European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Improved Maritime Statistics with Big Data

Abboud Ado[1], Annica Isaksson de Groote[2], Ingegerd Jansson[3], Marcus Justesen[4], Jerker Moström[5], Fredrik Söderbaum[6]

[1] *Transport Analysis, Stockholm, Sweden; Abboud.Ado@Trafa.se*
[2] *Statistics Sweden, Stockholm; Annica.Isaksson@scb.se*
[3] *Statistics Sweden, Stockholm; Ingegerd.Jansson@scb.se*
[4] *Statistics Sweden, Stockholm; Marcus.Justesen@scb.se*
[5] *Statistics Sweden, Stockholm; Jerker.Mostrom@scb.se*
[6] *Transport Analysis, Östersund, Sweden; Fredrik.Soderbaum@Trafa.se*

Abstract

In recent years, Big Data as a potential data source for official statistics has attracted much interest. In a joint project, Statistics Sweden and Transport Analysis (the government agency responsible for official transportation statistics) have explored the potential of a type of geographical Big Data, AIS data, for improving maritime transport statistics. In this paper we focus on one important goal of the project: to evaluate the usefulness of AIS data for improving the quality of the distances between Swedish ports. Two different methods are tried and described in this paper. Our study suggests that AIS data have great potential to improve maritime statistics, but further work is needed.

Keywords: AIS, position data, distance matrix, ports.

1. Introduction

In recent years, Big Data as a potential data source for official statistics has attracted much interest. Big Data can basically be described as huge volumes of data generated at short intervals, often in unstructured form, which due to their complexity and volume require innovative methods and technological solutions to extract useful information. Our work is based on a special kind of Big Data with a geographical component: AIS data. AIS (Automatic Identification System) is a system that makes it possible for ships to identify and track other ships' movements. The system also provides detailed information about the ships, including their identity, size, position, course, speed, type of cargo, and destination. The system relies on digital information transmitted by the ships at short intervals and received by other vessels via their own AIS equipment. The information is also received on land through a network of AIS base stations. The use of AIS data is an application of Big Data undergoing strong development, and there are several national and international examples of research studies and pilot projects on the topic. So far, however, systematic implementation of AIS data in statistical production is rare. One exception is the Swedish Meteorological and Hydrological Institute which uses AIS data to calculate emissions from shipping (SMHI 2013; SMED 2012).

This paper is based on a pilot study in 2015, jointly conducted by Statistics Sweden and the Swedish government agency Transport analysis. The aim of the pilot was to investigate the potential of AIS data to improve maritime statistics. Swedish official statistics on maritime transports are based on survey data from Swedish ports. Data are collected quarterly and the port offices are asked to provide information on ships arriving at the ports. A variable of interest to most users of maritime statistics is *transport performance*: the amount of transported goods (or number of passengers) multiplied by the travel distance. In order to calculate transport performance, information about the distances between ports is needed. At present, flat-rate tables on the distances between ports are used in the calculations. The flat-rates are derived from a Port Distance Calculation Tool developed by Eurostat (Eurostat 2009). The Eurostat tool is however not ideal for this use. Above all, it is mainly developed to provide distances between ports in different European countries. The calculated distances cannot be divided between international and national waters. Also, between ports where traffic is known to occur, information on the distances is sometimes missing.

In this work, we have tested the use of AIS data in order to simplify the data collection from the ports, and to improve the quality of the calculated distances between ports.

2. Selection of data

The pilot study is based on a test period of two weeks in 2014: 7-20 September. Passenger traffic was excluded from the analysis. Each day generated about one million data points within the study area.

The geographical area used for the selected data was the entire Baltic Sea and Skagerrak and Kattegat. The purpose of this relatively wide geographical area was to capture all possible routes between Swedish ports.

3. Preparatory processing

In order to create a high-definition matrix using AIS data, we prepared the data in four steps:

1. A data set of port areas was created, containing:
    - Swedish ports areas,
    - a dummy area representing foreign ports, and
    - water (outside the ports).
2. AIS data points were categorized to fall into one of the three areas defined in Step 1.
3. For each transport that occurred between ports (port areas), a line connecting the AIS data points was created based on the available information on port, time stamp and Maritime Mobile Service Identity (MMSI).
4. Each line was updated with information about origin and destination of the vessel.

Step 1-4 created a population of routes which was used both to verify existing model-based distance calculations and for developing a new distance matrix. From the population of routes, it is possible to select all transport operations between Swedish ports, all transport operations to or from a specific port, etc., see Figure 1.

