

# Weighted flow diagrams for statistical output

Edwin de Jonge

Statistics Netherlands, e-mail: e.dejonge@cbs.nl

## Abstract

Many official statistics are flow data, e.g. population dynamics, System of National Accounts (SNA) and energy statistics. Figures of these statistics are usually presented on the internet using detailed tables that are hard to digest. An alternative but intuitive visualization for these data are weighted directed flow diagrams. These diagrams show the flow of quantities through a system. For small and nicely distributed data flow diagrams are an attractive means of visualisation. Data of NSI's however often is skewed and detailed which leads to a cluttered flow diagram where important details are obscured. This paper describes a visualisation experiment plotting internal migration in the Netherlands using Google Earth. We discuss some methods and their application that alleviate these problems. The methods include filtering and interaction.

**Keywords:** Visualisation, flow diagram, web presentation

## 1. Introduction.

Visualization of statistics recently has gained interest. Many official statistics institutions have started to use statistical visualisations on their web sites. The rise in the usage of visualisations is partly due to the improved web graphics technology: it is relatively easy to create interactive and animated graphics. Many official statistics increasingly utilize visualisation to show patterns in their data instead of publishing the data only in tabular form.

The visualisations commonly used are line and bar charts, scatter and bubble plots and thematic cartographic maps. Each of these visualisations has its merits but none can be used to visualise flow data effectively. Flow visualisations in official statistics are rare. In section 2 we pose that many official statistics can be seen as flow data and that flow visualisation is an option to be considered. Useful flow visualisations for official statistics are discussed in section 3.

In section 4 we present a visualisation experiment of internal migration, the problems encountered and possible solutions.

We end in section 5 with conclusions and suggestions for further research.

## 2. Flow statistics

Flow statistics are statistics of quantities that flow from one point to another point. Many regional statistics are flow statistics. Migration, transportation, trade and commuting are all flows statistics: people and goods flowing from one region to another. The statistics calculated are properties of these flows. If we change the perspective from flow to a region, then we see that each region has an inflow, internal flow and outflow. The flows form a balance sheet of the region, a region can be in balance if inflow equals the outflow. Typically flow data is published in large tables

or plotting margins. This form of publication does not show the flow patterns available in the data.

Many important official statistics are statistical balance systems. System of National Accounts (SNA), energy statistics and international trade are all balance systems. A statistical balance system is kind of a balance sheet: for example in the SNA import, production, consumption and export should be in balance. In energy statistics energy carriers are imported or produced, converted into other energy carriers, consumed (with loss), stored, or exported (with loss). In these balance systems flows are not (always) regional, but the flows point from one stage to another stage. Balance systems are mostly published in large tables. It can be difficult to get an overall picture of balance data.

The purpose of visualisation of statistics is to utilise the human visual system to detect or suggest patterns and relations in the data. A good and simple example of an effective visualisation is data plotted in a scatter plot: a correlation is much easier to detect in a scatter plot than in a data table. This is also true for trends and distribution of data.

For flow data we can make a similar argument. It is difficult to discern patterns in a table with flow data. Possible flow patterns are main flow through a system and distribution of flow sizes over the vertexes.

A good visualisation of flow data should therefore show the major flows and give an impression on the distribution of flow sizes in origins and destinations.

### **3. Weighted flow diagrams**

There are many types of flow charts, but most are used for modelling process designs. These do not fit our visualisation purpose. The flows we are interested in have two important properties. First of all each flow has a direction: it flows from an origin to a destination. Secondly each flow has a size: it is a flow of a quantity. The flow diagram should be able to model and express these properties.

#### **3.1 Directed weighted graph**

Flow data form a graph, where the regions or stage are vertexes and the flows are the edges. Since each flow has a direction and a size the resulting graph is a directed graph with weighted edges. It should be noted that for many statistical datasets the resulting graph contains cycles.

Graph visualisation has been an active research topic, but has mainly concentrated on efficient layout algorithms for large undirected and directed acyclic graphs (DAG). These visualisations emphasise the vertexes and how they are connected.

