can use the visualizations to identify largest migrations, compare migration among grades and schools, spot outliers and understand the migration network. The specific tasks that are probably most relevant to them and that are thus supported by our proposed visualization are:

Task for schools

- Is there any particular grade in which migration is more intense?
- To which schools are their students going to/coming from?
- Is it possible to identify a pattern in the geographic location of the main competitor schools?

Task for government

- Are students migrating from public to private schools (or vice versa)?
- Which schools are gaining/losing more students?
- Which grade is gaining/losing more students across schools?

Both types of users, specially school's directors, are not very technology oriented or computer savvy. Thus the visualizations should to be simple to use and preferably use interaction strategies similar to the ones they are already familiar with, such as spreadsheets with ordering and filtering features.

2.3 Data Abstraction

We used the three-part analysis framework for a vis instance to help us in the data and task abstraction process [12]. Figure 1 summarizes data abstraction (what). We have the following dataset types:

- Network: showing the migration flow to and from a selected school.
- Tables: bar and column charts showing the total number of students leaving and entering each school per grade in a given year and students migrating to and from a given school, as well the flow balance.
- Geometry: showing geographic location of schools in a map (school view panel).

The dataset we have access to is static. We have both ordered attributes (grades) and quantitative attributes (total number of students migrating to and from schools, balance of migration flow). The ordering direction is both sequential (ascending list by migration total) and diverging (balance of migration flow).

2.4 Task Abstraction

Users will be probably using our visualizations to identify largest migrations, compare migration among grades and schools, spot outliers and understand the migration network (from where and to where students are going). As we are dealing with a large dataset, users will also use derived dataset to get a summarized view of the migration flow, for example total number of students schools lost in a given year. Users can also explore the visualization to discover common points among schools, for example, a specific grade which always shows a large outflow of students.

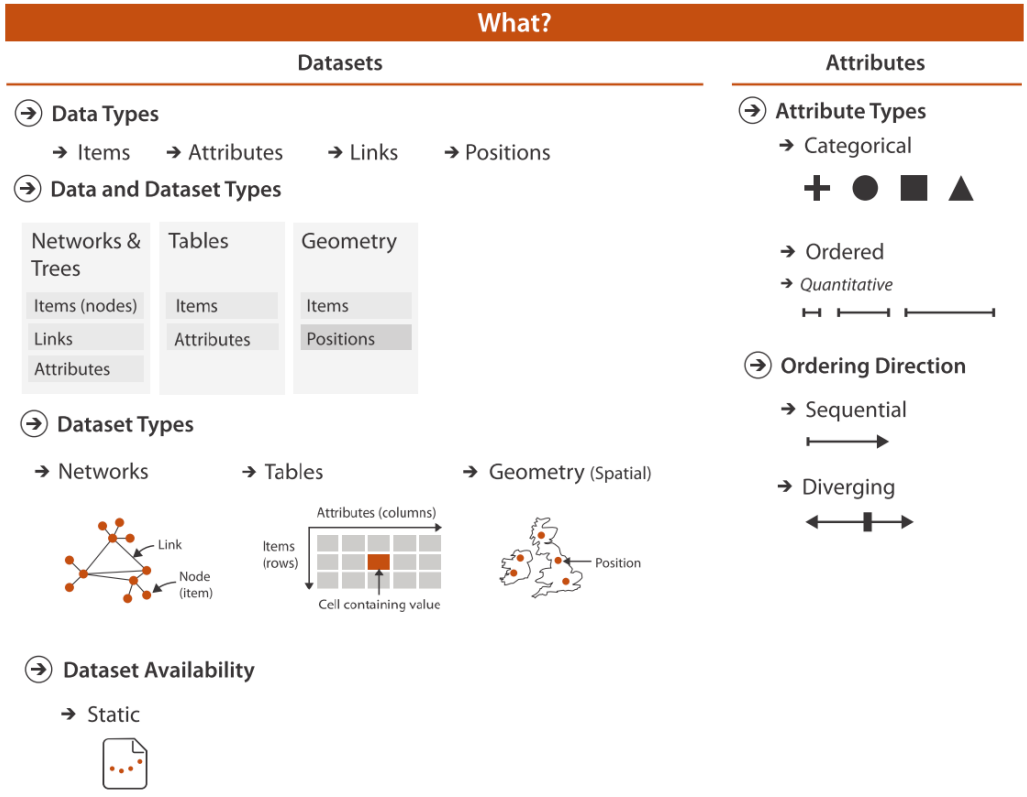

Figure 8 - What data the user will see in our visualization (adapted from Munzner, 2015).

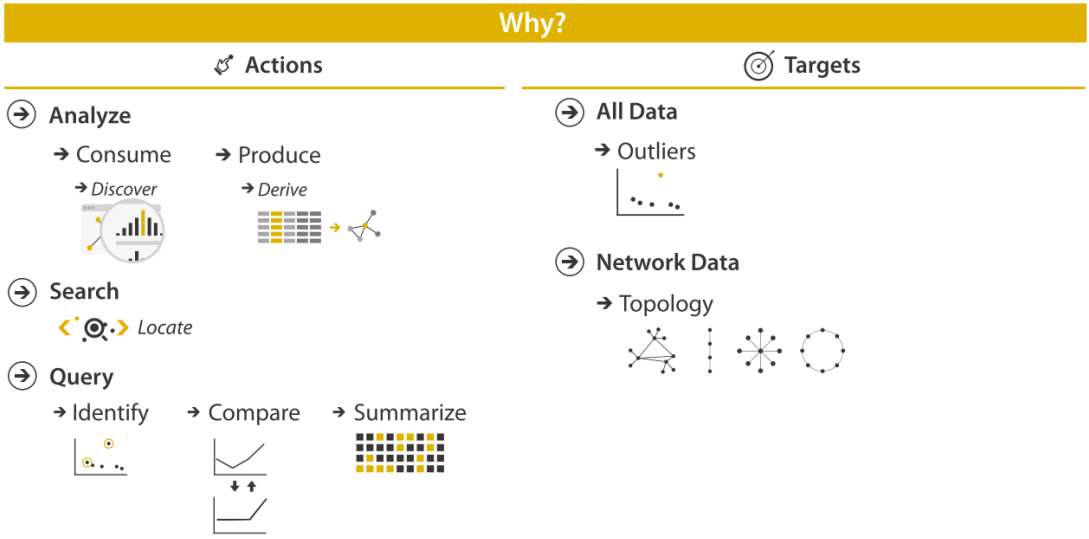

Figure 9 - Why the user will intend to use our visualization (adapted from Munzner, 2015).

3 Solution

Figure 10 summarizes how the visual encoding was constructed. To cope with the size of the dataset, we opted to make use of a combination of navigation, facets and filtering. The main function of the "Overview" panel is to show a ranking of student's migration among schools. To visualize the data, users have to define a pre filtering informing year, focus school type, flow type (inflow, outflow, balance), and type of competitor schools (private, public, both). Once in the "Overview" panel, the list is paginated; users can scroll and navigate among pages it in order to see the whole list of school. The number of student's migration (inflow, outflow, balance) is encoded by the length of the bar. For inflow and outflow data, the bar is drawn from the left side of the table cell. For balance data (inflow - outflow), positive value is represented by a bar drawing from the left side of the table cell while negative value is represented by drawing bar from the right side of the table cell.

