Road Safety Data, Collection, Transfer and Analysis

Deliverable 4.7.

Forecasting Road Traffic Fatalities in European Countries: Towards an integrated European model

Please refer to this report as follows: Lassarre, S. ,Dupont, E., & Antoniou, C. (Eds.) 2012. Forecasting road traffic fatalities in European countries. Deliverable 4.7 of the EC FP7 project DaCoTA.

Grant agreement No TREN / FP7 / TR / 233659 /"DaCoTA"
Theme: Sustainable Surface Transport: Collaborative project
Project Coordinator:
Professor Pete Thomas, Transport Safety Research Centre, Loughborough Design School, Loughborough University, Ashby Road, Loughborough, LE11 3TU, UK

Project Start date: 01/01/2010 **Duration 30 months**

Organisation name of lead contractor for this deliverable: Belgian Road Safety Institute (IBSR)			
Report Author(s): Antoniou, C.; Papadimitriou, E.; Yannis, G. (NTUA); Bijleveld, F; Commandeur, J. (SWOV); Broughton, J; Dupont, E.; Martensen, H. (IBSR), Giustianni, G.; Shingo, D. (CTL); Hermans, E. (Uhasselt), Lassarre, S. (INRETS), Perez, C. (ASPB)			
Due date of deliverable	**31/12/2012**	**Submission date:**	**21/02/2013**

Project co-funded by the European Commission within the Seventh Framework Programme	
Dissemination Level (delete as appropriate)	
PU	**Public**

Directorate-General for Mobility and Transport

Project co-financed by the European Commission Directorate General for Mobility and Transport

TABLE OF CONTENTS

1 INTRODUCTION

The main objective of the analysis work performed in the framework of the Work Package 4 of the DaCoTA project is to analyse the past evolution of the annual number of fatalities in the various member states, and to forecast this evolution up to 2020. The model applied for many countries was the Latent Risk Model, which defines the development of the annual numbers of fatalities as the result of the joint development of exposure and risk (see Bijleveld et al., 2008; Martensen & Dupont, 2010). Because it involves the simultaneous modelling of the risk and exposure trends, the LRT is a multivariate (bivariate) Time Series model.

In this report, demonstration is made of the usefulness and appropriateness of multivariate time series models to further improve the analysis of the developments of annual fatality numbers by exploring models (1) integrating the fatality time series of a *panel of countries*, (2) handling the trends of fatality numbers for subgroups of road-users and for various accident types simultaneously, and (3) including economic variables such as GDP.

The simultaneous modelling of multiple trends implies that attention is paid to possible correlations between the random variations of the trend components (level, slope among others). When these correlations are high, the components in question are said to be *common* for the different trends modelled. The identification of common components is important because it allows the improvement of models to make them more efficient, but also because common components are informative in themselves of the dynamics governing the evolution of the trends considered.

In the next section, the notion of “correlation between the random variations of the components of different trends” is formally related to concepts that are fundamental in Time Series analysis, such as the concepts of stationarity, integration and co-integration. This is described on the basis of the results of the investigation of the correlations between fatality and exposure time series that has been conducted previously for the different member states.

A first multivariate Time Series application is then presented and discussed, namely, the simultaneous analysis of the development of annual fatality series for the various member states – or subgroups of member states by means of macro panel data analysis. After having exposed the principles underlying the technique, an example application is presented for the development of fatality numbers in France and in the United Kingdom.

The next section describes how the multivariate time series framework can be applied to fit a disaggregated model of the fatality trends for subgroups of road users (defined on the basis of age, gender, transport mode and others). This allows identifying subgroups of road users for which the evolution of fatality numbers is governed by common processes or components, and also subgroups for which this evolution appears problematic (or not as encouraging as that of others). An example is then provided: the model of the evolution of the number of fatalities for 6 different age groups in Spain, taking into account the evolution of the size of the population. A second application explores the relationship between the number of fatalities and GDP on a macropanel of 30 countries and shows how to articulate short-term and long-term variations between them in a coherent time series model.

2 LESSONS LEARNED FROM UNIVARIATE AND BIVARIATE MODELS

2.1 Stationarity and order of integration

Stationarity is a fundamental assumption in some types of time series models (e.g., ARMA models). A time series is called stationary if the process generating the time series fluctuates around a constant value (the underlying mean), independent of time, and if the variance of the fluctuation remains essentially constant over time. When the stationarity assumption is met, the mean or variance of the series can be considered meaningful sample statistics, and can consequently be used to predict future behaviour of the process under investigation. Stationarity is, however, no natural property of most natural time series (e.g. economics), and those that are used in the road safety field are no exception (especially when the number of years of observation is large, e.g, >20). The number of fatalities, for example, has been continuously declining in most EU member states, while the number of vehicle-kilometres has been continuously increasing (see Dupont & Martensen, 2012). Thus, the means of the two series cannot be considered constant over time. Previous analyses performed in the Work Package have also shown that, for most countries, at least one component of the fatality trend (the level most often, but the slope as well in some cases) significantly varies over time.