Visualisation research for directed weighted graphs with cycles is scarce. An exception is [van Dongen (2000)] which presents a Markov Cluster Algorithm for weighted graphs. It calculates clusters of vertexes from a weighted graph and the resulting graph visualisation is a reduced graph where weights have been altered. This however does not result in a visualisation that has our desired properties.

#### **3.2 Sankey diagrams**

First used in 1898 by Sankey to describe the energy flow of a steam engine, Sankey diagrams [Sankey (1898)] are nowadays mainly used in material sciences to show the material transfers between processes. In Sankey diagrams the arrow widths are

proportional to the flow quantity. A Sankey diagram shows the distribution of flow sizes in a system. Sankey diagram also are able to present different material flows by using different arrow colours. This means that the arrows are typed: each arrow represents one type of material.

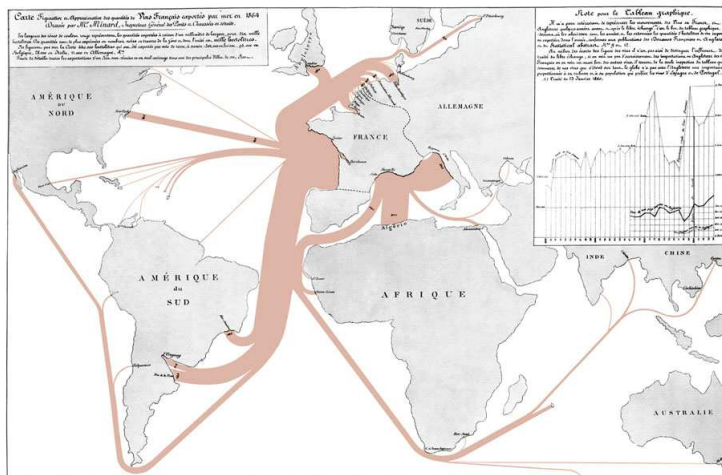
Sankey diagrams are very suited for visualising balance data. They are not very known but are used by energy statistics for a long time. Most of these Sankey diagrams are designed by hand.

When the data to be visualised is large or skewed, naive usage of Sankey diagrams is problematic. Naive application of a Sankey diagram to a large dataset results in a cluttered flow diagram. That is probably the reason that most energy Sankey diagrams restrict themselves to the main energy flows. If the data is skewed, large flows render all other flows nearly invisible. This happens for example with the Dutch energy balance. The energy balance of the Netherlands is dominated by oil import and export because most oil for European consumption is imported via the Netherlands.

A solution to these problems may be found in using the current web graphic technology to develop interactive Sankey diagrams, where users can select and filter data. If the data is hierarchal flow may be broken down into component flows. Another option is to offer different scaling options for arrow widths. For these interactive options a Sankey layout algorithm is needed that intuitively adapts to the changes made by the user. This is a direction for further research.

### 3.3 Cartographic flow maps

Flow mapping is a long existing practice in cartography.



Charles Joseph Minard. *Traité des Cartes Graphiques et Cartes Figuratives de M. Minard, 1845-1869*, a portfolio of his work held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris

**Figure 1** Export of French wine, Minard (1864)

An outstanding example of an early flow map is wine export of France designed by Charles Joseph Minard (1864), who is also the creator of the well known map of the march of Napoleon. Many effective flow maps are designed by hand because manually routing the flows gives a good result. Most flow maps contain at most 100 flows. Design principles are given and discussed in various text cartographic textbooks [Borden (1999), Ormeling (1997)]. Principles underlying such a flow layout are used only recently in computer generated flow layouts. The flow map layout algorithm described in [Doantam (2005)] gives an attractive example of computer

generated unidirectional migration visualization. This is accomplished mainly by clustering flows travelling in the same direction: the resulting visualizations have an artery structure. It should be noted that most literature on flow maps is on unidirectional flow maps. They restrict the flows to one origin or one destination, e.g. Minards flow map shows exports from France but no imports to France. Experiments with computer generated bidirectional flow maps have been conducted by Tobler (1987), where the number of flows was limited to 50.

