University of British Columbia
CPSC 547 - Information Visualization
Tamara Munzner

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

Status Report

Carolina Roman Amigo
carolamigo@gmail.com

Wenqiang Dong
wdong@cs.ubc.ca

23rd November 2015

1. Progress to date

We are succeeding in following the schedule proposed in Table 1. It has not been easy to keep up with the planning, as the project revealed itself more complex than we had foreseen. Herein we describe our progress to date.

Table 1 - Proposed schedule

Date	Milestone
11/9	Finish the proposal
11/13	Finish the front end design for the Overview interface
11/18	Finish the back end coding for the Overview interface Test Overview interface
11/22	Finish the front end design for the School View interface
11/23	Finish the Status Updates report

1. Data processing

The raw data, downloaded from the Brazilian Education Department Website, was in *.txt format and was formatted to be read in SAS or SPSS. Fields had a fixed number of characters and were separated by spaces. So we had to convert the original data to *.csv format and import it to a MySQL database. The tables we have in the database are listed in Table 2.

Table 2 - Description of tables contained in our database.

Table	Field	Description
school	school_id	primary key, foreign key, id for each school.
	school_name	the name of the school
	school_city	city code of the school
	school_district	district code of the school
	school_status	status of school (active, inactive)

	school_type	type of the school, public (federal, state, city) or private
	post_code	postal code of the school
	latitude	geographical coordinates of the school
	longitude	
student	year	enrollment year of a student, from 2012 to 2014
	enrollment_id	primary key, id for each enrollment
	student_id	primary key, id for each student
	education_grade	student’s educational grade
	school_id	id for each school
	new_grade	derived grade for the student

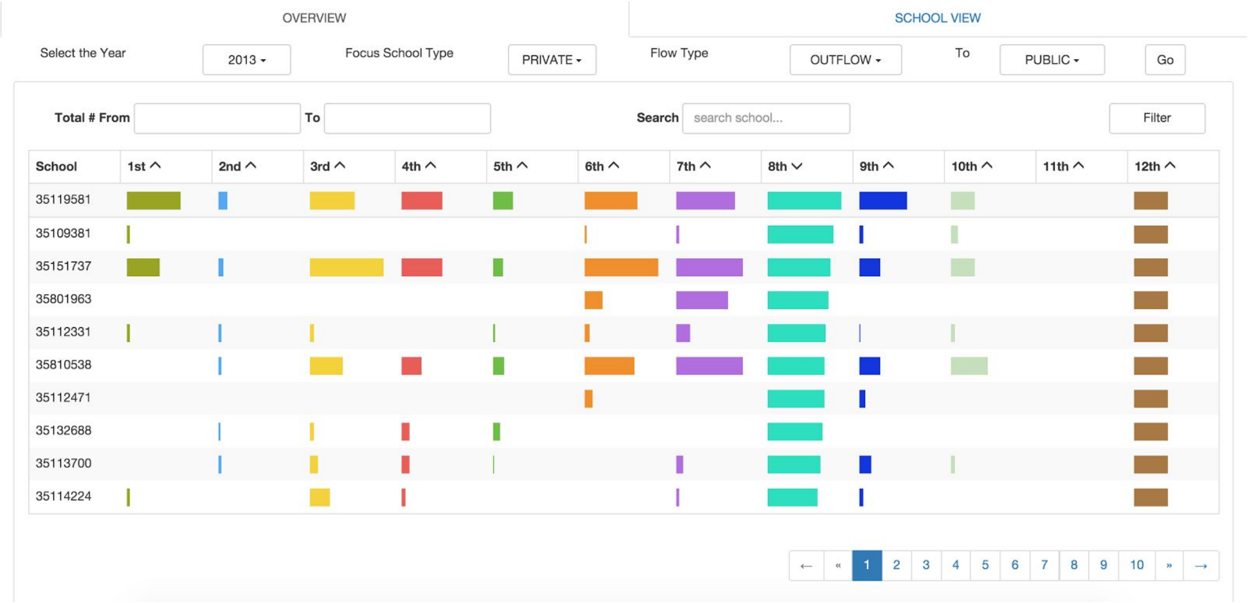
The “student” table is decomposed into three tables: “student2012”, “student2013” and “student2014”. We can calculate the inflow and outflow of students grouped by schools directly from these tables, which are used in our overview panel.

We only kept on the database students in elementary or secondary schools in the State of São Paulo, excluding the ones enrolled in professional degrees and late adult education. According to the data dictionary, we find that Brazil recently implemented a 9 years education system instead of the 8 years education system they had before. The tables reflect this change, containing two different coding systems, each one referring to one education system (the old and the new). A new column called “new_grade” was created in order to unify the education grades for the two different education systems.

After data cleaning, we have 34,876 schools in the “school” table, 7,690,966 students in “student2012” table, 7,565,158 students in “student2013” table and 7,465,901 students in “student2014” table.

2. Front-end and Back-end implementation of the “Overview” panel

Using HTML and JavaScript, we have completed the front-end implementation for the overview panel. The panel was written all by ourselves because we could not find any open source libraries that could serve for our purposes. For the back-end, we used Flask, a lightweight Python web framework based on Werkzeug and Jinja2, to implement the system. We have completed the communication of data between front-end and back-end for the overview panel, which we consider to be almost fully developed. Herein we show examples of how the panel is working right now.