In many cases, a series can be made stationary by means of *differencing*: the variable is expressed as being a function of its own value at previous time points. The *order of integration* of a series summarises the number of differences that is necessary for the series to become stationary and is denoted *I*(d). A series that is already stationary is denoted *I*(0). In the other cases, a univariate time series Y_t is said to be integrated of order *d*, denoted by *I*(*d*), if it needs to be differenced *d* times to make it stationary.

If a time series is *I*(1), then it can be analyzed with a local level model (LLM: the level of the series is defined as random, there may or may not be a *deterministic slope* in the model). It can indeed be shown that such a series needs to be differenced only once in order to make it stationary. If a time series is *I*(2), then it must be analyzed with a local linear trend model (LLTM: stochastic level and slope) or a smooth trend model (STM: fixed level and stochastic slope), as it can be shown that such a series needs to be differenced twice in order to make it stationary, see Commandeur and Koopman (2007, p.132-133) for details.

For the European countries analysed, most of the fatalities time-series are integrated of order 2 [i.e. I(2)], and modelled by a local linear trend model; same for exposure time series, when available.

2.2 Cointegration between exposure and fatalities

One of the main objectives of this Work Package was to apply a bivariate model to the developments of annual fatality numbers in the different member states, *taking the development of exposure into account*. As we use multivariate structural models with

unobserved components of the trends, such as the level and the slope, which are random walks and so integrated of order 1 I(1), the main point of the modelling is to assess the importance of the correlation between the random disturbances of the levels and the slopes of the trends of both time series.

If μ_t is a level we have

$\mu_t - \mu_{t-1} = \xi_t$

and ξ_t is the random disturbance which follows a normal distribution with mean 0 and variance σ^2.

We have seen that the disturbances of the unobserved level components of the fatalities and mobility can be correlated and that the disturbances of the unobserved slopes components of the fatalities and mobility can be correlated as well.

When such a correlation happens to be (close to) plus or minus one, then we have the special situation that the level and/or slope components are -as is said- *common*. When unobserved components are common this means that the changes or “shocks” driving the dynamics of the (two or multiple) time series are perfectly linearly related. Stated differently, the level and/or slope components then change in the same way at the same points in time, and the corresponding time series therefore display common behaviour.

There is an intimate relation between common factor models and what is known in the time series literature as *cointegration*. First, we will provide a short introduction into the concept of cointegration.

In that case we cannot explore relationship between both time-series by means of a classical linear regression, because such a regression is valid only between stationary time-series. The way to deal with such a problem is to explore the cointegration between time-series.

If two series Y_{1t} and Y_{2t} are both $I(d)$, then any linear combination ($Y_{1t} - \alpha Y_{2t}$) of the two series will usually be $I(d)$ as well. However, if two series Y_{1t} and Y_{2t} are both integrated of order d, and a linear combination of the two series, ($Y_{1t} - \beta Y_{2t}$) say, exists for which the order of integration is less than or equal to d, say (d-b), then the two series are said to be *cointegrated* of order (d,b), which is denoted by CI (d,b). CI(2,2) means that there exists a linear relationship of two integrated time-series of order 2 which is stationary I(0). We could extend this definition when a linear combination involving a linear deterministic trend ($Y_{1t} - \beta Y_{2t}$-$(a+ct)$) is trend stationary $I(0)$ (Hendry and Juselius, 2000).

As often fatalities and exposure time-series are I(2), we are looking for a cointegration of order 2 with a trend stationarity, because it gives a long-term linear relationship between the logarithms of the number of fatalities and exposure such that:

$\text{Log FAT}_t = \beta \text{LogEXP}_t + a + ct + \varepsilon_t$

This may sound pretty obscure, but translated in structural time series terminology all this is actually quite familiar. First of all, consider two time series both integrated of order 1 (see Table 1), meaning that both series can be modelled with a local level model or a local level with fixed slope model. If the two series are $CI(1,1)$, then this is the same as saying that a bivariate local level model or a bivariate local level with fixed slope model applied to both

series would reveal that the correlation between the level disturbances of the two series is (close to) plus or minus one, which implies that the two series have *common levels*.

		Y_{2t}	
		I(1) = LLM	*I*(2) = LLTM or STM
Y_{1t}	*I*(1) = LLM	*CI*(1,1) = *I*(0)	
	I(2) = LLTM or STM		*CI(2,1) = I(1)* *CI(2,2) = I(0)*

Table 1: Types of cointegration for two time series integrated of order 1and 2

Next, consider two time series both integrated of order 2 (Table 1), meaning that both series can be modelled with a local linear trend or a smooth trend model. There are now two possibilities: If the two series are *CI*(2,1), then this is the same as saying that a bivariate local linear trend model or a bivariate smooth trend model or a bivariate local linear and smooth trend model applied to both series reveals that the correlation between the slope disturbances of the two series is (close to) plus or minus one, implying that the two series have *common slopes*.

If the two series are *CI*(2,2), on the other hand, then this is the same as saying that a bivariate local linear trend model applied to both series reveals that the correlation between the slope disturbances of the two series is (close to) plus or minus one, *and* that the correlation between the two level disturbances is (close to) plus or minus one, implying that the two series have *common trends* (i.e., both common slopes and common levels).

