Data Mining and Big Freight Transport Database Analysis and Forecasting Capabilities

Massimiliano Petri, Antonio Pratelli, Giovanni Fusco

Data Mining and Big Freight Transport Database Analysis and Forecasting Capabilities

Massimiliano Petri[a], Antonio Pratelli[a], Giovanni Fusco[b]

Transport modeling in general and freight transport modeling in particular are becoming important tools for investigating the effects of investments and policies. Freight demand forecasting models are still in an experimentation and evolution stage. Nevertheless, some recent European projects, like Transtools or ETIS/ETIS Plus, have developed a unique modeling and data framework for freight forecast at large scale so to avoid data availability and modeling problems. Despite this, important projects using these modeling frameworks have provided very different results for the same forecasting areas and years, giving rise to serious doubts about the results quality, especially in relation to their cost and development time. Moreover, many of these models are purely deterministic. The project described in this article tries to overcome the above-mentioned problems with a new easy-to-implement freight demand forecasting method based on Bayesian Networks using European official and available data. The method is applied to the Transport Market study of the Sixth European Rail Freight Corridor.

KEY WORDS
~ Freight demand model
~ Bayesian networks
~ European freight corridor
~ Demand forecasting

a. Dept. Civil and Industrial Engineering, University of Pisa, Pisa, Italy
e-mail: m.petri@ing.unipi.it
b. Centre National de la Recherche Scientifique, Université de Nice Sophia Antipolis, Nice, France
e-mail: giovanni.fusco@unice.fr

1. INTRODUCTION AND GOALS

Transport modeling in general and freight transport modeling in particular are becoming important tools for investigating the effects of investments and policies, involving large amounts of resources. However, freight demand forecasting models are still in evolution stage (Arnone et al., 2011) for the following reasons:

- lower seniority (about 10 years) than the respective passenger models;
- high number of decision-makers to consider (companies, shippers, carriers, logistics operators, port operators, deposits, etc.);
- variety of products transported (in terms of categories, dimensions, weight, value, etc.);
- high variability in decision-making processes;
- limited availability of information (data often aggregated, dated, partial, heterogeneous, etc.).

To take into account the complexity of freight transport system, researchers have proposed a wide array of models belonging to the aggregate or disaggregate model types (Tavasszy, 2006) and to three different fields: the modeling of the relationship between transportation and economic activity, logistic decision making and processes and the link between traffic flows and networks (Ben-Akiva et al., 2013).

Recently, European projects like Transtools (Burgess et al., 2008) ETIS/ETIS Plus (NEA Transport research and training BV, 2005; Chen, 2011) have developed a unique modeling and data framework to forecast freight flows at large scale, so to avoid data availability and modeling problems (Albert and Schafer, 2013). Despite this, very important projects using these modeling frameworks have provided very different results for the same forecasting areas and years, giving rise to serious doubts about the results quality, especially in relation to their cost and

development time. For example, there is a very high divergence between the results of the two projects Prog-Trans and TransTools for truck flows (Germany in TransTools has an increase in freight transport tonnage in 2005-2020 of about 10 % while in Prog-Trans this value is about 50 %) (Petri et al., 2014).

This is a general problem for freight modeling and forecasting, with a high-complexity analysis level applied to a very large scale, bringing to errors many times uncontrollable.

Moreover, many of these models are purely deterministic in results, giving no information about their estimation errors or the probability of the occurrence of forecast values. Other problems include forecasting different scenarios with very long-term simulations. We think that projects of national/European importance would benefit from the contribution of probabilistic data-driven models that take into account the uncertainties and variability of attributes and scenarios, especially for long-term estimates, in order to have more truthful decision-support.

There are a lot of freight demand models (Chase et al., 2013), with some methods similar to the one adopted here like the use of Trend Analysis/Time series or Neural Networks (National Cooperative Freight Research Program, 2010), but Bayesian Networks have the advantage to allow the introduction in the model of expert knowledge and the possibility to verify the results (Onsel et al., 2013) that are in form of an easy-to-understand oriented causal graph among variables and not complex or black-box relations, as with Neural Networks (Floreano and Mattiussi, 1996).

The objective is mainly to understand quantitative and qualitative aspects of future traffic demand and evaluate possible future scenarios according to the most relevant and influencing variables of the freight market (Meersman and Van de Voorde, 2013). We also want to overcome the above mentioned problems with a new freight demand forecasting framework based on Bayesian Networks and using European official and available data. The model has to be easy to implement, not onerous and give probabilistic results in less time, with an estimation error similar to the more complex methods. It should be capable of giving order of magnitude of forecasted freight flows for strategic decision making at a very early phase of policy development, and be complementary to more traditional, more precise, but much more expensive freight models for later stages of the analysis.

2. NEW METHODOLOGY FOR FREIGHT DEMAND FORECAST

In our study, we applied the general demand forecast methodology to freight flows within the Transport Market study of the Sixth European Rail Freight Corridor. The European parliament and the Council adopted on 22 September, 2010 the EU Regulation 913/2010 concerning the European rail network for competitive freight. Within this framework, the EU identifies nine rail corridors; in particular, Rail Freight Corridor 6 (RFC6) allows railway connections among Spain, France, Italy, Slovenia and Hungary, also providing links with rail freight corridors 1, 2, 3, 4, 5 and 7 (see violet line in Figure 1).

Regulation 913/2010 sets two main goals:

- To develop the rail freight corridors in terms of infrastructure capacity and performance to meet market demand on both quantitative and qualitative layers;
- To lay the groundwork for the provision of good quality freight services to meet customer expectation.

Regulation 913/2010 requires the Transport Market Study (TMS) for each freight corridor, developed according to a clear "corridor perspective", with a coherent structure for the entire corridor and not as a collection of studies focused on individual Member States. The Transport Market Study is intended as the basis for the assessment of the customer needs.