Select the Year 2013 ▾ Focus School Type PRIVATE ▾ Flow Type OUTFLOW ▾ To PUBLIC ▾ Go

Total # From ___ To ___ Search search school... Filter

School	1st ^	2nd ^	3rd ^	4th ^	5th ^	6th ^	7th ^	8th ∨	9th ^	10th ^	11th ^	12th ^
35119581	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬		▬
35109381	▬					▬	▬	▬	▬	▬		▬
35151737	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬		▬
35801963						▬	▬	▬				▬
35112331	▬	▬	▬		▬	▬	▬	▬	▬	▬		▬
35810538		▬	▬	▬	▬	▬	▬	▬	▬	▬		▬
35112471						▬		▬	▬			▬
35132688		▬	▬	▬	▬			▬				▬
35113700		▬	▬	▬	▬		▬	▬	▬	▬		▬
35114224	▬		▬	▬			▬	▬	▬			▬

← « 1 2 3 4 5 6 7 8 9 10 » →

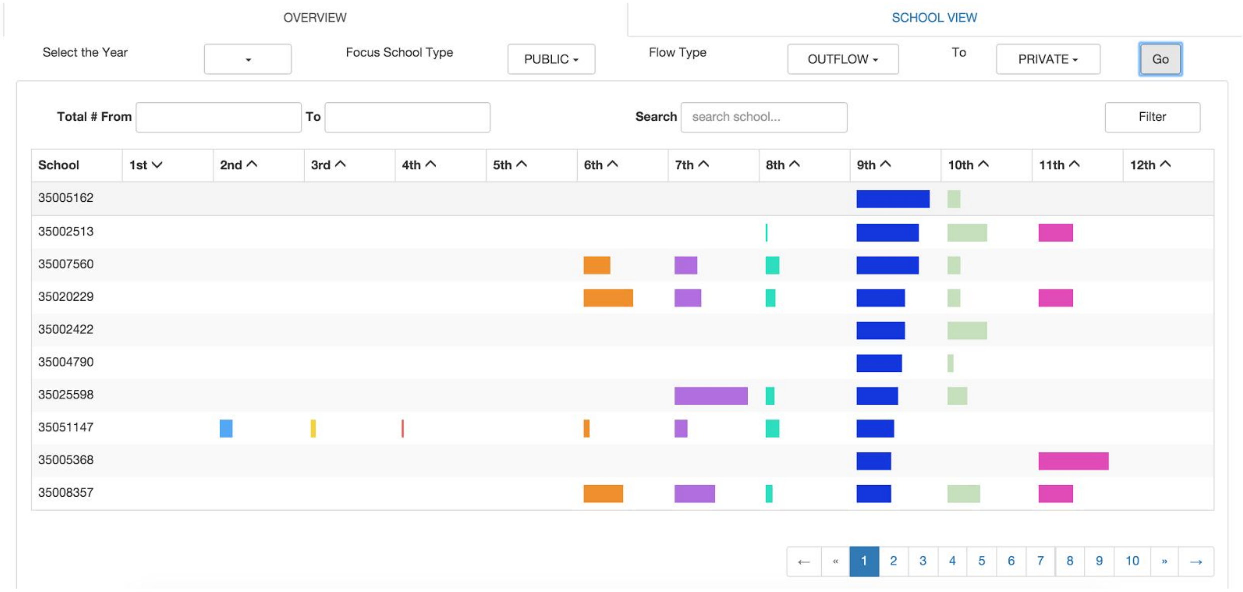
Figure 1 - First screen of the “Overview Panel”

Users first specify the year, then focus school type (public, private or both), then flow type (inflow, outflow or both) and finally destination school type (public, private or both). The inflow and outflow of the students migration is shown in the table (Figure 1). Because School’s names are very long and in Portuguese, we opted for showing just the School ID in the table as default. Users can see the name of the school by hovering over the School ID.

The number of inflow and outflow of students is encoded by the length of the bar. Each bar is aligned to the left side of the cell. Users can see the precise number of inflow or outflow by hovering over the color bar. Users can also filter the data by specifying the total number of the students flow or specifying the name of the school. To identify which schools are losing more students and which schools are gaining more students per grade, the data on the columns can be sorted by crescent or decrescent order. Because the data is really big, we used pagination in order to only load a small set of results at once.

By default, the year is set to “2013”. The focused school type is set to “both”. The flow type is set to “balance”. The destination school type is set to “both”.

Figure 2 shows the result of outflow from public schools to private schools. By default, we order the data by the 9th grade descendingly because it is the final year the students are in the elementary school. Most students will change schools in this year in order to be admitted to a better secondary school. From this figure, we can see that school 35005162 (PROFESSOR ALBERTO CONTE) loses the most students (21) in the 9th grade, which reminds the government that something should be done to improve the teaching quality of this school. We can reorder the same data by any grade (e.g 6th grade descendingly), as shown in Figure 3. Users can also click on “next page” to see more schools (Figure 4).

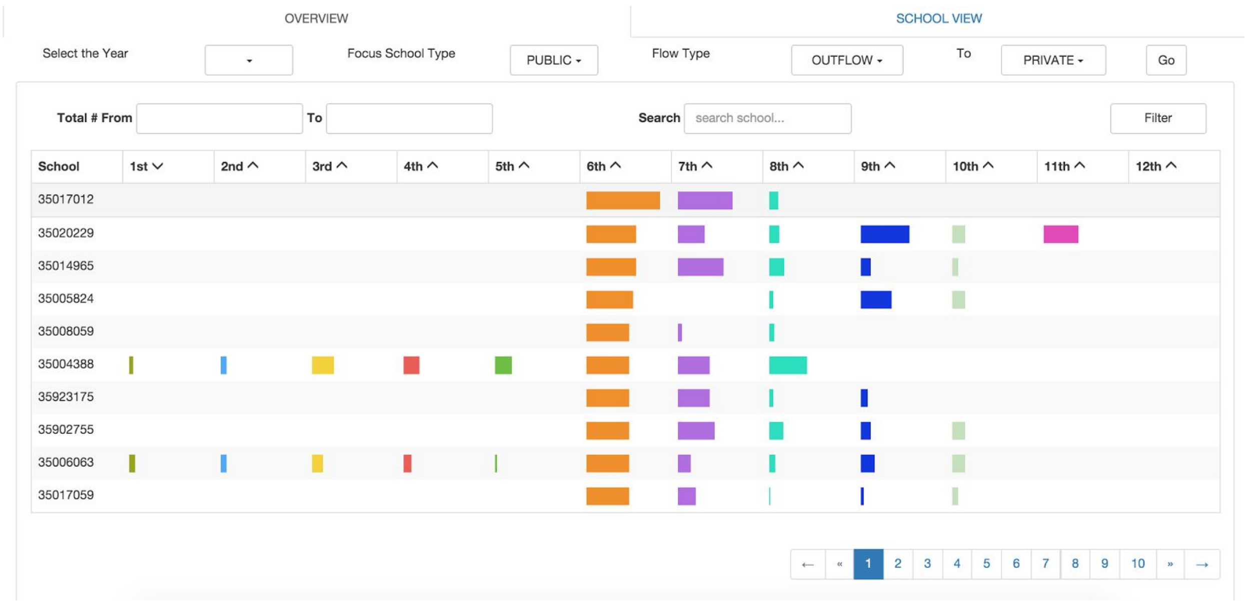


Select the Year ▾ Focus School Type PUBLIC ▾ Flow Type OUTFLOW ▾ To PRIVATE ▾

Total # From To Search search school...