Users can hover the mouse pointer over the bar to see the exact number of students of each bar. They can filter the data by the total flow number, school name or school id. They can order the overview list in ascending or descending order by grades per school. Colour hue is used to reinforce the different grades. Once users find a school of interest in the "Overview" panel, they can take note of the name of the school in order to look for it in the detailed dashboard ("School View" panel).

The "School View" panel shows the detailed information about a specific school selected by the user. We have a dashboard offering three facets of the dataset: a parallel coordinates graphic showing the size of flow to and from a school, per grade; a bar chart showing the total inflow, outflow or flow balance per grade; and a map showing geographic migration, also per type of flow. The parallel coordinates show the top ten schools from which the students are coming to the focus school, and the top ten schools to which the students from the focus school are going. These schools are aligned in descending order by the number of the students leaving or arriving. We only show the top 10 schools because they are more representative than the schools which have small migration number. The total number of schools in each axis is also shown at their top to help users see how many schools are interacting with the school they searched.

The grades are represented by ellipses and schools are represented by rectangles. If the school users selected has inflow or outflow from other schools at a certain grade, we connect the schools and the grade by a single line representing a link. The flow number is encoded by the thickness of the line. Sometimes there may be many lines between school's axis and grades axis, which can make it harder for the user to follow the link. To make it easier to follow, we highlight the line together with its corresponding schools and grades when the user hovers the mouse pointer over it.

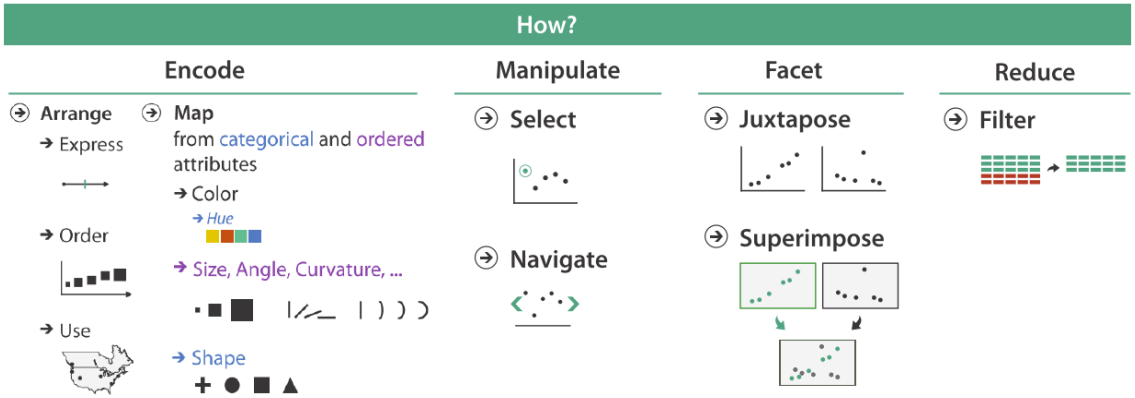

Figure 10 - How the visual encoding will be constructed in terms of design choices (adapted from Munzner, 2015).

The column chart in the school view panel shows the total inflow, outflow and balance of students for the focal school at each grade. We aligned the three charts vertically to help users to compare data easily. Users can hover over the grades in the parallel coordinates graphic to highlight the corresponding bars in the bar chart. Users can also click on inflow column chart and outflow column chart at a certain grade to open a pop-up with a node-link graph representing all the schools related to the focal school in the migration network. The flow number is encoded by the thickness of the link.

The geographic view shows the geographic distribution of the corresponding schools. We encode the flow type by colour and school types by different marks shapes and colours in the map. Users can click on a certain grade in the parallel coordinates to see the geographic location of the corresponding schools in that grade.

4 IMPLEMENTATION

The visualization system was built in Flask, which is very lightweight and reduces the amount of back-end work. The back-end code was written in Python. The front-end code was written in JavaScript and HTML. The visualization system runs on Mac OS platform. To accelerate the querying speed, we made use of indexes and joined the tables together in MySQL. We get the inflow and outflow of students per grade per school by using SQL statements, then we export the data to the csv format.

We added jQuery event handlers (click, mouse over, mouse out, change, etc.) to notify the back-end how the users are interacting with the front-end interface. The Ajax application sends the request data in the JSON format to the back-end, and the request is processed by the function binded to a particular URL by the route() decorator in Flask. The csv format data is loaded by Python in the dictionary. For the school information, the key is the school id and the value is the school's name, type, location, etc. The inflow and outflow information of each school is also stored in the dictionary format, where the key is the school id and the value is a nested array which stores the corresponding schools and the number of inflow, outflow of students per grade.

For the "Overview" panel, we developed it without using any third party libraries. The table in the overview panel is actually a template. After the data which is going to be displayed in the table is processed, the template is rendered by the render_template() function in Flask and sent back to the front-end. Then, Ajax receives the rendered data and binds it to the desired div.

For the "School View" panel, three HTML templates were used: parallel coordinates, total flow per grade and geographic view. The parallel coordinates and total flow per grade are drawn by d3.js. After the back-end gets the requested data from the front-end, the back-end function provides the data to be displayed and sends it to the JavaScript function which draws the graph by d3. Then, the HTML template which owns the JavaScript function is rendered and binded to the desired div. The geographic view is built on Google Map API. The back-end provides the data in the appropriate format and sends it to the Google Map API. Then, the map is rendered and binded into the desired div. The linkage between the three views in the "School View" panel is implemented by jQuery event handler.

5 RESULTS

With the help of scenarios, we demonstrate how the implemented visualizations may be used to answer the questions listed in item 2.2. First, we walk through the "School View" panel, using a private school scenario; then, we walk through the "Overview" panel, with a government scenario.

5.1 Private school trying to understand student's losses

A school director is noticing the number of enrolments falling in recent years. He would like to know the reason for students to be leaving his school, and it would help to know where their students are going to. So he decides to use the migration vis tool we are proposing. As he is interested in a specific school, he clicks on the "School View" panel (Figure 11), and searches for his school using the search field. After that, he selects the year he is interested in (2013), and sets to "private" the competitor type of school his school is receiving/sending students to, because this is the type of school he shares the market with. The visualizations appear in the dashboard a few seconds after he hits the "Go" button.

On the parallel coordinate view (Figure 12), he has an overview of the inflow/outflow of students per grade. In the middle axis, there are the grades; on the left one, the schools from which his school is receiving students; on the right axis, the schools for which his school is sending students to. Hovering over a link, he can highlight the corresponding school and grade. He notices that there are 21 students moving to school "35106756" at the 8th grade (Figure 13).

He notices that, in general, a large amount of students is leaving at the 8th grade. So he clicks on 8th grade to select it, and then the information about that grade is highlighted on the column chart view. Also, the schools to which the students at the 8th grade are moving to are shown in the geographic flow view (Figure 14).