Cartographic flow maps have an extra property besides direction and size: the vertexes have fixed coordinates. This property creates extra restrictions: the resulting flows are plotted on a map, where each flow should start and end in the correct region. These restrictions make it harder to avoid a cluttered visualization. Firstly small densely populated regions tend to have many incoming and outgoing flows resulting in crowded beginning and ending of flows. Secondly many flow phenomena are non local: flows between non-adjacent regions obscure the regions in between.

## 4. Experiment: Internal migration

As noted above, almost all flow maps have a small number of flows. Ideally a flow map visualisation of a large flow dataset shows relevant patterns in the data. None of the flow maps above deals with a large and skewed dataset. These type of datasets are common in official statistics. To experiment with this kind of datasets and to collect visualisation problems and possible solutions a dataset describing internal migration was selected. In a previous project a large number neighbourhood statistics were visualized and published using Google Earth [ten Bosch(2008), Beeckman (2008)]. Because of familiarity with this tool and its open format (KML) that makes it easy to add generated shapes Google Earth was used for the experiment and a regional dataset was selected.

### 4.1 Experiment

The dataset used is the migration between Dutch municipalities in 2006. This data is freely available from the statistical database of Statistics Netherlands, [StatLine]. This particular dataset has some interesting properties that are common in statistical flow data sets: the number of flows is large (however a fraction of the potential number of flows) and the distribution of flow size is very skewed.

Furthermore this dataset contains non-local flows, i.e. people migrating over a long distance. For many migration flows there is also a return flow, so most flows are bidirectional.

Migration can be modelled as a matrix  $M$  with elements  $m_{ij}$  that describes the migration from region  $i$  to region  $j$ . In 2006 there were 459 Dutch municipalities. This means that there are potentially  $210,222^1$  flows between the municipalities. Many of these flows are nonexistent. In 2006 the number of migration flows between municipalities was 60,073 (28.5% of the potential flows).

De maximum flow size in the flow data is 2,888, the median is 2 and the mean 10.86. So there are some huge migration flows, but the majority of migration flows has a

---

<sup>1</sup> The number of potential flows is  $n(n-1)$

very limited size. This is one of the aspects that make it difficult to find a comprehensive visualization of these migration statistics.

A successful visualization of this data set should have the following properties:

- It gives an impression of the overall migration flow and volume
- It shows clearly the important migration flows
- It reveals patterns in the flow data

The properties of the data make it hard to meet the goals. The number of flows is that large that the naïve implementation where every flow is drawn gives a cluttered result which is difficult to interpret. This problem is aggravated because (almost) every flow  $m_{ij}$  has a reverse flow  $m_{ji}$ . We address these problems with visual encoding and user interaction.

Before we can choose a proper visual encoding scheme and a successful user interaction mechanism, we first define an importance criterion of migration flows. The obvious criterion is that a migration size is large compared to all other migrations sizes. We use this criterion in a visual encoding for flow size and in data filtering.



**Figure 2 Internal migration 2006, naive implementation with 6000 flows**

## 4.2 Visual encoding

Migration flow size is a very important property to be encoded in the visualization. It is a measure for how important the migration flow is compared to the other flows and is therefore responsible for fulfilling at least two of the visualization goals. It gives an impression of the migration volume and it emphasizes large flows. An important issue is that the distribution of flow sizes is skewed; the encoding should deal with that. Because of the large number of flows multiple visual tricks are needed to encode the importance of the flows. We discuss a couple of them. Flow size is typically encoded