School	1st ∨	2nd ∧	3rd ∧	4th ∧	5th ∧	6th ∧	7th ∧	8th ∧	9th ∧	10th ∧	11th ∧	12th ∧
35005162									■	■		
35002513								■	■	■	■	
35007560						■	■	■	■	■		
35020229						■	■	■	■	■	■	
35002422									■	■		
35004790									■	■		
35025598							■	■	■	■		
35051147		■	■	■		■	■	■	■			
35005368									■		■	
35008357						■	■	■	■	■	■	

Figure 2 - Outflow of students from public schools to private schools at the 9th grade.



OVERVIEW

SCHOOL VIEW

Select the Year ▾

Focus School Type PUBLIC ▾

Flow Type OUTFLOW ▾

To PRIVATE ▾

Go

Total # From ___ To ___

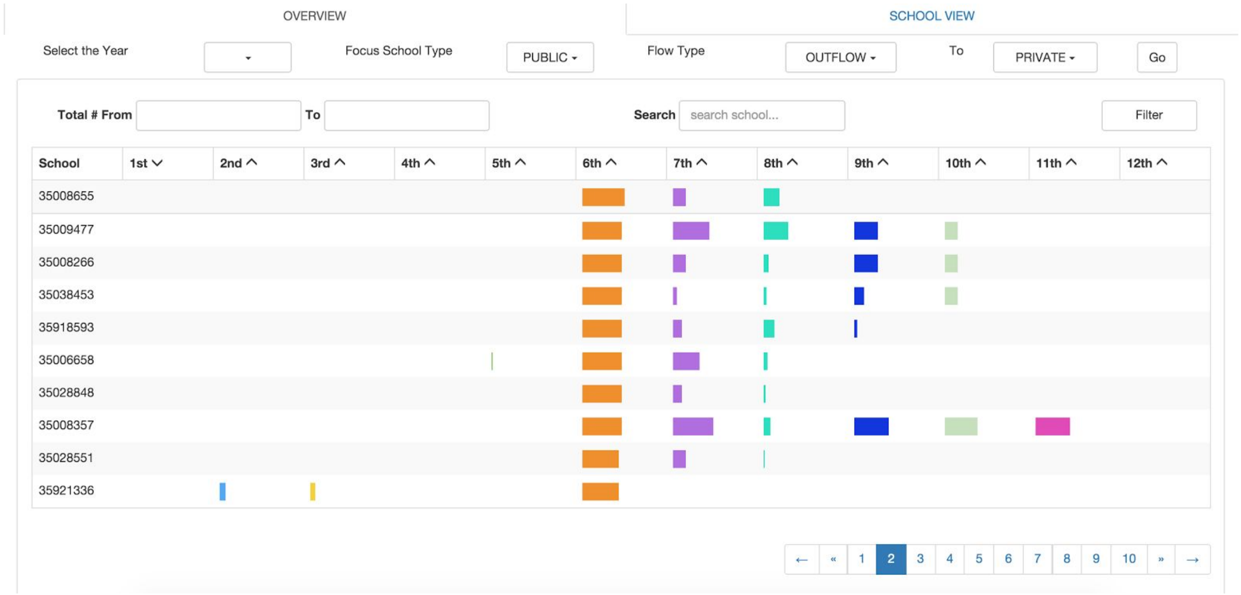
Search search school...

Filter

School	1st ∨	2nd ∧	3rd ∧	4th ∧	5th ∧	6th ∧	7th ∧	8th ∧	9th ∧	10th ∧	11th ∧	12th ∧
35017012						▬	▬	▬				
35020229						▬	▬	▬	▬	▬	▬	
35014965						▬	▬	▬	▬	▬		
35005824						▬		▬	▬	▬		
35008059						▬	▬	▬				
35004388	▬	▬	▬	▬	▬	▬	▬	▬				
35923175						▬	▬	▬	▬			
35902755						▬	▬	▬	▬	▬		
35006063	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬		
35017059						▬	▬	▬	▬	▬		

← « 1 2 3 4 5 6 7 8 9 10 » →

Figure 3 - Users can reorder the data simply clicking on another grade. This screen shows the outflow of students from public schools to private schools at the 6th grade.



OVERVIEW

SCHOOL VIEW

Select the Year ▾

Focus School Type PUBLIC ▾

Flow Type OUTFLOW ▾

To PRIVATE ▾

Go

Total # From ___ To ___

Search search school...