On the column chart at the top right, he can confirm in the "balance" section that he is losing a large amount of students at the 8th grade, more than in any other grade. Clicking on the 8th column of the outflow chart, he can see all the eight schools his students are moving to (Figure 15). On the geographic flow view at the bottom right, he can see the geographic location of the schools for which he is sending students to in the 8th grade (Figure 16). He can see the details about the schools by clicking on the pins (Figure 17).

He identifies a pattern that none of the students go to schools located at east or west side (Figure 16). It's probably because there is less traffic going south or north in the city than going west or east. Another potential reason may be there are real state being developed in the north-south regions with an affordable price, so people may be moving to these regions and changing schools as well. With these assumptions in mind, he searches on the internet to see why.

5.2 Education department trying to understand student's evasion for private schools.

An analyst from the Education Department of the State of São Paulo is working on a project to improve education quality in elementary and secondary public schools. He wants to identify specific schools in need of interventions. He knows that middle-class parents with limited budget for paying for their children's education use a combination of public and private education. So he would like to know which public schools are having the largest evasion to private ones, and if there is a pattern of evasion, for example one specific grade at which the evasion is higher.

He decides to use the migration vis tool we are proposing. As he wants to have an overview of schools, he selects the "Overview" panel (Figure 18). Then he selects the year he is interested in (2013), the type of focus schools he is exploring (public ones in this case), and the flow type (outflow as he is interested in evasion). Finally he selects the type of school students are going to (private).

A bar chart with all grades evasion for each public school appears in the screen after a few seconds (Figure 19). He can split the information by pages and also filter it, as it is large. He chooses to filter the total migration per school to more than 40, as he is only interested in the largest migrations (Figure 20). He can order the list by grades, in ascending or descending order, by clicking the arrows in the grade's columns headers (Figure 21). He can hover over the bars to see the exact number of students that are leaving each school (Figure 22). Among the schools at the top of the list, the analyst identifies one that is having a large outflow of students across almost all its grades. He clicks on it to see details on the "School View" panel (Figure 11, panel already explained in scenario 7.1).

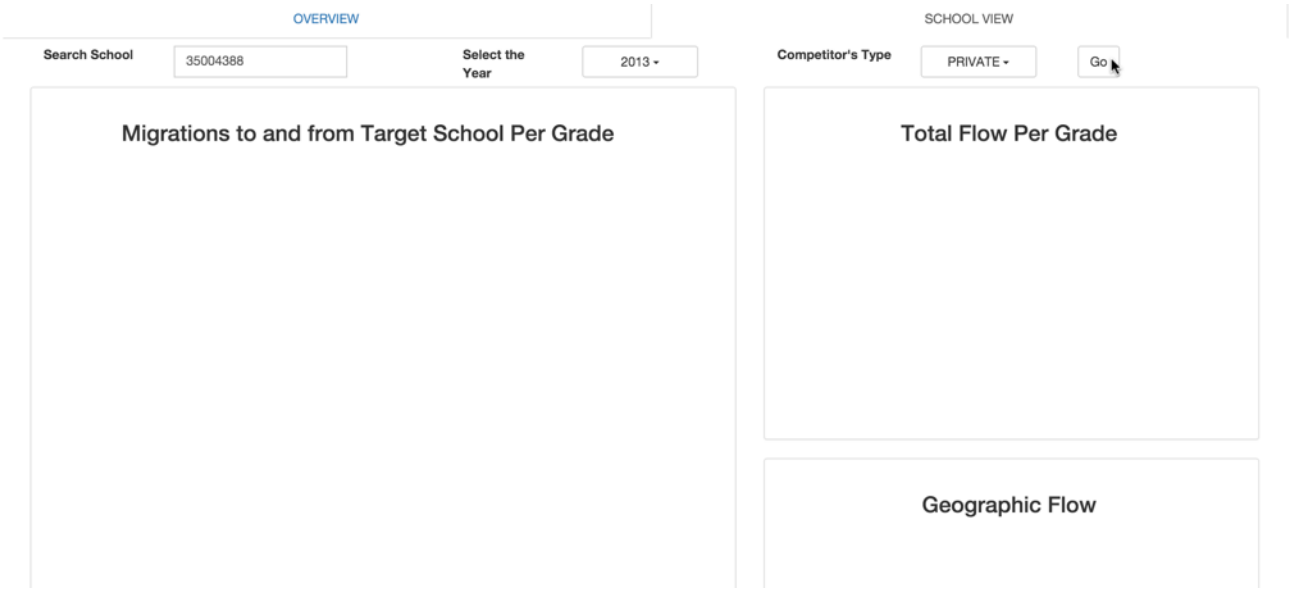

Figure 11 - “School View” panel before the user selects the information he wants to see.