2.3 LRT as a bivariate common factor model

An exploration of the correlation between the slope disturbances of the exposure and fatality series in the different member states has been carried out by means of SUTSE (Seemingly Unrelated Time Series Equations) models. Details of the results can be found in Deliverable 4.4 of the DaCoTA project. SUTSE models are bivariate structural models with a complete variance-covariance structure of the disturbances of the slopes, levels, and irregulars of both of the logarithms of the annual numbers of fatalities and vehicle*kilometers:

$$\log TrafficVolume_t = Level(\log TrafficVolume)_t + \varepsilon_t^{\,e}$$

$$Level(\log TrafficVolume)_t = Level(\log TrafficVolume)_{t-1} + Slope(\log TrafficVolume)_{t-1} + \xi_t^{\,e}$$

$$Slope(\log TrafficVolume)_t = Slope(\log TrafficVolume)_{t-1} + \zeta_t^{\,e}$$

$$\log Fatalities_t = Level(\log Fat)_t + \varepsilon_t^{\,f}$$

$$Level(\log Fatalities_t) = Level(\log Fatalities)_{t-1} + Slope(\log Fatalities)_{t-1} + \xi_t^{\,f}$$

$$Slope(\log(Fatalities)_t = Slope(\log(Fatalities)_{t-1} + \zeta_t^{\,f}$$

Of course, this model was only applicable to the extent that exposure data were available for the different countries analysed.

The analyses conducted revealed that, when exposure data are available, there is a correlation between the stochastic slopes of the trends of both time-series, which means that they have a common slope or that they are cointegrated. Secondly, the exposure trend most often appeared to be a smooth trend, that is, a trend with a fixed level and a stochastic slope. The number of fatalities, on the other hand, had a fixed level in most instances. In that case, there is no possibility of a common level. The correlations between the slope disturbances of the two series were used to define the model in the next step of the analyses, as explained below.

In a next step, the *risk* (the number of fatalities per billion vehicle kilometres) was introduced in the model as an unobserved trend with a level and a slope. The reason is that risk is the major indicator of safety performance. This model is consequently referred to as to the Latent Risk Trend model (LRT), see Bijleveld et al. (2008). It is a multivariate model which can have as common factor the trend of exposure

$$\log TrafficVolume_t = Level(\log TrafficVolume)_t + \varepsilon_t^{\ e}$$

$$Level(\log TrafficVolume)_t = Level(\log TrafficVolume)_{t-1} + Slope(\log TrafficVolume)_{t-1} + \xi_t^{\ e}$$

$$Slope(\log TrafficVolume)_t = Slope(\log TrafficVolume)_{t-1} + \zeta_t^{\ e}$$

$$\log(Fatalities_t) = Level(\log TrafficVolume)_t + Level(\log Risk)_t + \varepsilon_t^{\ f}$$

$$Level(\log Risk)_t = Level(\log Risk)_{t-1} + Slope(\log Risk)_{t-1} + \xi_t^{\ r}$$

$$Slope(\log Risk)_t = Slope(\log Risk)_{t-1} + \zeta_t^{\ r}$$

The LRT model is chosen related to the estimated value of the correlation between the slope disturbances.

When the correlation is null, there is no relationship between fatalities and exposure. Knowing exposure does not bring any additional information to predict the number of fatalities. In that case, a univariate LLT model is appropriate. When the correlation between the slope disturbances is equal to 1 and the level components are deterministic, both time-series share the same stochastic slope and are trend stationary with a deterministic linear trend for the risk. By constraining the beta coefficient for exposure to 1, we get an LRT model. It remains to demonstrate that there is a class of equivalence of models by showing that the elasticity coefficient β is a linear function of the risk deterministic slope c.

When the correlation takes a medium value, then there is a weak correlation between the two time-series and an LRT provides a solution through the estimation of a relationship this time between risk and exposure.

The relationship between the observations made on the correlations between the exposure and fatality trends and the type of model adopted is detailed in Table 2.

Type of correlation between the slope disturbances	0: No correlation	1: Full Correlation	0.1 to 0.9: Medium correlation
Relationships	Independence between fatalities and exposure	Strong dependency: Cointegration	Weak dependency
Consequences	E(fatalities|exposure)= E(fatalities)	Common components (same stochastic slope) Long-term linear relationship Log FAT_t = βLevel(logEXP_t) + a +ct + ε_t	
Model	Univariate LLT	LRT with deterministic risk trend. By constraining $\beta = 1$	LRT with stochastic risk trend
Example	Greece	France	Slovenia

Table 2: Types of cointegration for two time series integrated of order 1 and 2

3 MULTIVARIATE ANALYSIS

In this chapter we discuss joint models of the evolution of the annual number of fatalities. Such models are available for the n=28 European countries. They are mainly LRT models relating the number of fatalities to the number of vehicles*kilometres through the risk equation. We have seen in the previous chapter that this is based on the observation that fatalities and exposure *within* are cointegrated many countries.