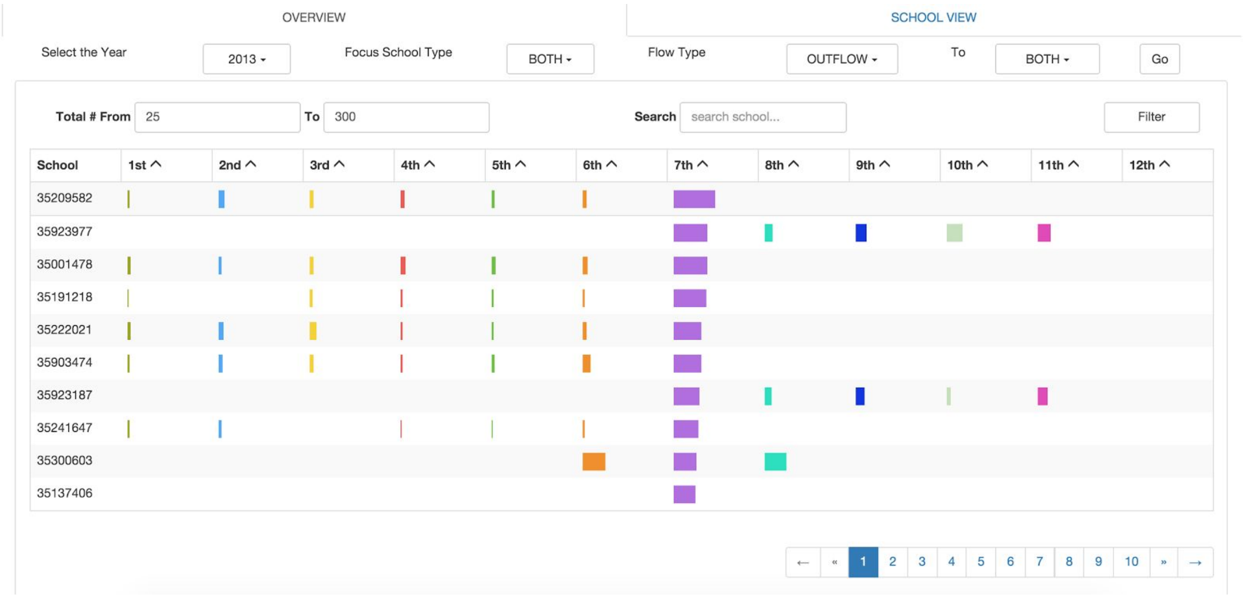
Filter

School	1st ∨	2nd ∧	3rd ∧	4th ∧	5th ∧	6th ∧	7th ∧	8th ∧	9th ∧	10th ∧	11th ∧	12th ∧
35008655						▬	▬	▬				
35009477						▬	▬	▬	▬	▬		
35008266						▬	▬	▬	▬	▬		
35038453						▬	▬	▬	▬	▬		
35918593						▬	▬	▬	▬			
35006658					▬	▬	▬	▬				
35028848						▬	▬	▬				
35008357						▬	▬	▬	▬	▬	▬	
35028551						▬	▬	▬				
35921336		▬	▬			▬						

← « 1 2 3 4 5 6 7 8 9 10 » →

Figure 4 - Users can see more schools clicking on “next page”.

Figure 5 shows the outflow result for students migrating from both (public, private) school types to both school types filtered by an outflow number between 25 and 300 and ordered by the 7th grade descendingly.



Select the Year: 2013

Focus School Type: BOTH

Flow Type: OUTFLOW

To: BOTH

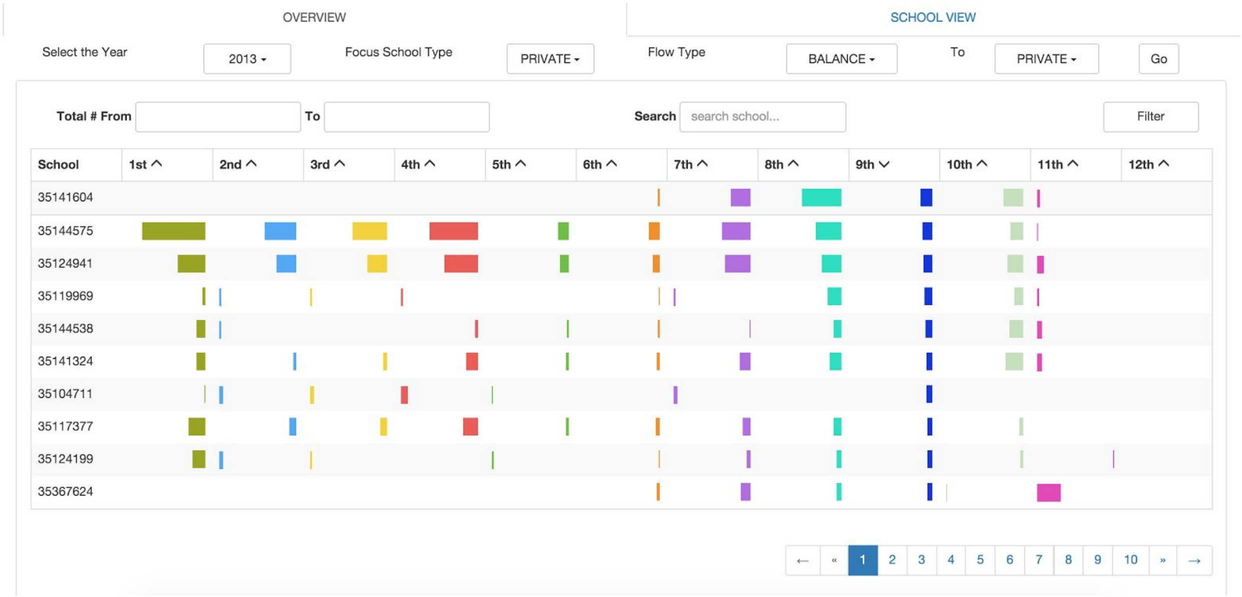
Total # From: 25 To: 300

Search: search school...

School	1st ^	2nd ^	3rd ^	4th ^	5th ^	6th ^	7th ^	8th ^	9th ^	10th ^	11th ^	12th ^
35209582												
35923977												
35001478												
35191218												
35222021												
35903474												
35923187												
35241647												
35300603												
35137406												

Figure 5 - Students migrating from both (public, private) school types to both school types filtered by an outflow number between 25 and 300 and ordered by the 7th grade descendingly.

For the “balance” data (inflow # - outflow #), there may be negative values, which means the school is losing more students than gaining students. In this case, we cannot encode the flow number by the length of the bar anymore. To visualize such data, we draw the bars from right to left. Focusing on the private schools, the “balance” result for these schools is shown in Figure 6.