The main goal of the TMS for RFC6 is to provide a clear understanding of the current conditions of the multimodal freight market along the corridor as well as to develop short-term and long-term traffic forecasts (volumes and modal split/modal shift) also including the effect of actions and measures related to the implementation of the Corridor itself. Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts as well, for math, etc. Consequently, the Transport Market Study is aimed to:

- Analyze the current situation in terms of transport demand and supply and economic context;
- Analyze the transport market in terms of customer needs and deliver information on modal choice process;
- Provide transport demand projections after the implementation of the corridor itself.

Figure 1.
Rail Freight Corridor 6 (RFC6).

2.1. Structure of the study

The study is organized in three Phases:

- Phase 1: Analysis of the present situation;
- Phase 2: Survey (Releaved Preferences-RP and Stated Prefereces-SP surveys);
- Phase 3: Short-term and long-term transport demand forecasts.

Phase 1 provides direct final results and creates the background to structure, design and implement Phase 2 and Phase 3. In particular, Phase 1 is aimed at providing a sound analysis of the present socio-economic situation and of the future scenario in the countries crossed by Corridor 6 within the wider EU framework, making clear the full picture and deriving first qualitative policy indications and guidelines.

Consequently, Phase 1 provides information in terms of:

- Present and future economic magnitude of countries and/or regions along Corridor 6;
- Present transport demand across the Corridor (macro-flows among countries and/or regions, including flows to areas not directly served by the Corridor itself);
- Future transport demand (at macro level) in terms of likelihood of increase (macro potential demand and macro role of the railway transport in terms of modal split, volumes and values of carried goods based on the evolution of the future competitive positioning of countries crossed by Corridor 6).

Phase 2 aims at engineering and implementation of surveys on the decision path in the choice of transport mode. In particular, the survey and its analysis provide a complete picture of the main factors affecting the choice of transportation including:

- cost of transport;
- travel time;
- risk of delay in delivery;
- risk of damage or theft.

The surveys aim at several transport market actors:

- manufacturing firms which directly organize the shipping / receiving of goods;
- intermediaries which organize the transport of goods on behalf of producers and /or final users;
- operators of rail transport networks and intermodal centers.

Based on the results of Phases 1 and 2, Phase 3 provides estimates of freight transport that could be carried out on Corridor 6 at different time horizons (2015 and 2030).

Phase 3 is divided into two distinct steps:

- estimate of the total (road and rail) freight transport demand in 2015 and 2030;
- estimate of the modal split road / iron as a function of hypothetical scenarios characterized by the variation of the main features of the transport (cost, time, delay, damage / theft).

The present paper is then aimed at explaining key quantitative and qualitative analysis of the Rail Freight Corridor 6 Transport Market Study. The methodology and its detailed results regarding the data mining methods are used for the actual state analysis (Phase 1) and for the freight trasport model implementation (Phase 3-first step).

The key steps of the different activities based on the data mining techniques performed and described here are:

- the input data analysis;
- the Decision Tree Induction model analysis (Witten and Frank, 2000);
- the final freight demand forecasting by Bayesian Network models ((Pearl, 2000) and more particularly (Fusco, 2010), (Fusco, 2012) for their use as spatial strategic forecasting tools).

These steps are logically connected. The input data analysis allows to know how each input variable influences the actual freight flow dynamics in terms of relative growth (i.e. percentage variation between reference years 2005 and 2010) so as to understand which variables are directly (with the uni-variate analysis) or indirectly (bi-variate and tri-variate analysis) related to it. The Decision Tree classification refines this preliminary analysis with a complex multi-variate elaboration having always the freight flow dynamics (the evolution of the freight flows between 2005 and 2010) as target variable. Finally, the Bayesian Network models use as input data only the most influencing variables in order to avoid irrelevant data in the model, resulting in errors and reduction in the forecasting capacity.

The Bayesian Networks models were finally used to forecast freight flows in different scenarios. More precisely, the final traffic forecasts were carried out according to three different estimates of GDP growth for the study area: basic, optimistic, conservative. The demand forecasting models were developed with reference to two different geographic areas: at first, the analyses were conducted with reference to the mobility data of the whole European O/D Matrix, later it was decided to focus only on the area interested by Corridor 6 and to calibrate the model accordingly in order to obtain more reliable estimates.

2.2. The Preliminary Data Analysis

A first socio-economic analysis was made to evaluate and estimate scenario for important input variables. For example, population and its evolution can be considered as a proxy of future trends for goods production and demand. Total population is about 184 million, against the European population of about 521 million. Corridor countries' population has been growing faster (CAGR +0,8 %) than Europe as a whole (CAGR +0,4 %) despite the negative trend in Hungary (Table 1).

Despite the negative impact of the economic downturn on the relevance of historical trends, medium term forecasts (in particular for year 2030) can provide a higher level of consistency, neutralizing short term fluctuations. For year 2030, in real prices GDP grows (base case) by about 28 % both for the countries

Table 1.
Gross domestic product (bn €) and population (m).

	GDP			Population		
Zone	**2008**	**2011**	**CARG % (2003-11)**	**2008**	**2011**	**CARG % (2003-11)**
Spain	1.087,70	1.063,40	3,9	45,3	46,2	1,3
France	1.933,20	1.996,60	2,9	64	65	0,6
Italy	1.575,10	1.579,70	2,1	59,6	60,6	0,7
Slovenia	37,3	36,2	4,3	2	2,1	0,3
Hungary	105,5	99,8	3,8	10	10	-0,2
Europe	13.152,80	13.499,50	3,1	515,9	521	0,4
Corridor	4.738,90	4.775,60	2,9	181	183,9	0,8

Table 2.
Gross domestic product growth rates (average % change over the previous year).

Zone	2012	2013	2014	2015	2020	2025	2030
Spain (F)	-1,40 %	-1,40 %	0,80 %	1,60 %	1,60 %	1,60 %	1,60 %
France (G)	0,00 %	0,10 %	1,20 %	1,70 %	1,70 %	1,70 %	1,70 %
Italy (H)	-2,20 %	-1,00 %	0,80 %	1,30 %	1,30 %	1,30 %	1,30 %
Slovenia (I)	-2,00 %	-2,00 %	0,70 %	1,30 %	1,30 %	1,30 %	1,30 %
Hungary (J)	-1,70 %	-0,10 %	1,30 %	1,20 %	1,20 %	1,20 %	1,20 %
Europe (K)	-0,20 %	0,20 %	1,60 %	1,40 %	1,40 %	1,40 %	1,50 %
Corridor (L)	-1,10 %	-0,60 %	1,00 %	1,50 %	1,50 %	1,50 %	1,50 %

Table 3.
GDP growth rates by scenario (average % change; 2011-x).