It could be of added value to consider a panel of several countries to estimate models. In this case we can investigate whether either fatalities or exposure are cointegrated between countries. There are three motives for the implementation of such a “macro panel” approach:

1.) Test the cointegration of different types of series (e.g. fatalities and exposure, or economic variables) on panel country data rather than on individual country data. The test is more powerful if we add a country dimension to the time dimension. Compare the structure of common factors between countries.

2.) Study the structure of common components (e.g. fatality slope and exposure slope) between countries. As an example, the dependence between the number of fatalities and the number of vehicles kilometres could take different forms according to groups of countries. Is direct proportionality the rule?

3.) Identify common developments. As mentioned in Koopman et al. (2007, p.93), in such models “ … some or all of the components are driven by disturbance vectors with less elements than the number of time series involved. The identification of common factors yields models which may not only have an interesting interpretation, but may also provide more efficient inferences and forecasts” (italics in the original text). This means there could be a number of prototypical developments (i.e. the factors) that play a role in many countries. Each country’s development can be seen as a combination of these factors, with factor weights determining how strongly the factor in question affects the development in that particular country.

3.1 Macro Panel data

As the country dimension is added, we could reduce the number of time points of the time series and work on a medium period of time such as a 20 years period (e.g. 1991-2010). This kind of data set is easier to provide and can still be considered a macro panel. Generally, we speak of a macro panel when we have few countries (the N dimension) each with many time units (20 or more). A micro panel, on the contrary, has a large number of units N (>100) and a small number time units (5 or less) and is therefore not suited to account for the complex time-related dynamics that usually govern time series. Note that, in the remainder of this section the subscript *i* will be used to refer to the country or N-dimension units and the subscript *t* will be used to refer to time units.

The statistical models are completely different for micro panel than macro panel data. With micro panel data, the time dimension is considered as a repeated measure with simple random structure and the N dimension is introduced through a random effect within a mixed general linear (or multilevel) model. It should be noted, however, that "The vast majority of

empirical research using 'macro panels' implements 'micro panels' methods!" (Eberhardt, 2011)

An important point is the nature of the time-series and in particular their order of integration. In our case, the time series of the number of fatalities[1] y_{it} are either integrated of order 2 I(2) (stochastic slope) or integrated of order 1 I(1) with a deterministic slope. Only for few countries (Germany, for example) the time series is stationary with a deterministic slope. The time series of the number of vehicle*kilometres x_{it} are rather integrated of order 2 (stochastic slope) and mostly smooth trend (stochastic slope and fixed level). It causes a problem of heterogeneity and some solution could be to differentiate the time series to get the same order 1.

3.2 Cointegration in panel and Common Factor models

Remember the various possible cases of Table 1 when considering cointegration between time series: In case of I(1), it means that a linear combination of I(1) time series is stationary I(0). In case of I(2), it means that a linear combination of I(2) time series is either I(1) (noted C(2,1)) or I(0) (noted C(2,2)). There could be also the possibility of multicointegration when a linear combination of time series and differentiated time series is I(0).

Besides, both intra- and inter-country cointegration deserve attention when working with macro-panel data. Intra-country cointegration means that the logarithms of the number of fatalities and the number of vehicle*kilometres *of one country* are cointegrated. There is a linear combination of both time series $logy_{it} - \beta_i logx_{it}$ which is rather stationary I(0) than I(1). The cointegration relationship could integrate a linear trend, which means that $logy_{it} - \beta_i logx_{it} -(a_i + b_i t)$ is I(0). This means that the combination of the two series is not stationary (i.e. does not have the same mean over the years), but follows a linear trend. We could be interested to test the homogeneity of the long-term relationship between fatalities and exposure and test if it is a risk type relationship ($\beta = 1$ for direct proportionality).

Inter-country cointegration means that there is either a cointegration of the logarithm of the number of fatalities $logy_{it} - \beta_{ij} logy_{jt}$ or of the logarithm of the number of vehicle*kilometres $logx_{it} - \beta_{ij} logx_{jt}$ across countries. The inter-country cointegration allows for dependence between variables (y or x) from different cross section units (countries). We could be interested to test the homogeneity of the long-term relationship between fatalities and exposure and test if it is a risk type relationship ($\beta = 1$ for direct proportionality).

Some interventions w_{it} could be introduced.

[1] After a logarithmic transformation. i varies from 1 to n=28.

3.2.1 Estimation and test in case of I(1)

A first method (residual-based approach or single equation) relies on the hypothesis of independence of the number of fatalities and of non cointegration between the exposures.

In case of I(1), the equations are

$logy_{it} = a_i + b_i\, t + \lambda_i\, w_{it} + \beta_i\, logx_{it} + u_{it}$

$logx_{it} = logx_{it-1} + \varepsilon_{it}$

With the variance-covariance matix of the residuals u and ε

$$\Omega_i = \begin{pmatrix} \sigma_{ui}^2 & \sigma_{u\varepsilon i} \\ \sigma_{u\varepsilon i} & \sigma_{\varepsilon i}^2 \end{pmatrix}$$

Here, the logarithms of fatalities $logy_{it}$ are expressed as a linear trend $a_i + b_i\, t$ plus a number of (weighted) interventions $\lambda_i\, w_{it,}$ plus the regression on the logarithm of exposure $\beta_i\, logx_{it,}$ plus the residuals $u_{it.}$ The index i indicates that each component is estimated for each country separately.