Select the Year: 2013

Focus School Type: PRIVATE

Flow Type: BALANCE

To: PRIVATE

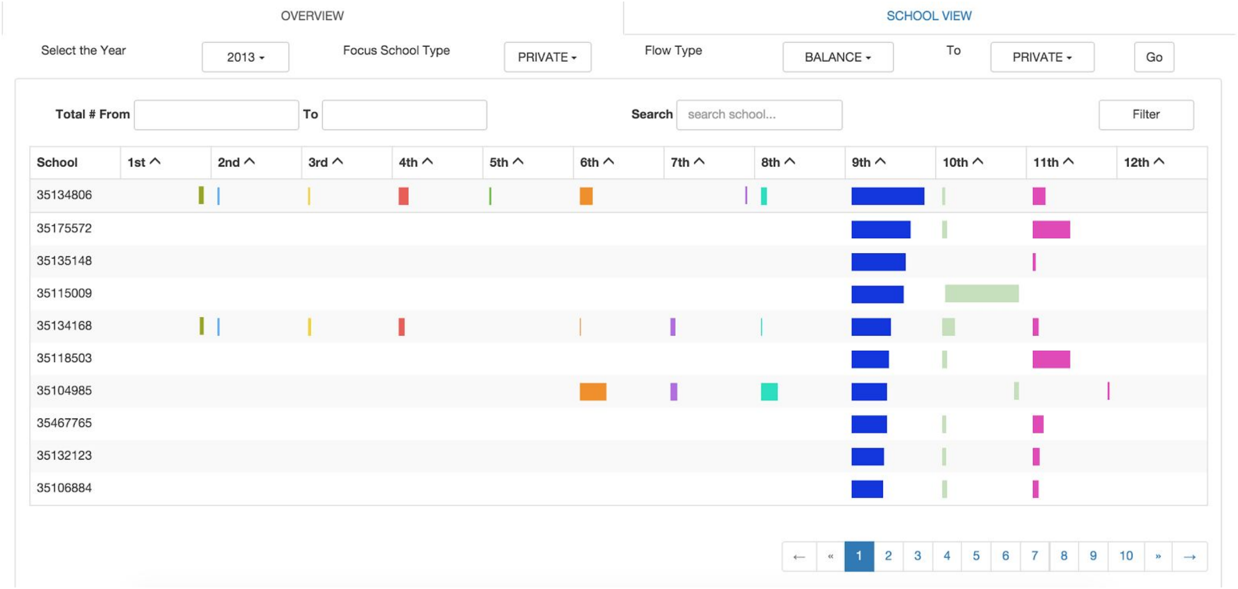
Total # From: To:

Search: search school...

School	1st ^	2nd ^	3rd ^	4th ^	5th ^	6th ^	7th ^	8th ^	9th ∨	10th ^	11th ^	12th ^
35141604												
35144575												
35124941												
35119969												
35144538												
35141324												
35104711												
35117377												
35124199												
35367624												

Figure 6 - Migration balance; negative values are shown from right to left.

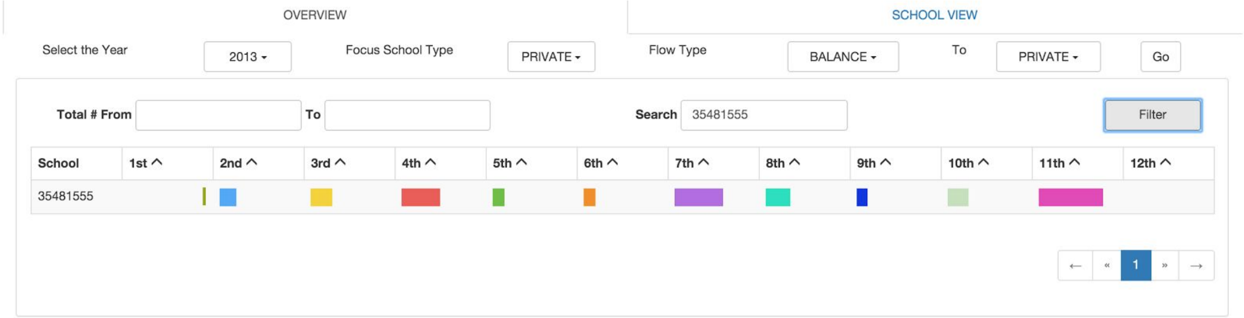
The data shown in the table is ordered by the 9th grade ascendingly. By hovering over the first cell of the 9th grade, we can see that school "35141604" (RIO CLARO SISTEMA DE ENSINO UNIDADE II) loses 62 more students than it is gaining. When ordering the 9th grade descendingly, we can find that the school "35134806" (ETAPA COLEGIO DE EFM) is gaining 411 more students than losing, what makes it the winner school. Figure 7 shows the results.



School	1st ^	2nd ^	3rd ^	4th ^	5th ^	6th ^	7th ^	8th ^	9th ^	10th ^	11th ^	12th ^
35134806												
35175572												
35135148												
35115009												
35134168												
35118503												
35104985												
35467765												
35132123												
35106884												

Figure 7 - Ordering the balance overview by 9th grade descendingly.

In addition, users can select a particular school to inspect, taking school "35481555" as example, the balance data is shown as in Figure 8. Users can either use the school id or school name to search the school.



School	1st ^	2nd ^	3rd ^	4th ^	5th ^	6th ^	7th ^	8th ^	9th ^	10th ^	11th ^	12th ^
35481555												

Figure 8 - Selecting a particular School to inspect.

2. Challenges

1. Processing time

All the challenges associated with this project are related to the scale of the data. The 2012 students file occupies 683.5MB. The 2013 students file occupies 675.8MB. These two files were already partially cleaned by Carolina before she shared them with Dylan, what means they were even bigger. The 2014 students file, which was taken raw as it was downloaded more recently for this project, is more than 1.5GB large. Cleaning data of such scale is time consuming, and we don't have much experience dealing with big data yet. When we finished coding, we were testing it using the whole data set, what was taking a long time. This caused us to waste lots of time for correcting bugs. We should pilot on a small sample data set first. This is a very important lesson we learned.

2. Extracting relevant fields

At first, we were not sure which fields were going to be useful for the project. We had to take time to add new columns to the table because some fields which we thought were not relevant turned out to be useful at last. We should make a very good plan before we extract relevant fields from the big data set.

3. Deriving data

We didn't know the dataset contained a different coding system for each education systems in Brazil. So when the first visual result came out, it looked very 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.

4. Front-end coding

Because we didn't find any open source libraries which is 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 lots of bugs, e.g, 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. In addition, when we conceived the mockup we didn't consider all the possibilities the users would have when using the overview panel. These problems pushed us to refine our panel. We learned lots of things in this process.