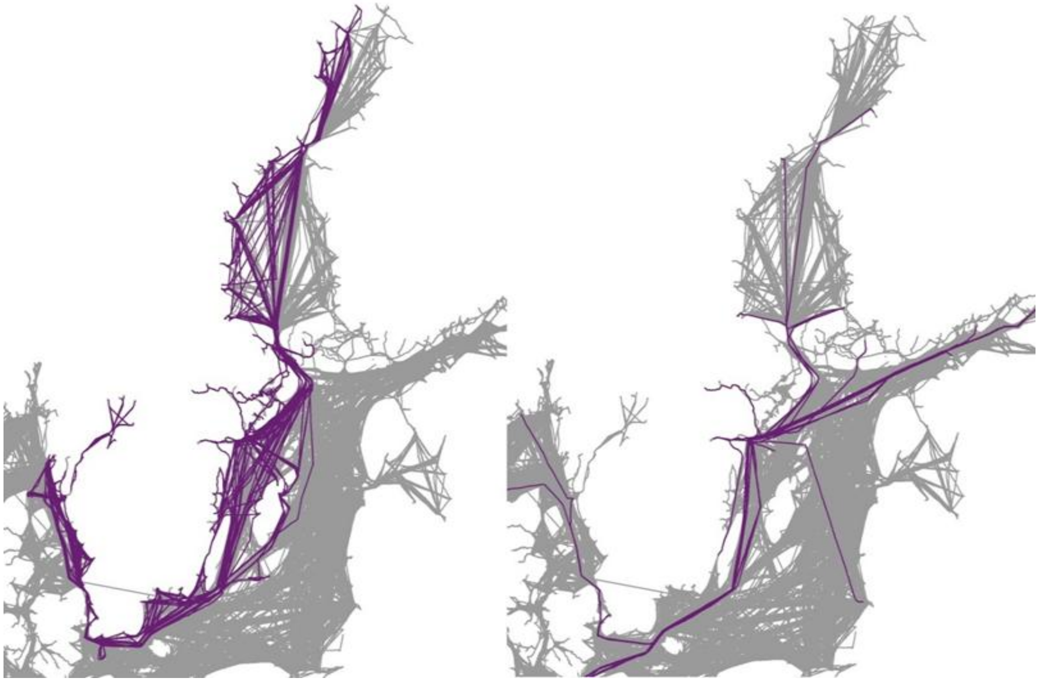

Figure 1 Left: map of all traffic between Swedish ports (in purple). Right: map of all traffic to Norrköping (in purple).Grey lines represent all traffic in the area

Using AIS data directly to calculate travel distance and transport performance for a certain period of time provides very accurate results as AIS data depicts the actual routes of all vessels in the study area. On the other hand, a direct use of AIS data is very time consuming as it involves repeated processing of large quantities of data.

Another option is to use a model approach. A model is easier to handle and gives an accurate result, however based on ideal distances between ports rather than actual measured routes. The model needs to be based on AIS data, but instead of recalculating the actual measured routes every time, the calculations can be made once and then fed into the model.

The basis for a distance matrix is the created population of routes between each pair of ports. It is thus important that the created routes correspond to the common transport routes from each port to all other ports. Hence, for a good performance of the model the population of routes needs to contain enough data to cover all possible routes, ideally with a large number of observations.

4. Development and modelling

Geographic information can be processed in two formats: vectors or raster. Typically, the vector format is used to create a network: the lines are connected by nodes and transport operations take place along the lines. We have however chosen to use raster. A raster in GIS can be summarized as a surface composed of cells with a specific definition, such as 1 x 1 meter, which contains a value for each cell. The value in the cell is chosen to represent any information necessary for modelling the network; for example, to exclude certain cells from a transport route. We have several reasons to believe that raster is more suitable than vectors for our particular purpose:

- Maritime transport operations take place on a *surface*, they are not limited to the roads the way road traffic is.
- Raster can contain a lot of data, because it is only a value in a cell.
- Raster is faster to process.
- Raster is flexible and easy to experiment with.
- New data can be added to a raster.

In order to define routes on a raster surface we have calculated the minimum cost to move from an origin to a destination, where the cost is equal to friction times distance. Friction can be anything, such as a road gradient, time, speed, or wind. Several different factors can be combined and weighted to achieve the desired friction. It is the friction that determines the route. We use the following steps to create a raster-based model:

1. All lines are converted to a raster with a one kilometer resolution.
2. For each raster cell, line density is calculated. Line density determines the waterway; if the line density is high, the traffic is high and a waterway is identified.

3. We use the waterway combined with destination (the port we want to calculate the route to) to define the cost to travel through each cell.

4. A distance raster is created from the cell to the port (source) of interest. Each cell in the raster is assigned a value equivalent to the least accumulated cost of being transported from the cell to the source.

5. Finally, the route is calculated as the path with the least accumulated cost from specified port or ports to the port of interest based on the distance raster.