There could, however, be a serial correlation ρ_i between the u_{it} (short term dynamics).

Therefore, the second model takes into account the cross section dependence (cointegration between countries' fatalities) and provides system estimators:

$logy_{it} = a_i + b_i\, t + \lambda_i\, w_{it} + \beta_i\, logx_{it} + u_{it}$

$logx_{it} = logx_{it-1} + \varepsilon_{it}$

$u_{it} = f'_t\, \gamma_i + \xi_{it}$

With,

$$\Omega_i = \begin{pmatrix} \sigma_{\xi i}^2 & \sigma_{\epsilon\xi i} \\ \sigma_{\epsilon\xi i} & \sigma_{\varepsilon i}^2 \end{pmatrix}$$

This time, the residuals - u_{it} – are expressed as a vector of common factors f_t of size k<n. f_t could be I(1) (common trend) or I(0) (common levels). If I(1), it could be correlated to $logx_{it}$. In that case, $logy_{it}$, $logx_{it}$ and f_t are cointegrated together.

A more complete model could integrate the cointegrations between the exposures, if any. The covariance matrix of the $logx_{it}$ residuals is not diagonal any more, and could be of lower rank than n.

The estimation methods are based on weighted iterative regression of the differentiated variables Δlogy and Δlogx, with a determination of factors through a PCA or singular value decomposition on the residuals of the long-term equation. Another technique is to use an I(1) Vector Autoregressive (VAR) model.

The first task is to adapt the system of equations within a structural multivariate model (state-space) by specifying appropriate structure for the variance-covariance matrices of the components, which is feasible as there is only one regressor variable, namely the exposure.

3.2.2 Estimation and test in case of I(2) or mixture I(2) and I(1)

The main problem is to extend the model to I(2) or a mixture I(2) and I(1) time series. Solutions exist such as an I(2) VAR model or an I(2) - I(1) transformation. We are in the domain of multicointegration.

3.2.3 First attempt on two countries

Consider France and UK from 1953-2010. The model form considered is:

$$\begin{aligned}
\log y_{1t} &= a_1 + b_1 t + \beta_1 \log x_{1t} + \mu_{yt} + \varepsilon_{1t} \\
\log y_{2t} &= a_2 + b_2 t + \beta_2 \log x_{2t} + \boldsymbol{\theta_2} \mu_{yt} + \varepsilon_{2t} \\
\log x_{1t} &= \mu_{x1t} + \varepsilon'_{1t} \\
\log x_{2t} &= \mu_{x2t} + \varepsilon'_{2t}
\end{aligned}$$

Where the logarithms of fatalities are not only the sum of a country-specific linear trend and regression on the logarithm of exposure, but also have a component μ_{yt} that is common for both countries. The question is therefore: is it possible to estimate a model of this form for the fatalities in France and the UK?

At first, we considered fatalities only (i.e. without regression on exposure and linear trend) and found France and UK to have a common slope. Moreover, with a fixed level component for France and a stochastic level component for UK.

Once we introduce the regressors (exposures and linear deterministic trend), there is no common slope any more. The remaining trends for GB and France are random walks. But these levels are not correlated, so there is no common factor μ_{yt}, but instead two local level trends for each countries.

$$\begin{aligned}
\log y_{1t} &= a_1 + b_1 t + \beta_1 \log x_{1t} + \mu_{y1t} + \varepsilon_{1t} \\
\log y_{2t} &= a_2 + b_2 t + \beta_2 \log x_{2t} + \mu_{y2t} + \varepsilon_{2t}
\end{aligned}$$

Concerning the logarithms of exposures $\log x_t$, the slopes are not correlated and the level is stochastic only for G-B (smooth trend for France). There are no common factors. Each has its own way.

When we estimate the multivariate model with interventions (interventions not shown)

$$\begin{aligned}
\log y_{1t} &= a_1 + b_1 t + \beta_1 \mu_{x1t} + \mu_{y1t} + \varepsilon_{1t} \\
\log y_{2t} &= a_2 + b_2 t + \beta_2 \mu_{x2t} + \mu_{y2t} + \varepsilon_{2t} \\
\log x_{1t} &= \mu_{x1t} + \varepsilon'_{1t} \\
\log x_{2t} &= \mu_{x2t} + \varepsilon'_{2t}
\end{aligned}$$

we introduce common slopes by making the fatalities dependent of the exposures, meaning a dimension 2 instead of 4 for the covariance matrix of the slopes. We opt for a diagonal matrix for the levels' covariances (no common factor at that level). We get, for the slope disturbance factor loading matrix (with a slight correlation of 0,5 between both exposures):

	LVKMGB	LVKMFrance
LVKMGB	1.77	-0.28
LVKMFrance	0.011	1.05

which are the beta's, and:

	LVKMGB:	LVKMFrance:	LVKMGB:	LVKMFrance
Constant	-0.06653	-0.05089	0.0000	0.0000

which are the b1 and b2.