	2015			2030		
Zone	**Low**	**Basic**	**High**	**Low**	**Basic**	**High**
Spain (F)	-0,50 %	-0,10 %	0,30 %	0,80 %	1,20 %	1,70 %
France (G)	0,50 %	0,70 %	1,00 %	1,00 %	1,50 %	1,90 %
Italy (H)	-0,70 %	-0,30 %	0,10 %	0,60 %	1,00 %	1,40 %
Slovenia (I)	-1,00 %	-0,50 %	-0,10 %	0,50 %	0,90 %	1,30 %
Hungary (J)	-0,20 %	0,20 %	0,50 %	0,60 %	1,00 %	1,30 %
Europe (K)	0,40 %	0,70 %	1,10 %	0,90 %	1,30 %	1,70 %
Corridor (L)	-0,10 %	0,20 %	0,50 %	0,80 %	1,30 %	1,70 %

crossed by Corridor 6 and for Europe, but with significant internal differences (France and Spain grow more; Italy, Slovenia and Hungary grow less). GDP growth rate is assumed according to specific annual forecasts (made available in winter 2013) for the years 2012, 2013 and 2014 and on average trends since 2015 on (average official long term trends to 2060, to neutralize short terms fluctuations) (Table 2 and 3).

To cope with uncertainty in long term forecasts, low and high sensitivity scenarios (GDP growth higher or lower than in base case) are introduced.

The Statistical initial data analysis was carried out on the whole road and rail ETIS Origin-Destination Freight Flows Matrix in Europe for 2005 and 2010 years. Origins and Destinations in this database are known at the NUTS 2 level. The original road 2005 O/D matrix has thus about 134,000 O/D pairs while the corresponding 2010 matrix has only 102,000 O/D pairs. 88,000 O/D couples are common to the two matrices. Taking into account only these common data (88,000 O/D pairs), we lose around 4 % of the total flows (containing also the flows not interesting directly for the Corridor 6). For each O/D couple, an evolution rate between 2005 and 2010 could thus be calculated. Together with freight flows, the starting data include twenty variables belonging to different fields like economy, geography and transportation summarized in Table 4.

Table 4.
List of the initial variables (in rose color the variables changing for the three scenarios are indicated (best, regular and worst).

ID	Indicator	Starting year	Forecast year 1	Forecast year 2	Scale start year	Scale forecast year
1	GDP (Gross domestic product) of NUTS2i	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS2	NUTS0
2	GDP (Gross domestic product) of NUTS2j	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS2	NUTS0
3	Total population of NUTS2i	2010	2010	2010	NUTS2	NUTS0
4	Total population of NUTS2j	2010	2010	2010	NUTS2	NUTS0
5	GCF (Gross capital formation) of NUTS2j	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
6	PMGS (Production of Manifactured Goods Sold) of NUTS2j	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
7	PV (Production value by industry) of NUTS2j	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
8	IG (Import of goods) of NUTS2i	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
9	EG (Export of goods) of NUTS2i	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
10	Total Freigth floes between NUTS2i and NUTS2j	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS2	NUTS2
11	Minimum distance between NUTS2i and NUTS2j	2010	2010	2010	NUTS2	NUTS2
12	Macroregion Name of NUTS2i	2010	2015	2030	NUTS1	NUTS1
13	Macroregion Name of NUTS2j	2010	2015	2030	NUTS1	NUTS1
14	NMF - Net migration flows of NUTS2i	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
15	NMF - Net migration flows of NUTS2j	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
16	Unemployment rate of NUTSi	2010	2010	2010	NUTS0	NUTS0
17	Transport taxation revenues of NUTS2i (milion of €)	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
18	Transport taxation revenues of NUTS2j (milion of €)	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
19	Diesel price of NUTS2i (€/litre)	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0
20	Diesel price of NUTS2j (€/litre)	DELTA 2005 - 10	DELTA 2005 - 15	DELTA 2005 - 30	NUTS0	NUTS0

This general data analysis phase explores the freight flow dynamics and its correlation with the main variables, some of which are normally used in Transport Distribution Models (like distance, population and GDP) while others are not included in these models but can be used in data-driven Bayesian Network learning (for example, unemployment rate, variation of origin export, and destination import, or binary variables, like the belonging to the EU) (Caplice and Phadnis, 2010).

The initial data analysis is divided into three parts of increasing complexity: orthogram, bi-variate and tri-variate analysis. The following analyses concern only the road and rail freight flows because they are the most interesting for Corridor 6 study area (Figure 1).

The first part of the preliminary data analysis uses some correlation tools at different complexity level; for the simplest part, some bi-variate correlation analysis has been elaborated, for example:

- Distance – Delta flow 2010/2005
- Population 2010 – Delta flow 2010/2005
- Unemployment 2010 – Delta flow 2010/2005
- Delta Export 2010/2005 – Delta flow 2010/2005

Increasing the complexity, a tri-variate analysis has been elaborated as for the correlation between Origin Delta GDP, Destination Delta GDP and Delta flow 2010/2005.

Before using data mining methods, there has been, moreover, implemented an Orthogram analysis, e.g. for the two following variables: UE Belonging - Delta flow 2010/2005.