5. Portuguese

It is really difficult for non-Portuguese speaker to deal with the data. There are more than 50 fields in the dataset, whose names are in abbreviation. Even though Carolina provided Dylan with translations for the main fields, Dylan had to use Google Translate to understand any extra information, or turn to Carolina for help. That takes more time than if the dataset was in English. But it is an interesting dataset and Dylan likes Portuguese :)

3. Changes to the previous plan

1. We build our system on Flask instead of Apache. Flask is very lightweight and it reduces lots of work on the back-end. The back-end code now is written in Python.

2. (Changes in the future) As Tamara mentioned in our proposal, we should try to review our overview panel to provide a more thorough overview. We want to reduce the amount of filtering and offer more information. If we still have time after we complete the school view panel, we want to add the function of clustering schools into groups based on similar flow patterns. It is really interesting coming up with new ideas for the project.

4. Previous Work Section Draft

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 analyze 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.

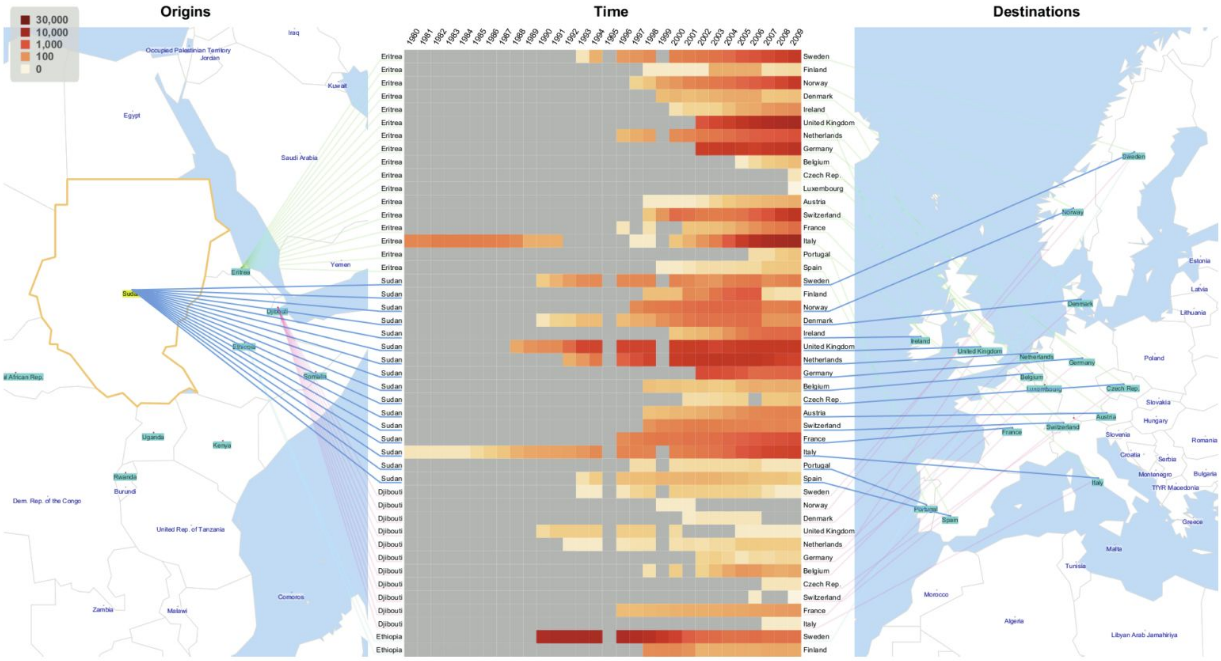


Figure 1 - Flowstrates example showing refugees migration among East Africa and Western Europe from 1980 and 2009 [1].

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 [24]. 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.

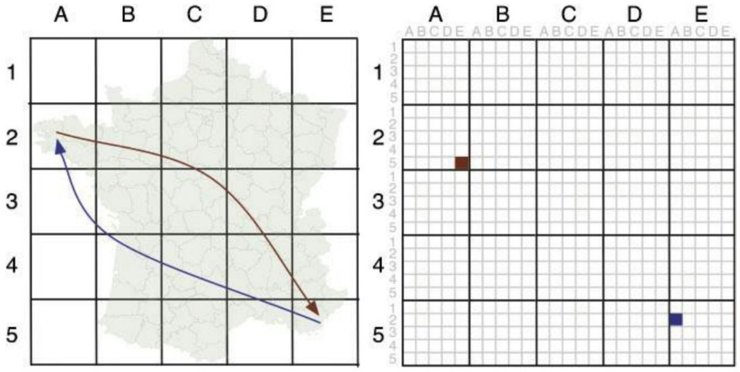


Figure 2 - Geographic space partitioned into a regular grid (left) and a origin-destination map space (right) [24].

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.