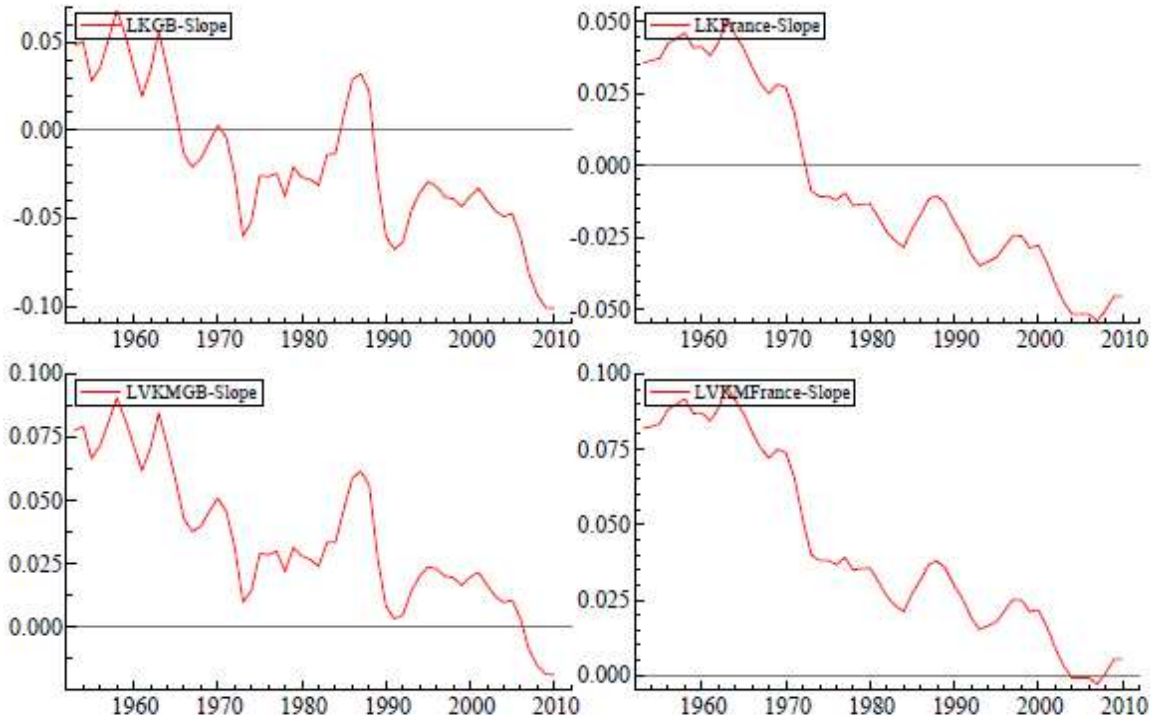

Figure 1: Slopes of fatalites (LK) and exposure (LVKM)for Great-Brtitain and France.

The slopes of fatalities are equal to: constant +betaF*slopeVKMF+betaGB*slopeVKMGB. For France, this is equivalent in percent to -5,09+1,07*slopeVKMK+0,01*slopeVKMGB. In France the number of fatalities is directly proportional to the exposure (pure risk model). The risk decreases by -5,1 % per year.

Both profiles are different, even if they are decreasing. In 2010, the slopes of the number of fatalities are equal to -10% for GB and -4,5% for France.

Finally, if we restrict some factor loadings and the correlation between exposure slopes to 0, we obtain the required form on country exposures, with no common factor f (as we have no common level between the μ_{yt}s)

$$\log y_{1t} = a_1 - 0{,}067t + 1{,}44\,\mu_{x1t} + \mu_{y1t} + \varepsilon_{1t}$$
$$\log y_{2t} = a_2 - 0{,}05t + 1{,}05\,\mu_{x2t} + \mu_{y2t} + \varepsilon_{2t}$$
$$\log x_{1t} = \mu_{x1t} + \varepsilon'_{1t}$$
$$\log x_{2t} = \mu_{x2t} + \varepsilon'_{2t}$$

We could go on and estimate such models with STAMP by adding other countries if their size is not too big.

3.2.4 Discussion:

The case of France and Great-Britain give us an opportunity to test the formulation of the macropanel model by means of structural models with trend components and apply such kind of modelling to assess the possibility of common factors. Unfortunately in this case, there are no common factors, because the exposures are not similar or correlated; neither are the risk levels. Nevertheless we have a tool to proceed for further multivariate analysis of the evolution of risks in European countries.

4 TWO SPECIFIC APPLICATIONS

4.1 Disaggregation

4.1.1 Risk and exposure within and between classes of road users

Road-users are not exposed to an abstract risk on the road. Road users get killed in traffic, because of collisions of moving vehicles with objects along the road, with pedestrians or with other moving (or stopped) vehicles. How can we model such interactions between classes of road-users? Road-users can be classified by of age (sex), mode of transport (car, powered 2-wheelers, etc.), and by type of collisions.

Risk models definition:

In case of single vehicle accident, the basic equation works

$$\text{fatalities} = \text{risk} * \text{exposure}$$

and can be applied to any road-user class.