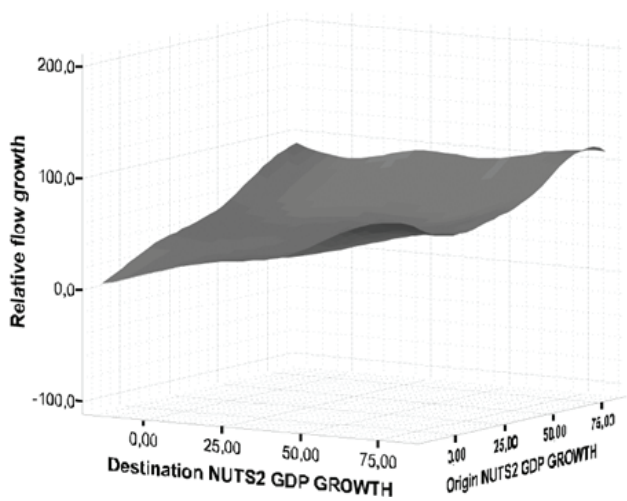


Figure 2.
Smoothing interpolation of 3D Scatterplot between Origin and Destination GDP growth and freight flow variation.

The bi-variate analysis shows that the correlation of freight flow dynamics is practically absent both with the distance between Origin and Destination (measured in kilometers on the transportation networks), with Origin Population, with unemployment rate at the Origin or with the Origin Export Variation and with the Destination Import Variation (Import and Export variations are known at the country level).

The tri-variate analysis correlates simultaneously the freight flow variations with Origin and Destination GDP variations. The 3D scatterplot, with a smoothing interpolation effect (Figure 2) indicates an overall positive correlation between these three variables with more specific local trends. The results of the tri-variate variables analysis is shown in Figure 2.

To better understand this last point, two 2D scatterplots are extracted from the 3D diagram (Figure 3). The correlation is similar for the destination and origin GDP variations. Curiously, a positive flow growth characterizes even the negative GDP Variations, showing for some countries an inverse correlation, which could indicate a profound restructuring of the economy following the integration in the European market. More than linear flow growths are to be observed beyond 75 % of GDP increase rate (more evident for Origin than Destination).

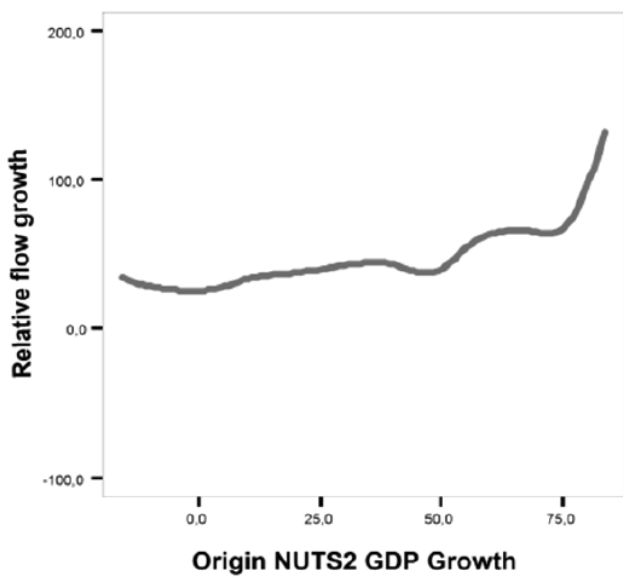
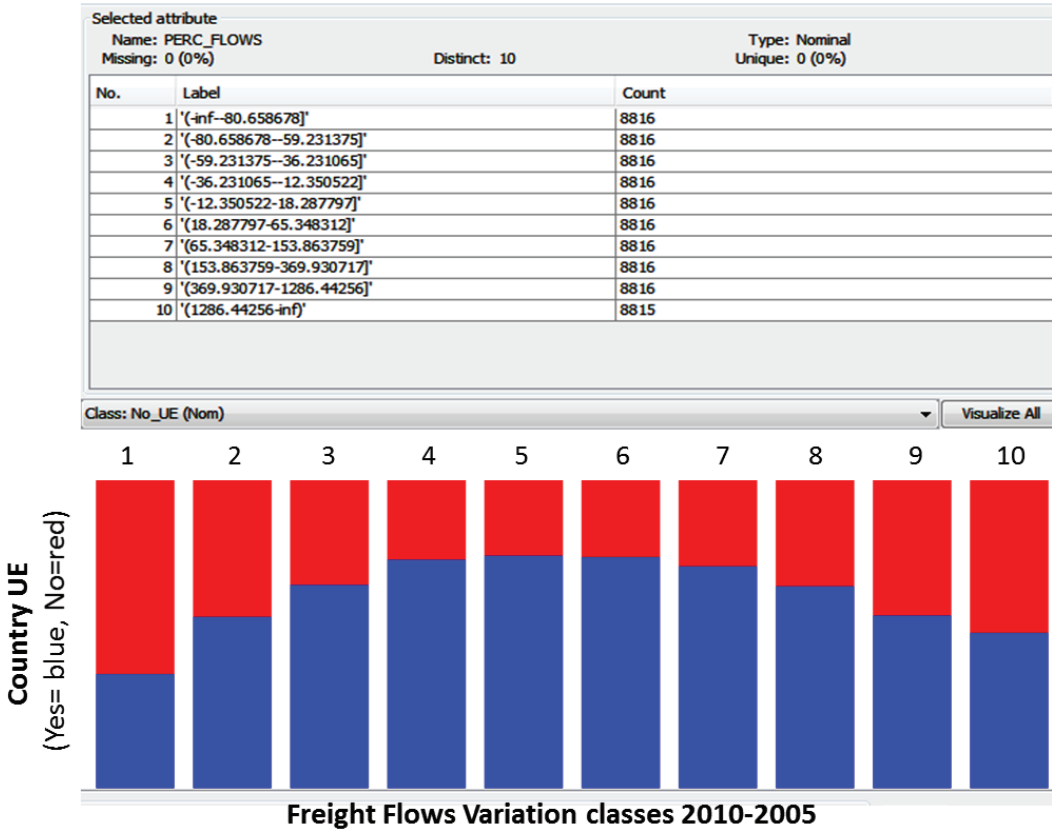


Figure 3.
The results of two bi-variate analysis corresponding to the previous tri-variate one.

The orthogram analysis (Figure 4) allows studying the correlation between different kinds of variables (categorical and numeric for example).