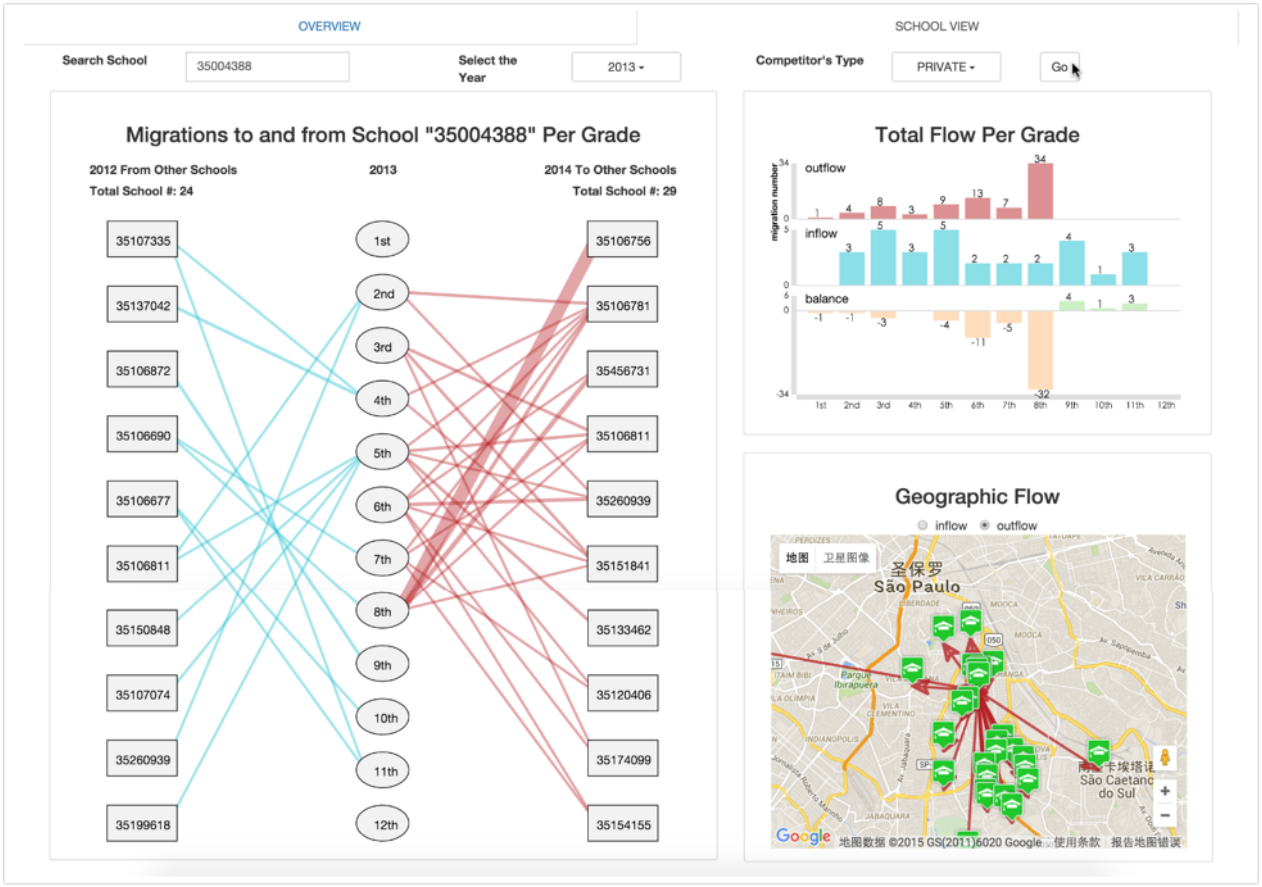

Figure 12 - “School View” panel showing information for School “35004388”, for all grades.

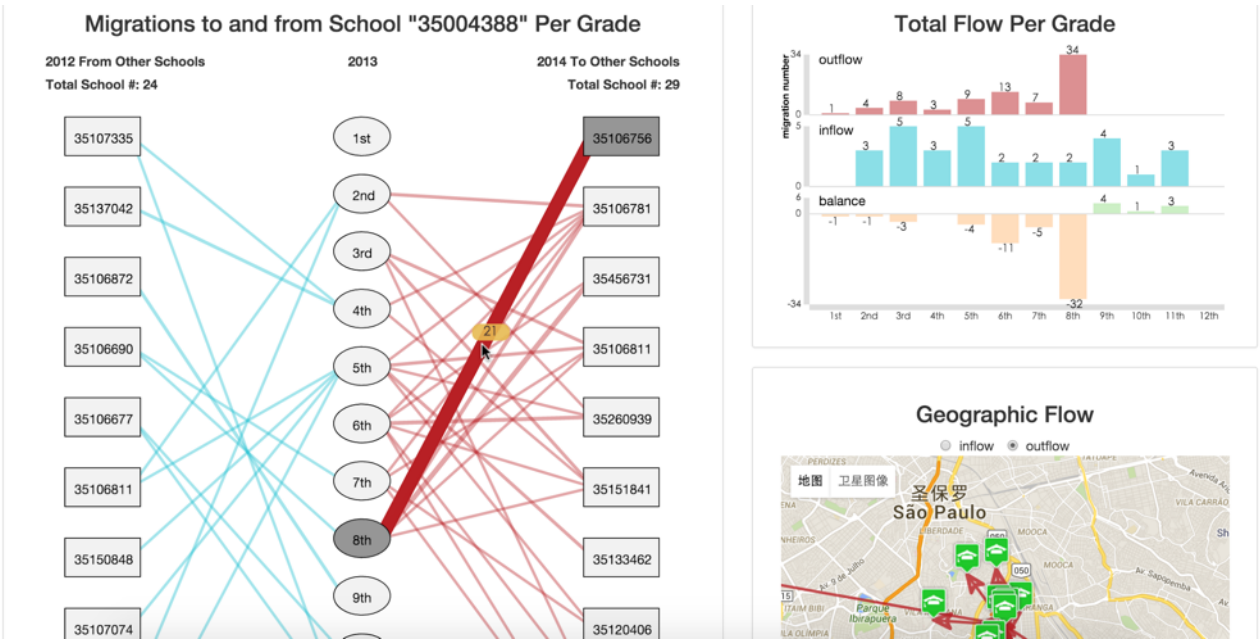

Figure 13 - Major outflow of students highlighted in the parallel coordinate view, by hovering the mouse.

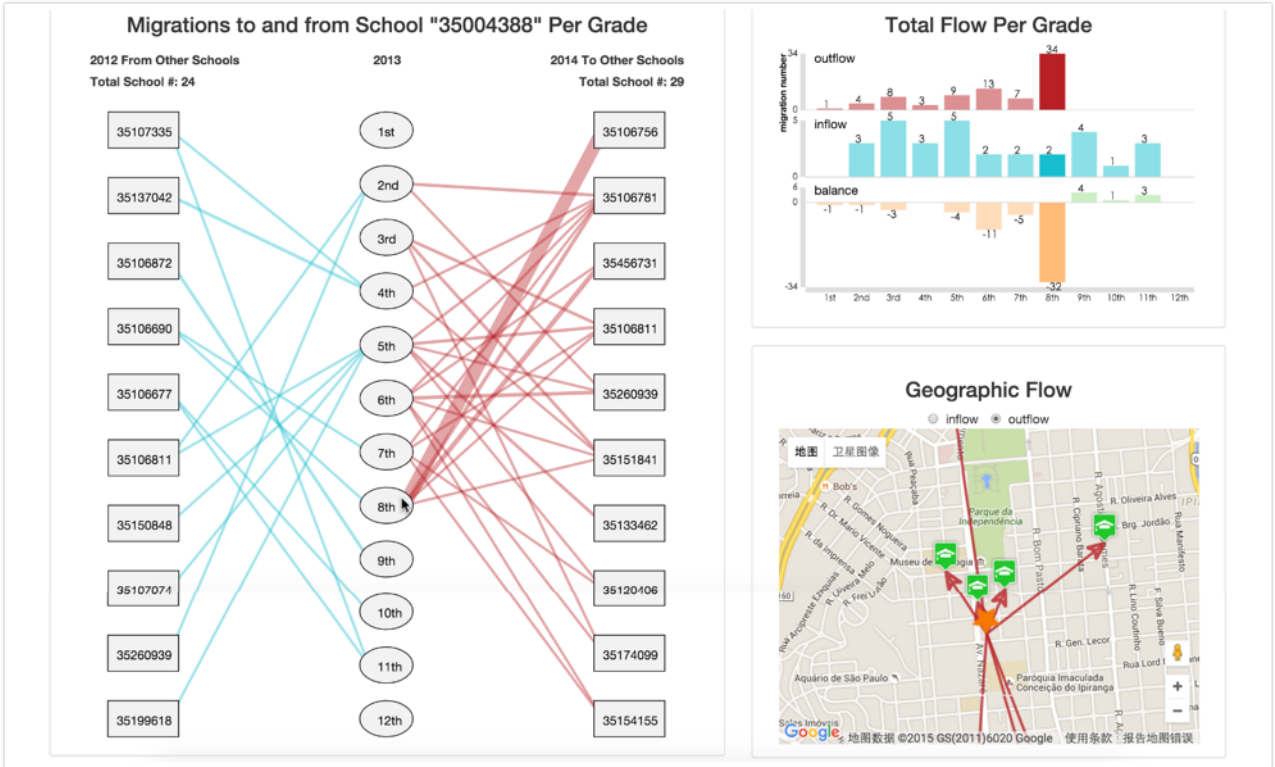

Figure 14 - Clicking on a given grade (in this case 8th) highlights the schools they are going to in the map view and in the column chart.