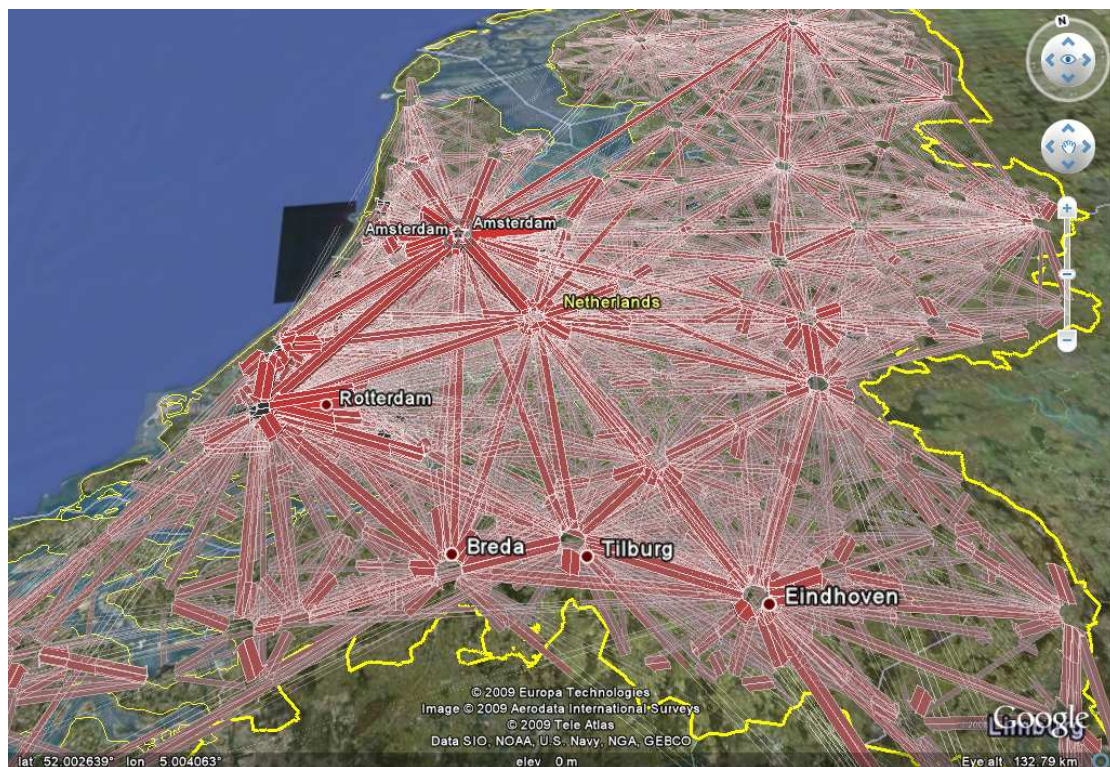
by the width of the flow arrow. The migration flows range from 1 person to 2,888 persons. This means that a linear proportional width is not an option. As said before the migration data is very skewed. The mean of migration size  $m_{ij}$  is approximately 10 persons. So even when we restrict ourselves to the migration flows with a size above average, the size would range from 10 to 2,888 (1 to 300). For the flow width we should therefore use a different scale. In our visualization we chose a log scale for arrow width.

Tobler (1987) puts more important flow in front of less important flows. This Z-order technique is also applicable in our case. It underlines the importance of flows with respect to less important flows. We implemented the Z Order as height of the flow using the 3D functionality of Google Earth. This technique puts the larger and significant flows in front of the visualization. Height also gives an indication of the size of the flow. Height can use a linear scale.

Our visualization has a large number of flow arrows and these obscure the overall picture. To reduce this clutter we used a technique that is commonly used in graph visualizations: we make less important flows more transparent than important flows. For the same reasons as given above a linear (transparency) scale is not an option. In our visualization we chose a logarithmic scale.

The direction of the flow is encoded using an arrow head.

We use the simple straight line routing scheme, with a small adaption. Flow arrows start and end at a small distance from the real origin and destination: Arrow heads are better visible and it reduces clutter. The alternative routing scheme of Doanton (2005) works only on unidirectional flow data. It is of course possible to generate for each region an outgoing flow map, but the combination of all these maps would be difficult to interpret. Since the flow layout algorithm changes flow routings, it would be difficult to spot where the flow originates.



**Figure 3 Internal migration 2006, slightly improved**

### 4.3 User interaction

Most flow map visualizations are static pictures. An interactive visualization gives a user control to solve visualization problems. A disadvantage of using Google Earth is that it is limited in possible user interactions.