No.	Label	Count
1	'(-inf--80.658678]'	8816
2	'(-80.658678--59.231375]'	8816
3	'(-59.231375--36.231065]'	8816
4	'(-36.231065--12.350522]'	8816
5	'(-12.350522-18.287797]'	8816
6	'(18.287797-65.348312]'	8816
7	'(65.348312-153.863759]'	8816
8	'(153.863759-369.930717]'	8816
9	'(369.930717-1286.44256]'	8816
10	'(1286.44256-inf)'	8815

Figure 4.
The Orthogram analysis (freight flow variation are indicated in percentage).

The analysis shows that 2005-2010 freight flow variation is correlated with the belonging or not of each area to the EU: there is a clear distinction between areas belonging to the EU and other areas. EU countries have more stable freight flows, while non-EU countries have opposite behaviors with some of them showing a big increase of freight flows and others a considerable decrease. These bi-modal behaviors are difficult to model with classical Transport Distribution Models. This first analysis already shows the interest of using different, more exploratory methods like Decision Tree Induction and Bayesian Network modeling.

2.3. The Decision Tree Induction Classification

Decision Tree models are useful multivariate classification instruments allowing analysis of data correlation on the basis of a target variable, O/D freight flow relative growth. Moreover, instead of regression models, where we need to hypothesize a shape of the correlation (linear, cubic, exponential, etc.), Decision Tree models do not require any assumption and give more than one type of correlation. Finally, the IF THEN framework is very useful and understandable for users and Decision Tree models can be used as a preliminary phase for the Bayesian Network modeling in order to understand the most influential variables to simulate the target one. Decision Trees Induction is an inductive classificatory technique belonging to the Data Mining and to the Knowledge Discovery in Database fields. It will be applied to the complete list of variables (Table 4), keeping the O/D freight flow relative growth as target variable.

The extracted classifier has a percentage of Correctly Classified Instances of about 38 %, which appears relatively inaccurate. However, the analysis shows two main points:

- the classification ability is higher for the first and last flow variation classes and for the class nearest to zero;
- once again, distance (DIST_2010) between the individual Origins and Destinations does not have a relevant influence.

The analysis suggests introducing new variables so as to add detail in the information (GDP at NUTS 2 Level, Internal, Belonging to EU and others) and to add interaction between territorial dimensions at NUTS2 and NUTS0. The new variables are:

- Internal (indicates if an O/D couples belong to the same

country);

- No_EU (indicates if an O/D couples belong to EU countries or not);
- Delta GDP 2010-2005 at NUTS2 level;
- Flow 2005 (to indicate flow level before the 2008 economic crisis);
- EU15_CH_NO (indicate whether a flow belongs to the 15 EU member states before 2004 plus Switzerland and Norway);
- Weight of the exit flow for a given origin = Fij/Fi.;
- Weight of the entry flow for a given destination = Fij/F.j;
- Weight of exports to Country J from i = FiJ /Fi.;
- Weight of imports from Country I to j = FIj/F.j

where FiJ means total flows from NUTS2 i to all Country J, while F.j means total exit flows to NUTS2 j.

Introducing these new variables, the extracted Decision Tree identifies the variable "weight of the exit flow" as the most important one and shows the relatively chaotic evolution of flows for the non-EU countries. Decision Trees results for the whole ETIS O/D Matrix describe a non-unique freight traffic evolution with different variables explaining the flow growth for each country and mainly different from the countries being or not among the early EU member states. The only shared important variable is the weight of the exit flow (Fij/Fi) showing the relative importance of the economic relation between the origin and destination areas with respect to all the exit flows.

The Decision Tree extracted from the same variables but including only O/D flows belonging to the area of interest for Corridor 6 shows clearly two different dominant behaviors:

- the first is related to the countries with more stable economy and freight market, where the only element that explains the freight dynamics is the actual weight of the outgoing flows (this concerns more than 50 % of the total flows);
- the second is the already noted bi-modal behavior.

2.4. The Bayesian Network Forecasting Model

The Decision Tree technique produces knowledge only for the pre-processing phase. This limit of the technique is mainly due to the difficulty of the application of the rules extracted from the sample to the whole population:

- firstly, it is possible that a combination of conditional attributes never occurred in the extracted rules (IF part), whereas it can be present in the prevision dataset; the problem would then be to compute the relative conditional probability distribution;
- secondly, it could also be possible not to find a rule exactly identical (in the IF part) with the record to be classified: this problem can be solved only with the search of an attribute set close enough to the one to be classified.

Due to these possible situations, the extracted influent variables were used as input variables to implement a Bayesian Network. Bayesian Networks are more suitable to predict phenomena due to their robustness (they can couple statistical robustness from data mining to expert knowledge directly implemented in the model, whereas Decision Trees are only based on data frequencies) and the possibility to make probabilistic inference so as to have probability values attached to predictions. Even in the absence of expert knowledge (as in our application), prior probabilities in the network initialization produce non-null probabilities for combination of attributes that are not present in the learning data-base. Through Bayesian learning algorithms from data (Jensen, 20011), the model links the variables in acyclic and directional graphs, showing their reciprocal influence in a cause-effect relationship between "parent" and "child" nodes. Finally, a conditional probability table is calculated for each dependent variable (with incoming link in the node), detailing the probabilistic relationship between the values of the "parent" and "child" variables. Unconditional probability tables are calculated for independent variables (without incoming links in the node). Learning algorithms search for the best possible combination of structure (links among nodes) and parameters (probability values in the tables) within a subspace of possible solutions. The best solution is found through likelihood maximization, knowing the empirical data.

Different Bayesian Network models were calculated from data covering the whole ETIS O/D Matrix, or just the area of interest for Corridor 6. Continuous variables were discretized in eight classes of equal frequencies (other discretizations were also attempted). Each model allows probabilistic inference of O/D freight flow relative growth between 2005 and 2010 from 2005 and 2010 data. Under the assumption of model stationarity, the probabilistic relationships embedded in the model can be used to infer O/D freight flow relative growth between 2010 and 2015 (end hence 2015 freight flows) from 2010 data and scenarios on 2015 data. A more problematic stationarity assumption was also used in order to forecast 2030 freight flows.