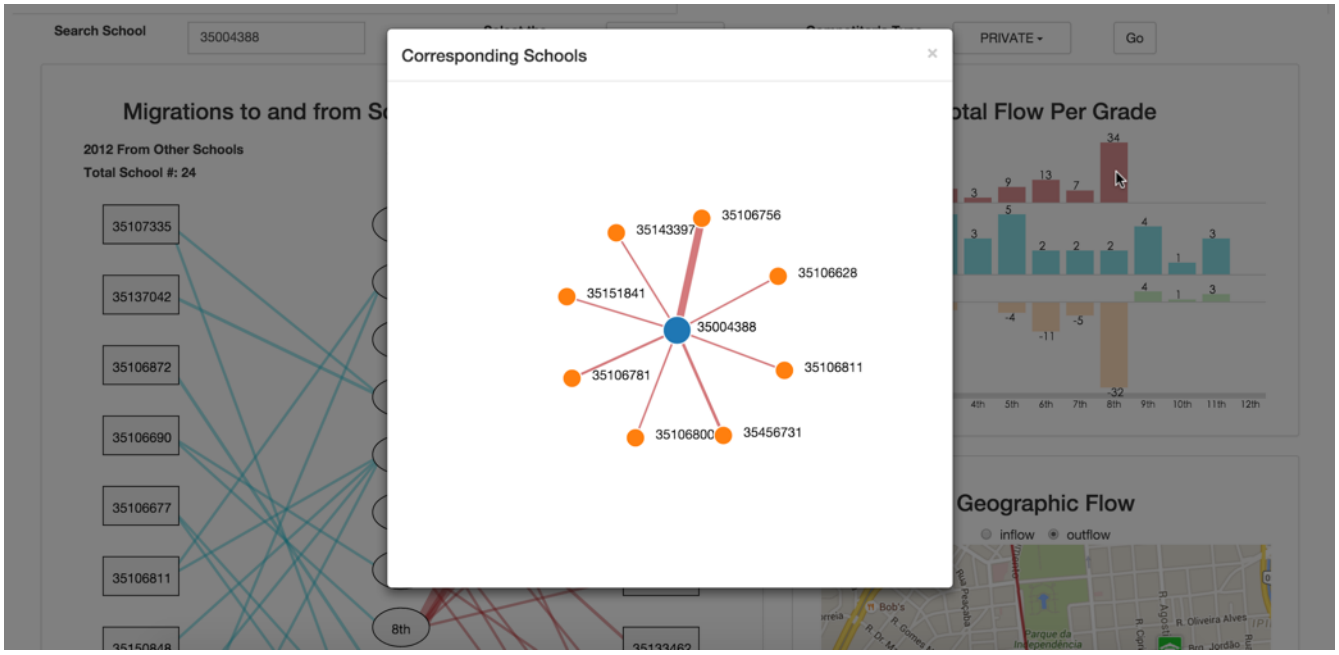

Figure 15 - All schools for which the focal school is sending students to in the 8th grade.

Figure 16 - Geographic location of the schools the focal school is sending students to in the 8th grade.

Figure 17 - Clicking on the pins, users can see details about the schools.

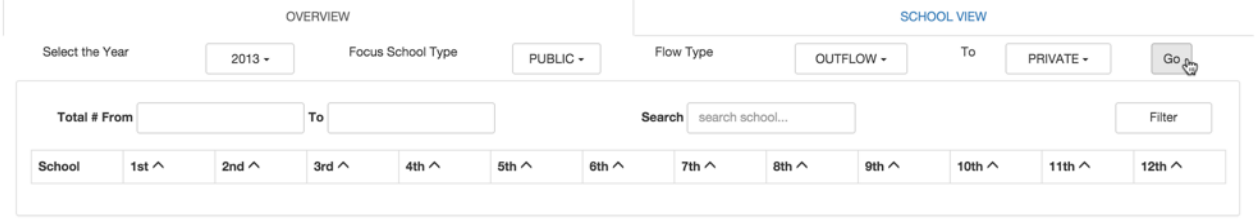

Figure 18 - Pre-filtering options for the overview panel.

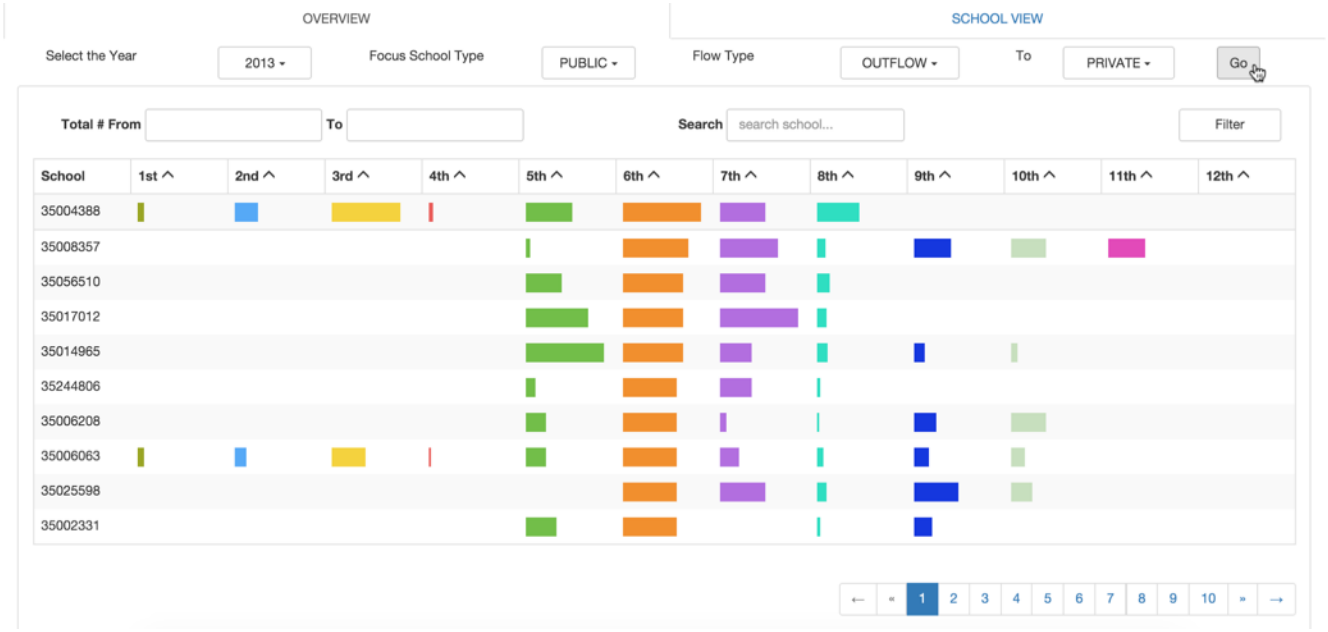

Figure 19 - Overview panel showing results for outflow of students from public to private schools in 2013.