In case of pedestrian accident, the collision occurs between a pedestrian and a moving vehicle. The basic equation depends on the environment composed by moving vehicles

$$\text{fatalities(pedestrian)} = \text{risk(vehicle)} * \text{exposure(pedestrian)}$$

Risk(vehicle) on a particular network is a function of the flow of vehicles (if necessary classified by mode), which is characterized by a volume, a density and a speed, knowing that volume=speed * density and speed = g(density). Exposure of vehicles is equal to the product of the volume by the length of the network. Usually, the number of fatalities is not directly proportional to the volume, rather to a power α less than 1

$$\text{fatalities(pedestrian)} = \text{risk} * \text{exposure(vehicle)}^{\alpha} * \text{exposure(pedestrian)}$$

In case of two-vehicle collision, the basic equation is usually transformed to take into account the interaction by a product of exposures

$$\text{fatalities} = \text{risk} * \text{exposure(vehicle)} * \text{exposure(vehicle)}$$

Is it justified?

In fact, if we distinguish vehicle 1 and 2, we have, by taking one vehicle as a moving obstacle:

$$\text{fatalities(vehicle1)} = \text{risk(vehicle2)} * \text{exposure(vehicle1)}$$

$$\text{fatalities(vehicle2)} = \text{risk(vehicle1)} * \text{exposure(vehicle2)}$$

It depends on the form of the risk function. Suppose a dependence like in the pedestrian accident risk

$$\text{fatalities(vehicle1)} = \text{risk12} * \text{exposure(vehicle2)}^{\alpha} * \text{exposure(vehicle1)}$$

$$\text{fatalities(vehicle2)} = \text{risk21} * \text{exposure(vehicle1)}^{\alpha} * \text{exposure(vehicle2)}$$

If we sum, we get:

$$\text{fatalities} = r12 * e2^{\alpha} * e1 + r21 * e1^{\alpha} * e2$$

If $\alpha = 1$ and r12 = p1 and r21 = p2, we get Koornstra's model (1973):

$$\text{fatalities} = (p1 + p2)\ e1 * e2$$

which becomes the basic interaction model if p1 = p2= r12 = r21= risk.

Estimation

Let us take three classes (young adult, adult, senior) and focus on single-vehicle accidents and accidents between two vehicles.

On cross-section data, we start from a special symmetric contingency table counting the number of fatalities. 0 is a phantom vehicle implicated in single collision, 1 is young adult, 2 is adult and 3 is senior, d11 is the number of fatalities in collisions between young adults, d12 is the number of fatalities in vehicles driven by young drivers in collisions with adult drivers, and so on. S1 is the number of fatalities in single accidents (against phantom vehicles) and so on.

	1	2	3	0
1	d11			
2	d12	d22		
3	d13	d23	d33	
0	s1	s2	s3	0
Total	n1	n2	n3	

Table 3: Form of the contingency table.

In Koornstra's model, one is free to use the single accident row with six unknown parameters called proneness p and exposure e by means of a Poisson model, or to ignore it. The exposure can be dependent on some values E like the population size for example.

On longitudinal data, we start from the marginal counts of the total number of fatalities per class. The complete model can be written as:

$$n1 = (r1 + r11 * e1^{\alpha} + r12 * e2^{\alpha} + r13 * e3^{\alpha}) * e1$$

$$n2 = (r2 + r21 * e1^{\alpha} + r22 * e2^{\alpha} + r23 * e3^{\alpha}) * e2$$

$$n3 = (r3 + r31 * e1^{\alpha} + r32 * e2^{\alpha} + r33 * e3^{\alpha}) * e3$$

The first consequence of this formulation is that the exposure for a group is not strictly proportional, but rather with an exponent greater than 1 (r1 e1 + r11 * $e1^{1+\alpha}$ # r'1 $e1^{1+\beta}$).

As the ni's depend on the ei's, the reciprocal is true. It means that there is a dependency between the ni's. So

$$n1 = (r'1\ e1^{1+\beta}) * f1(n1,n2,n3)$$

$$n2 = (r'2\ e2^{1+\beta}) * f2(n1,n2,n3)$$

$$n3 = (r'3\ e3^{1+\beta}) * f3(n1,n2,n3)$$

if we suppose a multiplicative form for the fi's. Taking the logarithm, we end with the simultaneous equation model

$$\log(n1) = r''1 + (1+\beta)\log(e1) + \gamma 1 \log(n2) + \gamma' 1 \log(n3)$$

$$\log(n2) = r''2 + (1+\beta)\log(e2) + \gamma 2 \log(n1) + \gamma' 2 \log(n3)$$

$$\log(n3) = r''1 + (1+\beta)\log(e3) + \gamma 3 \log(n1) + \gamma' 3 \log(n2)$$

This set of equations justifies the use of a multivariate structural model for the reduced form of the number of fatalities according to 6 classes of age as presented in the note (Lassarre, 2012). And we expect a correlation between the unobserved components and the existence of common factors because of these dependencies through the exposures, which are too complicated to model directly, but could be subsumed via some common factors.

4.1.2 Disaggregation by age group for Spain

We consider six age intervals: 0-14, 15-17, 18-24, 25-49, 50-64, 65+ over a period of 20 years from 1991 to 2010. We estimate a multivariate structural model with 6 equations relating, for each age interval, the logarithm of the number of fatalities to the logarithm of the size of the population plus a stochastic trend.