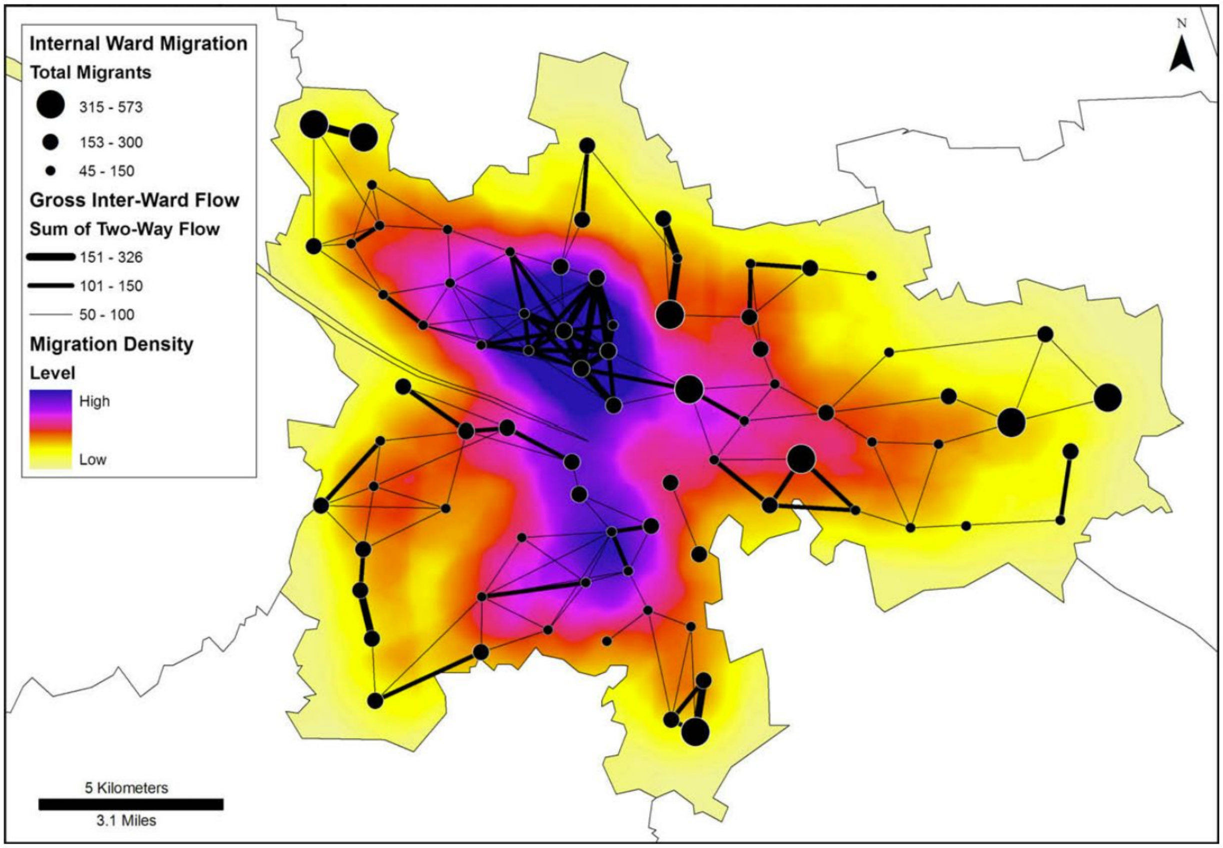


Figure 3 - Intra-city migration in Glasgow [17].

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.

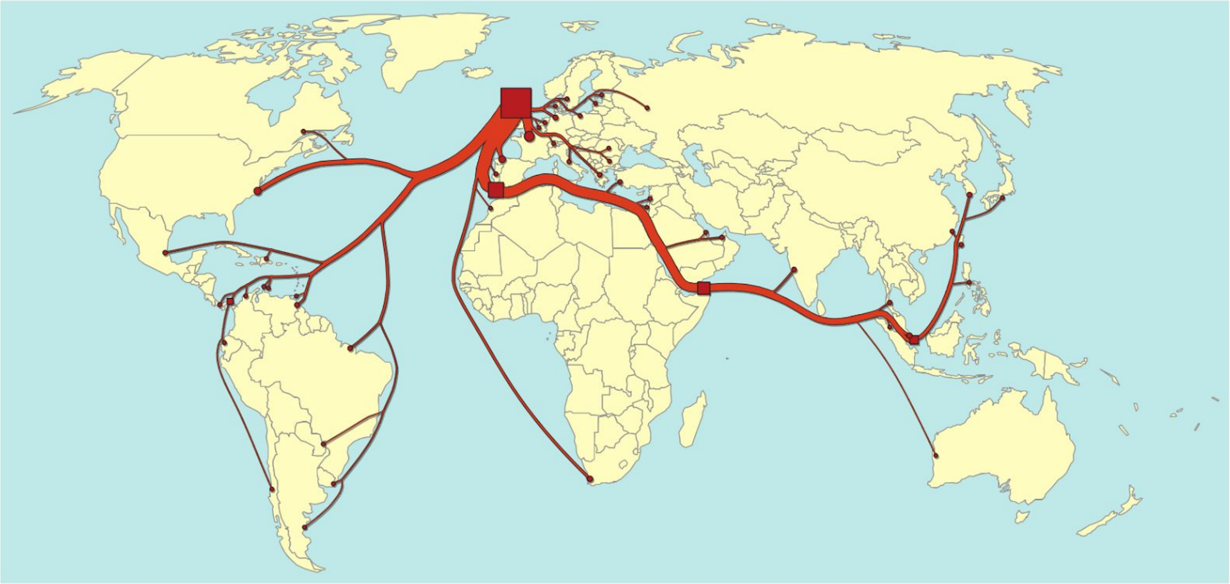


Figure 4 - Top 50 whisky exports from Scotland in 2009 by volume, example of flow mapping using spiral trees [2].

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.