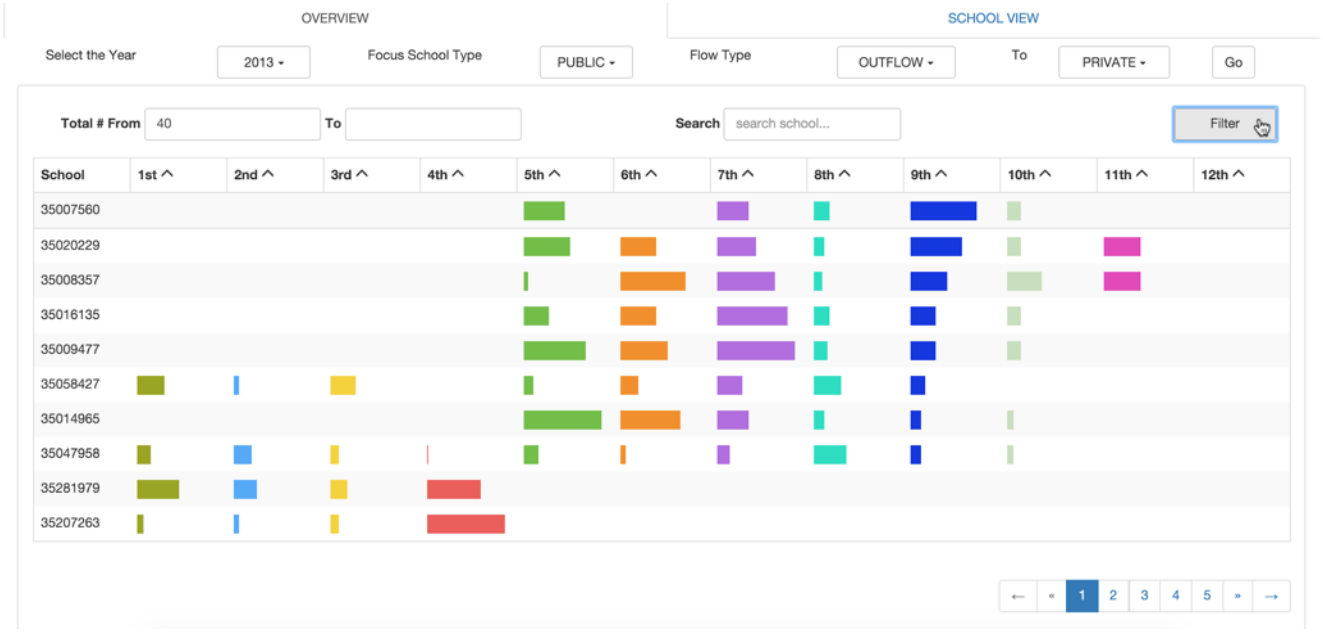

Figure 20 - Users may filter the results by more than 40 students migration, for example, in order to show only the most significant outflows.

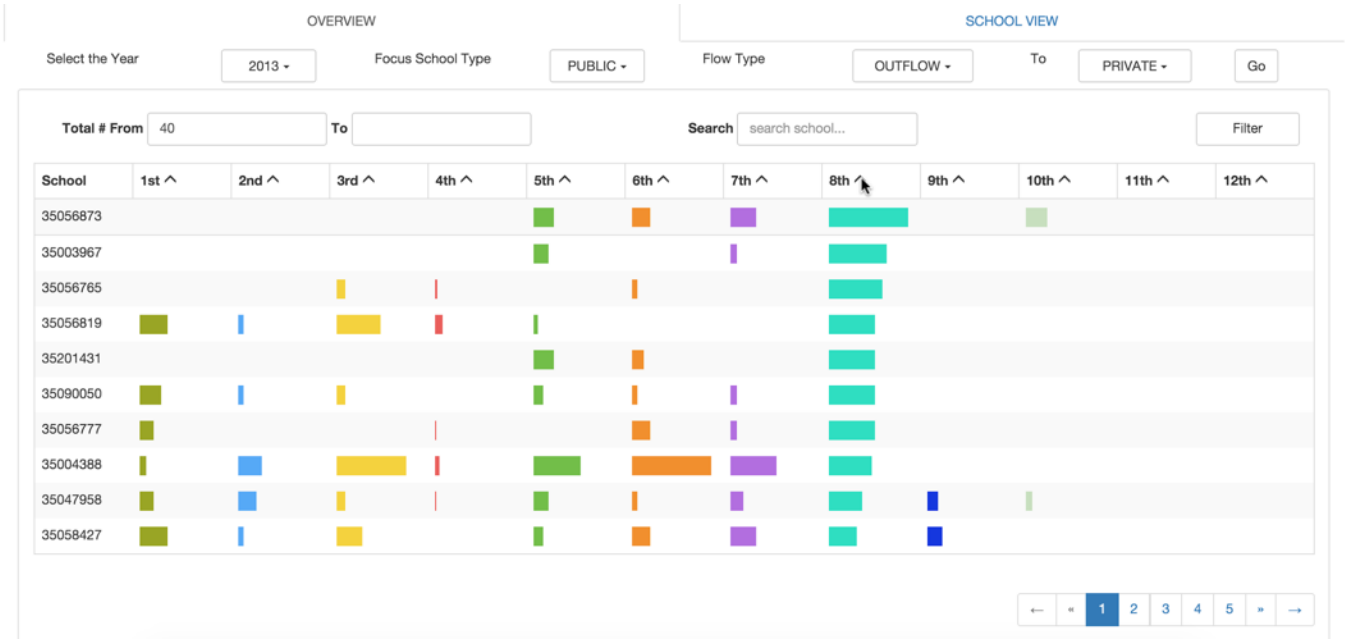

Figure 21 - Users can order the results in ascending or descending order by grade.

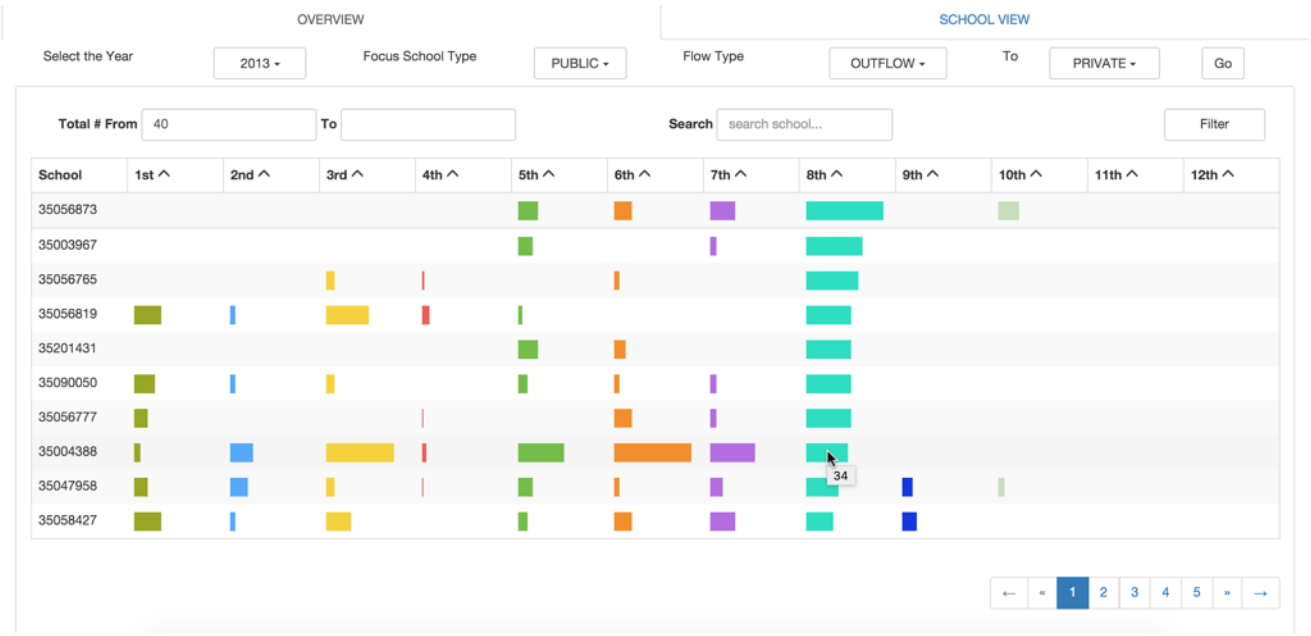

Figure 22 - By hovering the mouse over the bars, users can see the specific number of students leaving the school at that grade.