1) The forecast for the whole ETIS O/D matrix

The final model for the whole ETIS O/D Matrix (Figure 5) shows that the most important variables in order to forecast freight flow relative growth are the GDP national growth in the country of origin (NUTS2 GDP growth had too many missing data to produce statistically significant links in the model) and the relative importance of the outflow for the origin (weight of the exit flow Fij/Fi.). The mutual information analysis (resumed by the position of each node within the model) shows a clear clustering of economic (with internal circle in dark grey) and geographic (without internal circle) variables.

A first validation of the extracted Bayesian Network concerns its predictive power in inferring the value of the target variable of flow relative growth knowing the other variables. The resulting confusion matrix shows that the model can predict values of the target variable with a total precision of 25 % when considering the prediction of the exact variation class, but of

more than 50 % when considering prediction of the right class or of the two (eventually one) nearest ones (flow growth rates are discretized in eight classes). The second validation tests the model generalizability (or presence of over-fitting problem) through a ten-fold cross validation (i.e. the iterative use of 9/10 of the total O/D data to build the network and 1/10 of the total O/D data to validate it). The results of cross validation are very similar to the initial model, which leads us to the conclusion that the model does not have particular over-fitting problems. During the cross validation, another validation of the model regards the stability of its network structure (called confidence analysis) and relative variable dependencies (represented by the arc connections) in the ten simulated networks. The arcs directly connected with the target variable (flow growth) remain always the same and are present in all the networks produced within the cross validation (100 %).

The first problem of this methodology arises when we need to use the probabilistic results of the Bayesian Network inside the Discrete Choice model (Ben-Akiva and Lerman, 1985), which is based on deterministic values of total demand and, on the basis of Revealed Preferences / Stated Preferences interviews (RP/SP, (Danielis and Rotaris, 1999)), elaborates probabilistic results on the modal split. In our application, the modal split predictions are carried out using the weighted average of the median value of each flow variation class. An example is shown in Figure 6 with the probability distribution for the target variable freight flow growth. For each of the eight classes, the central value is reported in the right column and is used to calculate the expected mean value (-26.17 % in the example) as the weighted average (on the predicted probabilities) of the mean class values.

Once the Bayesian Network model is calibrated for 2010 (base year), scenario values can be defined for 2015 and 2030 for the main economic variables. Subsequently, the most probable values of freight flow growth can be inferred through the Bayesian Networks for the O/D couple in 2015 and 2030. The scenarios for the economic variables are as follows:

- Base scenario: 2015 and 2030 forecast baseline (natural development of the market from the current situation);
- Optimistic scenario: GDP growth forecast increased by 30 %;
- Conservative scenario: GDP growth forecast decreased by 30 %.

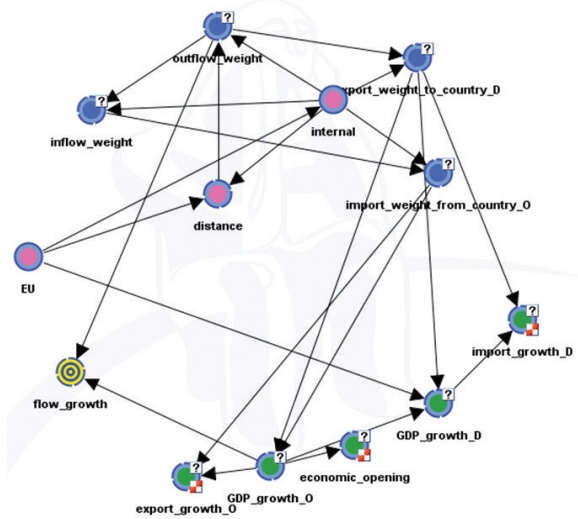


Figure 5.
The Bayesian Network model (whole ETIS O/D Matrix).

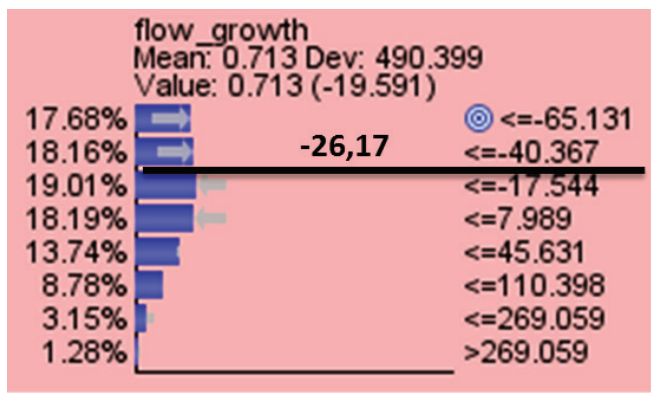


Figure 6.
Bayesian Network (whole ETIS O/D Matrix): evaluation of the mean flow prediction.

Table 5.
Demand forecast (whole etis o/d matrix) for 2015 and 2030 scenarios.

YEAR	Freight flows (road and rail) of the whole ETIS O/D Matrix	
2005	17.752 milions of tons	
2010	16.229 milions of tons	
2015	16.367 - 17.037 milions of tons	Delta 2010 - 15: 0 %: +5 %
2030	19.530 - 26.167 milions of tons	Delta 2010 - 30: +20 %: +61 %

2) The forecast for the Corridor 6 study area

A second Bayesian Network model was developed more specifically for the area concerned by Corridor 6. Flows are grouped as follows:

- Internal, with Origin AND Destination in Corridor zones;
- Exchanges, with Origin OR destination in Corridor zones;
- Transits, with Origin AND Destination outside of Corridor zones.

Once again, on 5-year and 20-year stationarity assumptions freight flows were inferred for 2015 and 2030, using the most probable values of flow relative growth. The forecasts for Corridor 6 flows (see Table 6 and Figure 7) show that the flow variation in 2015, relative to 2010 base year and considering the three scenarios, lies between -1 % (conservative scenario) and +10 % (optimistic scenario) with very low probability of having total flow decrease and high probability of having total flow increase, although small in quantity. The results of the demand forecast for 2030 show a general long-term increase of traffic flows with high percentage variation from the conservative scenario, with a 27 % of increase to a 96 % of increase for the optimistic one. It is very difficult to verify these results. We thus tried to compare our results with those produced by a recent work by the French Ministry of the Environment (Ministère de l'Écologie, du Développement durable et de l'Énergie, Direction des Transports Terrestres-Bureau d'Informations et de Previsione Economiques, 2005). This is one of the few comparable works to ours, in terms of geographical extension of the study area.