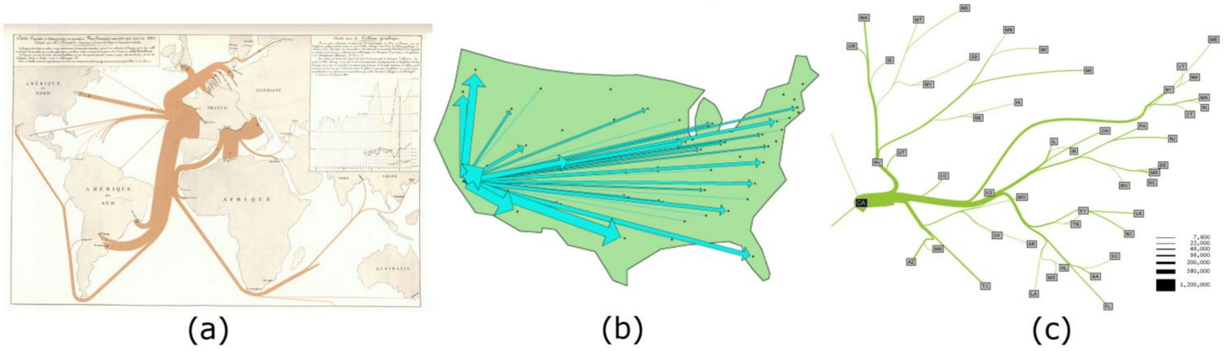
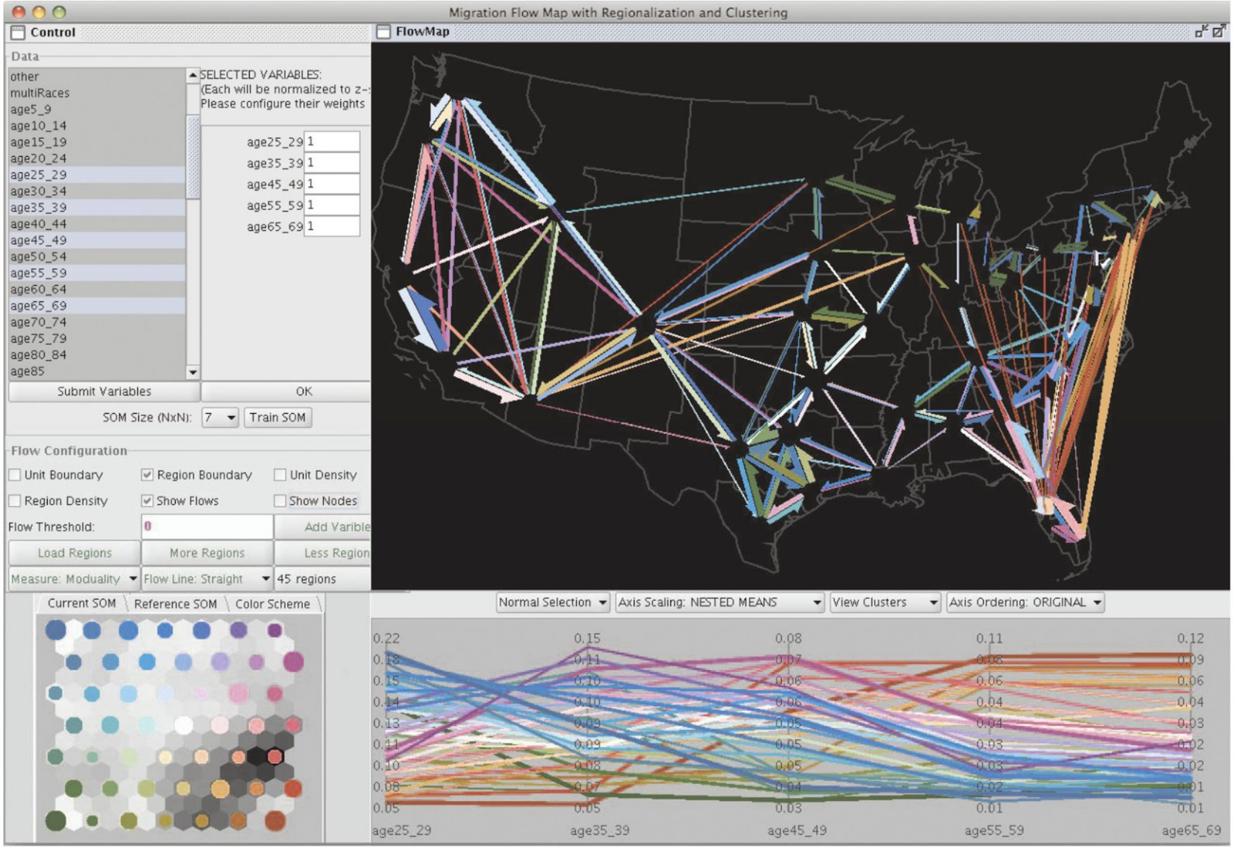


Figure 5 - (a) Minard's 1864 flow map of wine exports from France [21] (b) Tobler's computer generated flow map of migration from California from 1995 - 2000. [19; 20] (c) A flow map proposed by Phan et al. [16] that shows the same migration data.

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.



SELECTED VARIABLES:
(Each will be normalized to z-
Please configure their weights

age25_29 1
age35_39 1
age45_49 1
age55_59 1
age65_69 1

SOM Size (NxN): 7

☐ Unit Boundary ☑ Region Boundary ☐ Unit Density
☐ Region Density ☑ Show Flows ☐ Show Nodes

Flow Threshold: 0

Measure: Moduality
Flow Line: Straight
45 regions

Normal Selection
Axis Scaling: NESTED MEANS
View Clusters
Axis Ordering: ORIGINAL

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.

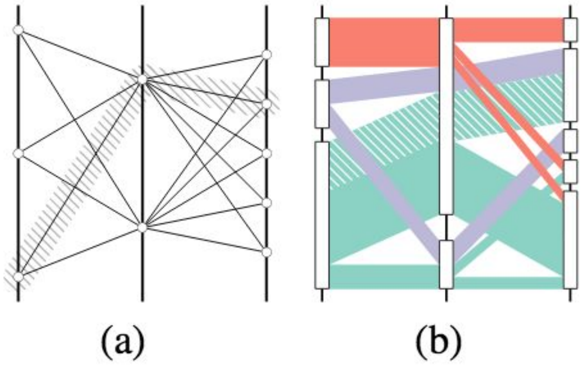


Figure 7 - (a) Parallel coordinates represent categories by points on continuous axes; (b) Parallel sets show the frequencies of categories and relations [11].

8. 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, Fernanda. “Public Education Expenditures and Private School Enrollment.” *Canadian Journal of Economics/Revue canadienne d'économique* (2015): Web.

[6] Fua, Ying-Huey, M.o. Ward, and E.a. 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, Diansheng. “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, Danny, and Jarke J. Van Wijk. “Force-Directed Edge Bundling For Graph Visualization.” *Computer Graphics Forum* 28.3 (2009): 983–990. Web.

[11] Kosara, R., Bendix, F., & Hauser, H. (2006). Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, *12*(4), 558–568. http://doi.org/10.1109/TVCG.2006.76

[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, Tamara. *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, Hwee-Leng, and Hing-Yan 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, Alasdair. “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. American Cartographer, 1987.

[20] Tobler, W. Movement Mapping. http://csiss.ncgia.ucsb.edu/clearinghouse/FlowMapper. 2004.

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

[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.o. “XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data.” *Proceedings Visualization '94* (1994): 326-333. Web.

[24] Wood, Jo, Jason Dykes, and Aidan Slingsby. “Visualisation Of Origins, Destinations and Flows with OD Maps.” *The Cartographic Journal Cartogr. J.* 47.2 (2010): 117–129. Web.