6 Discussion

Overall, our prototype meets most of the user's tasks and requirements listed in section 2.2. We were able to complete the project as it was initially proposed. We had the opportunity to show it to the team who is developing a related product at the company for which one of us is working and the feedback was positive. As they are still entering the conceptual design and mockup process, they found it very valuable to have our prototype as a reference, and might make use of some of our solutions in the migration section of their product.

The "School View" panel works as expected and serves well to the tasks we identified as the most relevant to schools: identify grades in which migration is more intense, identify to which schools are their students going to/coming from and observe the geographic location of the main competitor schools. The linked faceted view helps user to understand several aspects of the data regarding a specific school migration quickly. The overview panel clearly shows rankings, something users explicitly wanted, and employed a navigation logic familiar to users, important in this case as they are a mostly non computer savvy group.

However, from the information visualization discipline point of view, we didn't fully explore the potential this dataset had regarding its overview. We show the data in small slices, and we don't provide the ideal conditions for uses to discover patterns. When designing the visualizations, we perceived we had a trade-off among allowing comparison among grades and allowing comparison among years, and we prioritized the former as we only had 3 years of data and users tasks had put more emphasis on comparing grades. Although the users requirements we had access to did not mention the need to have a big picture of the whole set of schools, neither emphasize the need for identifying patterns of migration among years, we believe that this has the potential to show relevant information for them. Also, aggregating results per municipality, for example, for sure would be useful for the government analyst, as at least part of resources administration for public education happens in the municipality level in Brazil. If we had more time to work on this project, we would have tried one of the approaches depicted in the next section "Future Work".

7 Lessons learned

This project was a rich experience full of learning opportunities. Herein we list the main lessons we learned.

7.1 Test and get feedback before coding

It is hard to predict all needed interactions in conceptual design and when drawing the mockups, and it is also hard to predict how well the visualization is going to meet users needs. So getting early feedback from users or specialists and making use of discount evaluation techniques such as a cognitive walkthrough with a low-fi prototype could have helped us to foresee issues early in the process. In this project, for example, as we had a tight schedule we started to code the overview panel as early as we could. By the time we got the first feedback on it, a fair amount of coding effort had been done already. So even we were aware that it could be improved, we made the decision to finish it according to the first plan, as we were already very committed in the coding process to make any substantial changes. That meant less available time for investing in one of the alternative options we show in the future work section.

7.2 Data cleaning and wrangling

Most of the challenges associated with this project were related to the scale of the data. Each educational census files (one per year) were roughly 1.5GB large. Cleaning data of such scale was time consuming, and it was even slower because we don't have much experience with big data processing and we had to learn the right tools to deal with it. Most of the fields were not relevant for our purposes (such as the ones related to school's infrastructure). Also, the dataset had some particularities we had to figure out during the process, for example it contained two different coding systems for grades, one for the new educational system, implemented in 2011, and another one for the old educational system. We just found this out when the visual results that were coming out were looking weird. After searching for an eventual bug for a long time, we found out that that was happening because of the data. Then, we derived a new field called "new_grade" to unify the education grade of these two education systems.

7.3 Processing time

When we finished coding, we started to test it using the whole data set. However, for a data that large, the processing times were very long. After wasting some time proceeding this way, we learned that it is better to use a small sample of data for looking for bugs, and only them connect the visualization to the whole dataset.

7.4 Front-end coding

Because we didn't find any open source libraries relevant to our project, we had to do all the coding by ourselves. The front-end design may look simple, but it took a long time to write the HTML and jQuery code. After the first version of the overview panel was done, we found bugs such as: the sorting icon failing to change correctly, the text value cannot be show in the dropdown menu correctly, clicking on the pages in the pagination returning wrong result, etc. We had to solve such small problems one by one. So we learned that we have to allocate a fair amount of time for front-end coding.

8 Future Work

The "Overview" panel could be improved by introducing a filter by municipality, or by showing totals per municipality. This would offer an intermediate level of aggregation of the data in between specific schools and all schools of the state, useful as at least part of resources administration for public education happens in the municipality level in Brazil. Also, more elaborated overview visualizations can be developed to explore further the potential of this dataset, for example, support comparison of years at the overview level, and show clusters of schools among which migration is happening. Herein we suggest two possible solutions to explore, which can be used in combination with the "Overview" panel developed in this project.

The Flowstrates [1] view showed at Figure 1 in the section 3 (related work) could be used to show students migration among years. The origin schools could be at the left side of the view, and the destination schools could be at the right side of the view. The middle column could be segmented by year; if there was migration among those schools in that year, the cell is colored according to an intensity scale based on luminance. The list of schools could be ordered in ascending or descending order to show rankings. The map on the left and the right extremes of the view could show the geographic location of the schools. Users could choose to see total migration flow among schools or the migration among schools for a specific grade (figure 23). Although in this visualization we would also be able to just show a small slice of the dataset at once and use pagination or scrolling to allow the user to explore the rest, it would be easier to the user to identify patterns of migration among the years. He would also be able to see geographic location of schools. However, this visualization is not ideal for comparing migration among grades, so it would be harder to the analyst to answer the question: in which grade is migration more intense? For that reason, we indicate this visualization to be used in combination to the overview panel we developed.