Google Earth's main user interaction is panning, zooming and tilting. A user can use zooming and panning to magnify flows of a region. Tilting can be used to get an indication of the height of the flow.

A user can disable and enable flows per region, thus allowing the user to apply a filter based on region. It is however not possible to interact with flows.

A special version of the map has been created that shows in- and outflow per region. The flows for a region are shown when a user moves his mouse on the region. The flows are hidden when a user moves his mouse out of the region.

### 4.4 Discussion

In Figure 3 the result of the visual encoding and user interaction is shown. The resulting visualisation is not tested on users, but it shows a clear improvement over the naive implementation. The visualisation shows the major flows and shows many flows while still being readable. Non local flows are also visible (Groningen-Amsterdam, Rotterdam-Amsterdam).

The visualisation is however still far from optimal. Many detailed flows are hidden and a user has not the controls to make them visible. It is not possible for users to interact with flows themselves and get detailed information on a flow. Furthermore net migration is not clearly visible in the visualisation. Net migration could be plotted in a separate plot, but it would be nice if differences between in- and outflow would be expressed more clearly in the visualisation itself. Final point is that the patterns this visualisation reveals focus on large flows. It would be interesting to see a pattern of migration clusters but these are not visible. This is due to design and choice of the importance criterion. An alternative criterion could be to compare normalized flows. Normalizing means outbound flows are normalized to the population size of a region. The interpretation of a normalized outbound flow of a region would be the probability that an inhabitant migrates to the specified region.

## Conclusions and future work

Many official statistics are flow data or can be modelled as flow data. Simple flow data can be visualised using static Sankey diagrams or static cartographic flow maps. Large and skewed data which are common in official statistics are not easily visualised. To use Sankey diagrams for large datasets, an interactive Sankey diagram could be developed in which users can select and filter data. The layout algorithm for the Sankey diagram should be able to respond smoothly to the user interactions. This is direction for further research.

For unidirectional cartographic flow data an attractive flow visualisation exists that can be used. Visualising large and bidirectional flows is promising but still

problematic. First, user interaction should be improved. Users should be in control of the visualisation and be able to restrict and change the flow map very interactively. The options for user interaction in Google Earth are very limited, so a different graphics technique should be considered.

Secondly visual clutter should be further reduced. This might be done by creating different levels of detail using cluster and aggregation techniques. Future work would be to examine visualization of multiple hierarchical levels of aggregation, although Masser and Slater (1976) report that fine details get lost when doing hierarchical aggregation. An alternative approach would be to use the cluster information to visual encode the original flow map.

Not mentioned earlier but a very interesting direction for research is to add animation to show flow time series.

## References

- Beeckman, D., de Jonge, E. (2008), CartoGraphy with a capital G: Google and more, IAOS Shanghai Conference 2008
- Borden D. Dent. (1999), *Cartography: Thematic map design*. McGraph-Hill. New York.
- Doantam Phan, Ling Xiao, Ron Yeh, Path Hanrahan, Terry Winograd (2005), Flow Map Layout, *InfoVis 2005*
- Google Earth, available at <http://earth.google.com>
- Masser, I. (1976), The design of spatial systems for internal migration systems, *Regional Studies*, 10, 39-5
- Ormeling, F., Kraak, M.J. (1997) *Cartography, visualisation of spatial data*, Harlow Longman, 2<sup>nd</sup> edition.
- Sankey M.H.P.R (1898), The thermal efficiency of steam engines, *Minutes of Proceedings of The Institution of Civil Engineers*. Vol. CXXXIV, Session 1897-98. Part IV
- StatLine, Online Statistical Database of the Netherlands, available at <http://statline.cbs.nl>
- van Dongen, S. (2000), *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht
- van Dongen, S. (2000), *A cluster algorithm for graphs*. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam
- ten Bosch, O., de Jonge, E. (2008), Visualisation of neighbourhood statistics using Google Earth, *Meeting on the Management of Statistical Systems (MSIS 2008)*
- Tobler, W. (1987), Experiments in migration mapping by computer, *American Cartographer*