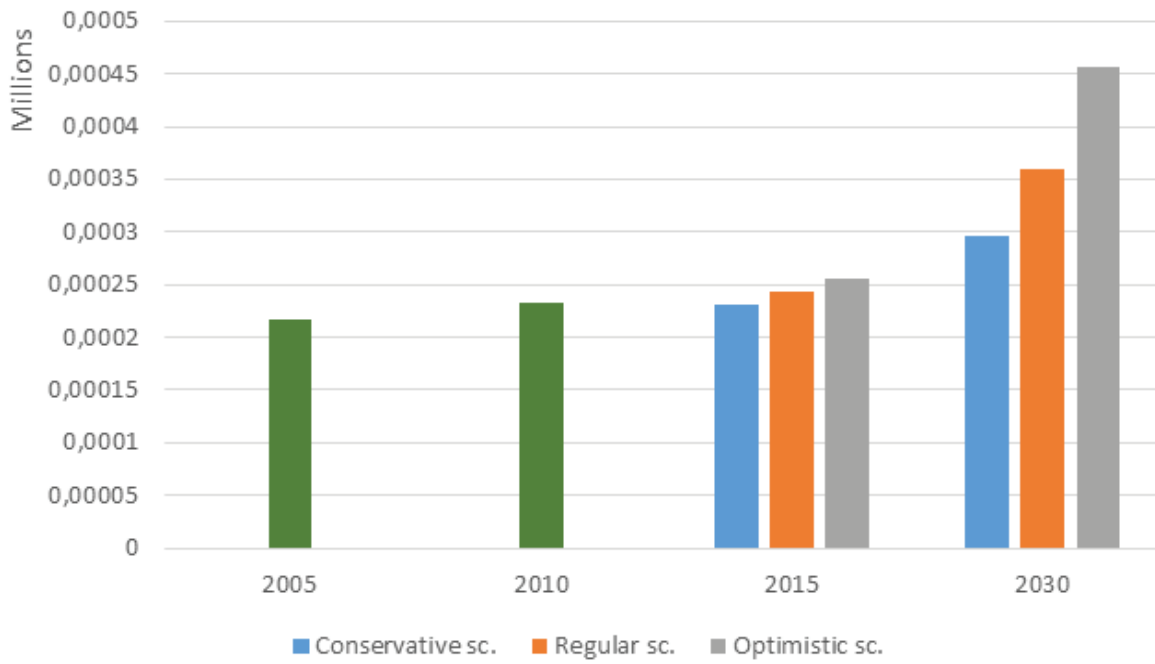


Figure 7.
Road and Rail flows in the Corridor 6 catchment Area (including transit).

Table 6.
Evolution of freight flows concerning corridor 6 catchment area (including transit).

YEAR	Freight flows (road and rail) of interesting O/D couples	
2005	217 milions of tons	
2010	233 milions of tons	Delta 2005 - 2010: +7,3 %
2015	230 - 256 milions of tons	Delta 2010 - 15: -1 %: +10 %
2030	297 -457 milions of tons	Delta 2010 - 30: +27 %: +96 %

Table 7.
Comparison of evolution of freight flows through pyrenees between the french study and the etis real values (values in million tons).

ETIS 2005	Study on freight flows through Pyreneer' - 2010 estimates		ETIS 2010
	Low scenario	High scenario	
10,86	12,5	13,5	11,68

The study on freight flows through the Pyrenees predicts the following annual average freight flow growth rates between the Iberian Peninsula and the rest of Europe between two scenarios: 2.9 % (low scenario) and 4.5 % (high scenario).

By applying these growth rates to the observed 2005 freight flows within the area of interest for Corridor 6 (data derived from the 2005 ETIS O/D matrix), the estimated 2010 road and rail freight flows from the Iberian Peninsula to the rest of the catchment area of Corridor 6 would be much higher than the ones actually recorded within the ETIS 2010 O/D Matrix (Table 7).

Table 8 provides a comparison between 2015 and 2030 forecasted freight flows in the two studies (the study of freight flows through the Pyrenees provides estimates for 2025, but due to the hypothesized linearity of the evolution, it was possible to determine the "most probable forecast" for 2030).

Table 8.
Comparison of prevision of freight flows through the pyrenees between the french study and our results (values in million tons).

YEAR	Freight Transport of Pyreneer' study - Estimates 2010		Our study on RFC 6		
	Low scenario	High scenario	Conservative	Regular	Optimistic
2015	14,8	16,9	11,5	11,9	13,6
2030	22,7	32,6	16,3	17,3	24,4

3. CONCLUSIONS AND FUTURE DEVELOPMENTS

The data-driven methodology applied in this work seems to be very promising from many points of view. Firstly, the data it needs are easy to find from official European level sources (even if more complete economic data-bases at the NUTS 2 level could have improved the performance of our models). Secondly, because of its simplicity, the methodology is applicable in the short term through model updating by incremental learning or new model development; it will thus be possible to update forecasts as new data are available and to follow multi-temporal economic dynamics. Moreover, the Bayesian Network framework adopted allows the recognition of different flow evolutions (which is similar to having multiple transport distribution system equations with different calibrated parameters) and their application in the forecasted scenarios. In addition, a comparison of the results with some official studies shows that our results are acceptable estimates.

The starting database for this first application covers two base years, 2005 and 2010, which are a very particular period for the European economy (arrival of new member states in 2004 and deep economic crisis after 2008), with some peculiar correlations and dynamics among economic, transportation and social variables. Availability of the 2015 version of the ETIS database will allow data-driven model development over the 2005-2015 period, which should produce more reliable results. Of course, the development of new infrastructures or geo-economic dynamics (entrance of new member states in the EU) will always be exogenous to the model and the use of time or cost-distances could be used instead of km-distance to better model the impact of transportation networks on the study area. Finally, the stationarity hypotheses on the links between economic, geographic and transportation variables are much more appropriate for a short-term forecast (5 years) than for long-term ones (20-30 years).