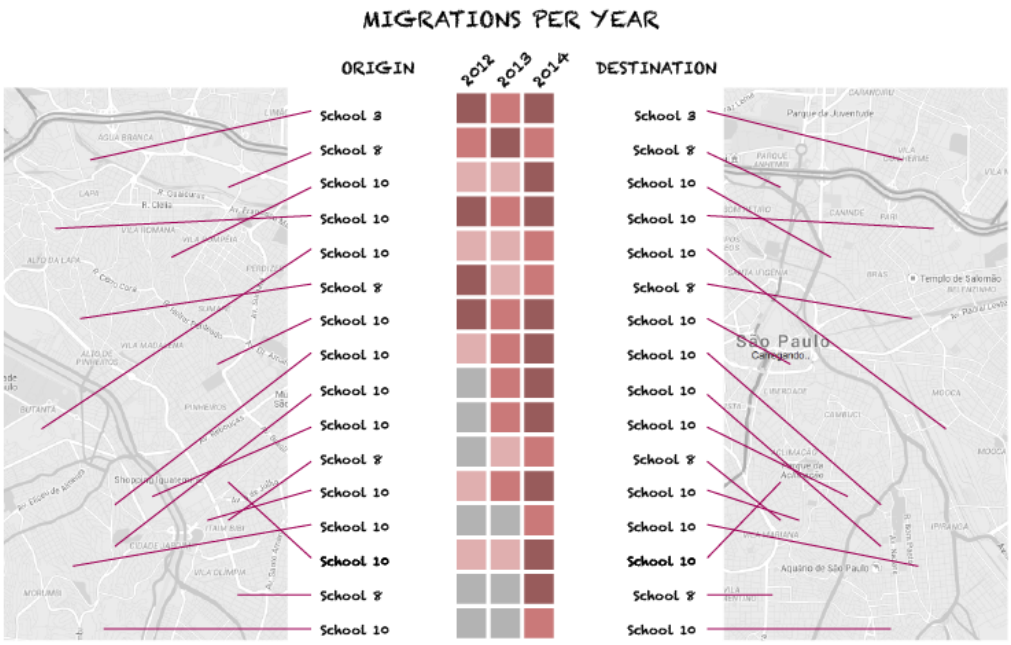

Figure 23 - Sketch of what a “Migration among years” panel would look like.

Another visualization that might be useful to provide a broad overview of the data is a node-link graph such as the ones used in social network analysis. Figure 24 provides an example. Each node would be a school, and links would represent if there was any migration happening among these schools. We could make separate views for inflow, outflow and balance. Balance could be encoded with a diverging colour scale, from red to negative values to blue for positive values. The total number of students migrating, or the balance, could be encoded using the size of the nodes. This view would be probably useful to identify clusters of schools among which migration is more significant.

Probably some kind of filtering by number of students migrating, grades or municipality would be needed, as the total visualization would be too cluttered considering we have 20.000 schools. Also, a zooming strategy like a magnifying glass would be useful to help the user to explore the visualization in order to select a school or a group of schools to know more about. Drawbacks of this visualization are the fact that there is no way to compare migration among years or grades, and it is also does not show a ranking. Maybe it can be used in combination to the “Flowstrates” view described above, or the “Overview” panel we developed in this project.

Figure 24 - Example of node-link graph that could be used to show a big picture of the data [24].

9 Conclusions

Understanding student's migration is useful for both public and private schools as it can help them to identify and understand eventual losses of students. We defined a set of tasks each stakeholder would be interested in doing with this dataset. We proposed an "Overview" panel with migration ranking of schools per grade, and a "School View" panel containing a dashboard showing migration details for a specific school. Through a walk through a scenario, we demonstrate that the prototype has the potential to serve to all tasks we defined. However, we didn't fully explore the potential this dataset had regarding its overview. We show the data in small slices, and we don't provide the ideal conditions for uses to discover patterns. Although the user's requirements we had access to did not mention the need to have a big picture of the whole set of schools, neither emphasize the need for identifying patterns of migration among years, we believe that this has the potential to show relevant information for them. Alternative visualizations such as a node-link graph are proposed for future studies in order to fulfil this gap.

Acknowledgments

The authors wish to thank Tamara Munzner for her valuable feedback and suggestions.

References

[1] Boyandin, Ilya et al. "Flowstrates: An Approach For Visual Exploration of Temporal Origin-Destination Data." *Computer Graphics Forum* 30.3 (2011): 971–980. Web.

[2] Buchin, K., B. Speckmann and K. Verbeek. "Flow Map Layout Via Spiral Trees." *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011): 2536–2544. Web.

[3] *Chicago Area Transportation Study: Final Report*. Chicago: CATS, 1959. Print.

[4] Cui, Weiwei et al. "Geometry-Based Edge Clustering For Graph Visualization." *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008): 1277–1284. Web.

[5] Estevan, F. "Public Education Expenditures and Private School Enrollment." *Canadian Journal of Economics/Revue canadienne d'économique* (2015): 1540-5982. Web.

[6] Fua, Y., M. Ward, and E. Rundensteiner. "Hierarchical Parallel Coordinates for Exploration of Large Datasets." *Proceedings Visualization '99 (Cat. No.99CB37067)* (1999): 43-50. Web.

[7] Gilbert, M. et al. "Cattle Movements and Bovine Tuberculosis in Great Britain." *Nature* 435.7041 (2005): 491–496. Web.

[8] Guo, D. "Flow Mapping And Multivariate Visualization of Large Spatial Interaction Data." *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009): 1041–1048. Web.

[9] Holten, D. "Hierarchical Edge Bundles: Visualization Of Adjacency Relations in Hierarchical Data." *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006): 741–748. Web.

[10] Holten, D., and Jarke J. Van Wijk. "Force-Directed Edge Bundling For Graph Visualization." *Computer Graphics Forum* 28.3 (2009): 983–990. Web.

[11] Kosara, R., F. Bendix and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics* 12.4 (2006): 558–568. Web.

[12] Luo, Sheng-Jie et al. "Ambiguity-Free Edge-Bundling For Interactive Graph Visualization." *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 18.5 (2012): 810–821. Web.

[13] Munzner, T. *Visualization Analysis and Design*. (2014). Print.

[14] Novotny, M., and H. Hauser. "Outlier-Preserving Focus Context Visualization In Parallel Coordinates." *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006): 893–900. Web.

[15] Ong, H. and H. Lee. "Software Report: Winviz—A Visual Data Analysis Tool." *Computers & Graphics* 20.1 (1996): 83–84. Web.

[16] Phan, Doantam et al. "Flow Map Layout." *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (2005): 29. Web.

[17] Rae, A. "From Spatial Interaction Data to Spatial Interaction Information? Geovisualisation and Spatial Structures of Migration from the 2001 UK Census." *Computers, Environment and Urban Systems* 33.3 (2009): 161–178. Web.

[18] Ravenstein, E. G. "The Laws Of Migration." *Journal of the Statistical Society of London* 48.2 (1885): 167. Web.

[19] Tobler, W. Experiments in Migration Mapping by Computer. *The American Cartographer* 14.2 (1987). Web.

[20] Tobler, W. Movement Mapping. Available at http://www.csiss.org/clearinghouse/FlowMapper/. 2004. Web.

[21] Tufte, E. The Visual Display of Quantitative Information. Graphics Press. Chesire, Conneticut. (2001). Print.

[22] Vandenberghe, V., and S. Robin. "Evaluating The Effectiveness of Private Education across Countries: a Comparison of Methods." *Labour Economics* 11.4 (2004): 487–506. Web.

[23] Ward, M. "XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data." *Proceedings Visualization '94* (1994): 326-333. Web.

[24] Wikimedia Commons. By Martin Grandjean [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)]. Available at https://commons.wikimedia.org/wiki/File%3ASocial_Network_Analysis_Visualization.png. Web.

[25] Wood, J., J. Dykes, and A. Slingsby. "Visualisation Of Origins, Destinations and Flows with OD Maps." *The Cartographic Journal Cartogr. J.* 47.2 (2010): 117–129. Web.