Note that the traffic density is not the main factor in the weighting. It only acts as a complement to the *destination* which is the most important factor. With a computation of routes based only on density and distance, a calculated route can differ significantly from the actual route. This is illustrated in Figure 2. Adjustments are needed to handle this, such as destination that also can be added as a friction.

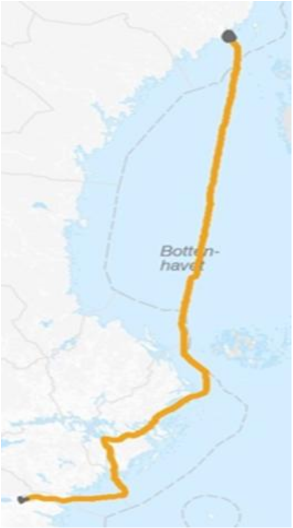

Figure 2 Route from the north of Sweden to Norrkoping calculated using only line density.

Another aspect is the position of the original port when routes to a destination port are calculated. Depending on the choice of traffic; only from Swedish ports, or all traffic, the results of the route calculations differ. Figure 3 illustrates these differences. Note how the route from northern Gotland changes because of the transport operation coming from Lithuania.

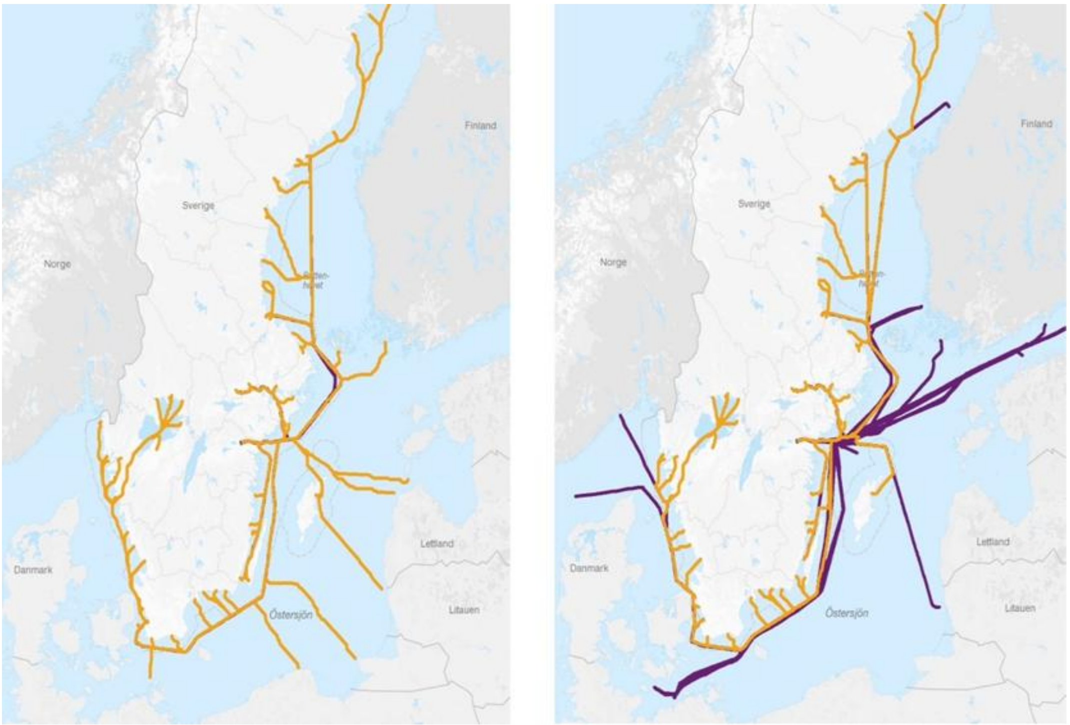

Figure 3 Calculated routes from 70 ports to Norrköping. In the left image, only transport operations from Swedish ports to Norrköping have higher weights. In the right image, all transport operations to Norrköping (even from foreign ports) have higher weights.

5. Conclusions

We have developed preliminary methods for identifying routes between Swedish ports, and between Swedish and foreign ports. The same methods are likely to be useful also for processing other types of mass-generated position data (such as data from other mobile devices). The raster-based model for calculating distances has shown promising results. The model needs however further development. In summary, our study suggests that AIS data have great potential to improve maritime statistics, but further work is needed.

6. References

Eurostat (2009), Methodology for Maritime Network description, Unpublished report.

SMED (2012), Uppdatering av typfartyg för svensk inrikes sjöfart, SMED rapport Nr 135.

SMHI (2013), A dynamic model for shipping emissions. Adaptation of Airviro and application in the Baltic Sea, METEOROLOGY, 153.