A further point to be developed is the link between the total demand forecast and the following modal split scenarios. The use of average prediction values necessary for this further methodological step involves the loss of the richness of the Bayesian Network results, i.e. the probability distribution of

the estimated flow demand. We are presently trying to use Montecarlo simulation approaches (Train, 2009) in order to extract a large number of possible deterministic demand values from the demand probability distribution. Subsequently, a modal choice probabilistic distribution will be derived from each of these values. It will then be possible to estimate an overall probability distribution for flows by mode and the results will be expressed in terms of values accompanied by statistical parameters such as mean, variance and quartiles. The methodology is similar to that used in Mixed Discrete Choice Models.

Another option would be to develop the entire demand forecast, i.e. the generation and the modal distribution of freight flows, within the Bayesian Network framework. It will then be possible to preserve a consistent probabilistic approach for flow estimation by transport mode.

REFERENCES

Albert, A. and Schafer, A., (2013), Demand for Freight Transportation in the U.S.; A High- Level View, Journal of Transportation Statistics pp. 103-116, available at: http://ageconsearch.umn.edu/bitstream/206946/2/3 %20DEMAND %20FOR %20 FREIGHT %20TRANSPORTATION %20IN %20THE %20U.S. %20A %20HIGH-LEVEL %20VIEW.pdf

Arnone, M., Inaudi, D. and de Jong, G., (2011), Un modello matematico per la valutazione degli scenari di sviluppo del sistema del trasporto merci nel Nord-Ovest, in Proc. XXXII Italian Conference of Regional Science, Torino, Italy, September 15-17.

Ben-Akiva, M. and Lerman, S. R., (1985), Discrete Choice Analysis, Boston: MIT Press.

Ben-Akiva, M., Meersman H. and Van de Voorde E., (2013), Freight Transport Modelling, Bingley: Emerald Group Publishin Limited.

Burgess, A., Chen, T. M., Snelder, M., Schneekloth, N., Korzhenevych, A., Szimba, E., Kraft, M., Krail, M., Nielsen, O., Hansen, C., Martino, A., Fiorello, D. and Christidis, P., (2008), TRANS-TOOLS (TOOLS for TRansport Forecasting ANd Scenario testing), Deliverable 6: Final Report, Deft, Netherlands: TNO Inro.

Caplice, C. and Phadnis, S., (2010), Driving Forces Influencing Future Freight Flows-NCHRP, web-only document 195, Washington D.C.: Transport Research Board.

Chase, K. M., Anater, P. and Pelan, T., (2013), Freight Demand Modelling and Data Improvement - The Second Strategic Highway Research Program, Washington D.C.: Transport Research Board.

Chen, M., (2011), ETIS and TRANS-TOOLS v1 Freight demand, in CTS-seminar – European and National Freight demand models, 1 March 2011, Stockolm.

Danielis, R. and Rotaris, L., (1999), An Analysis of Freight Transport Demand Using Stated Preference Data: a Survey and a Research Project for the Friuli-Venezia Giulia Region, Transporti Europei, 1999(13), pp. 30-38.

Floreano, D. and Mattiussi C., (1996), Manuale Sulle Reti Neurali, Bologna: Mulino Editrice.

Fusco, G., (2010), Handling Uncertainty in Interaction Modelling in GIS: How will an Urban Network Evolve?, in: Prade, H., Jeansoulin, R., Papini, O. and Schockaert, S. (eds), Methods for Handling Imperfect Spatial Information, pp. 357-378., Berlin: Springer.

Fusco, G., (2012), Démarche géo-prospective et modélisation causale probabiliste, Cybergeo: European Journal of Geography, Systems, Modelling, Geostatistics, document 613., http:/dx.doi.org/10.4000/cybergeo.25423

Jensen, F. V., (2001), Bayesian Networks and Decision Graphs, New York: Springer.

Meersman, H. and Van de Voorde, E., (2013), The Relationship between Economic Activity and Freight Transport, in: Ben-Aktiva, M., Meersman, H. and Van de Voorde, E. (eds), Freight Transport Modelling, Bingley: Emerald Group Publishing., http://dx.doi.org/10.1108/9781781902868-002

Ministère de l'Écologie, du Développement durable et de l'Énergie, Direction des Transports Terrestres-Bureau d'Informations et de Previsione Economiques, (2005), Analyse et évolution des flux de transport de marchandises à travers les Pyrénées, Issy Edition BIPE.

National Cooperative Freight Research Program, (2010), Freight-Demand Modeling to Support Public-Sector Decision Making, Washington D.C.: Transport Research Board.

NEA Transport research and training BV, (2005), Core Database Development for the European Transport Policy Information System (ETIS), Final Technical Report v1, NEA Transport research and training BV, available at: http://www.transport-research.info/sites/default/files/project/documents/20100120_125530_51920_ETIS %20 BASE %20- %20Final %20Technical %20Report.pdf

Onsel, S., Ulengin, F., Kabak, O. and Ozaydin, O., (2013), Transport Demand Projections: A Bayesian Network Approach, 13th World Conferenze on Transport Research, Rio de Janeiro, Brazil, July 15-18, available at: http://www.wctrs-society.com/wp/wp-content/uploads/abstracts/rio/selected/812.pdf

Pearl, J., (2000), Causality – Models, Reasoning and Inference, Cambridge: Cambridge University Press.

Petri M., Fusco, G. and Pratelli, A., (2014), A New Data-driven Approach to Forecast Freight Transport Demand, Berlin-Heidelberg: Springer-Verlag., http://dx.doi.org/10.1007/978-3-319-09147-1_29

Tavasszy, L., (2006), Goederenvervoer: Verre Vriend én Goede Buur!, Nijemegen: Radboud Universiteit.

Train, K. E., (2009), Discrete Choice Methods with Simulation, 2nd edition, Cambridge: Cambridge University Press.

Witten, I. H. and Frank, E., (2000), WEKA – Machine Learning Algorithms in Java, University of Waikato, Morgan Kaufmann Publishers.