The univariate models are smooth trend models with fixed level and stochastic slope. There are common slopes, maybe 2 or 3. One common slope is not sufficient. With 3 common slopes, the model becomes unstable. With 2 common slopes, the log-likelihood, equal to 230, is minimized. We keep as common factors the two intervals 25-49 and 65+. The trend for the other classes is a linear combination of these two.

	25-49	65+
0-14	0,27	0,66
15-17	2,40	-1,25
18-24	1,27	0,11
50-64	0,46	0,65

Table 4: Slope disturbances factor loading matrix.

The trend for the 18-25 is similar to the one of the 25-49 age group. The trend for the 50-64 depends more on the 65+ trend than on the 25-49 trend.

The dependence on the size of the population varies according to the age intervals. It is significant only for the 15-17 and 65-+ with an elasticity of 1,9 and 2,9, respectively.

	coefficient	t-value
0-14	-0,17	-0,49
15-17	1,86*	2,46
18-24	-0,64	-1,31
25-49	1,24	1,82
50-64	-0,88	-1,63
65+	2,86*	2,77

Table 5: Regression coefficients and t-values on population size for the different age intervals (significant coefficients are denoted by a star).

All the trends are decreasing, because the majority of the slopes have negative values, with two leading patterns given by the 25-49 and 65+ age slopes (Figure 2). The risk has increased in the recent years for the elderly and for children. It has decreased, however, for the 15-17.

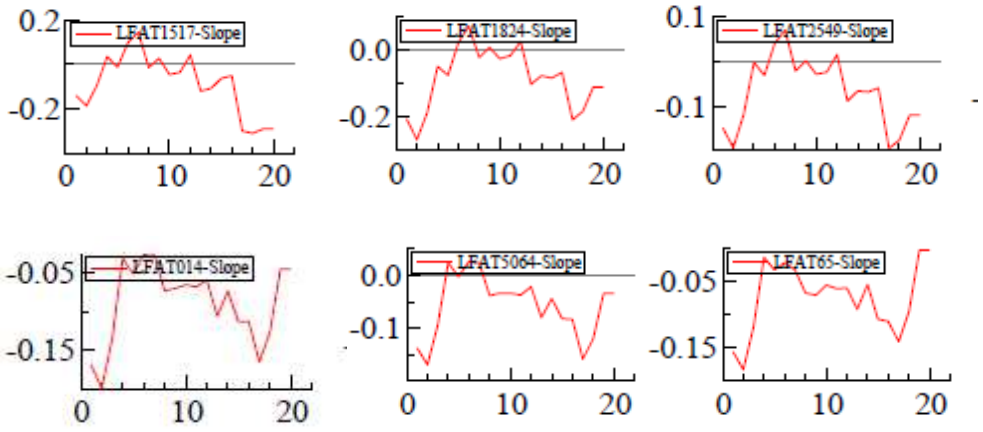

Figure 2: Risk slopes for the different age classes

The population size is decreasing for young people between 15-17 and 18-25, increasing for adults and the elderly and follows a U shape for children between 0-14 (Figure 3). Only the

15-17 and 65+ regressions are positively sensitive to the evolution of their population size (Table 5). Deterministic trends could be considered for these two age classes: downwards for the 15-17 and upwards for the elderly.

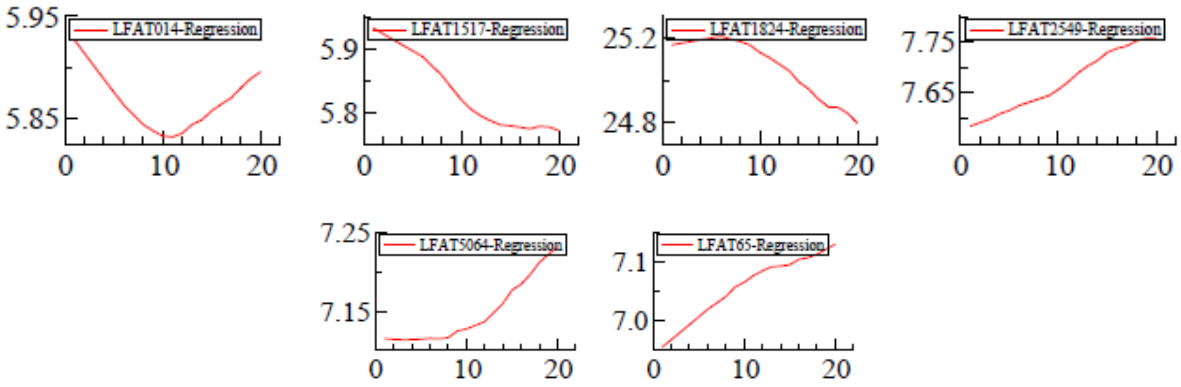

Figure 3 : Population sizes (log scale).

The mortality trends have three patterns: regular decrease for 0-14 and 65+, decrease with a plateau for 18-24, 25-49 and 50-64, plateau ending with a sharp decrease for 15-17.

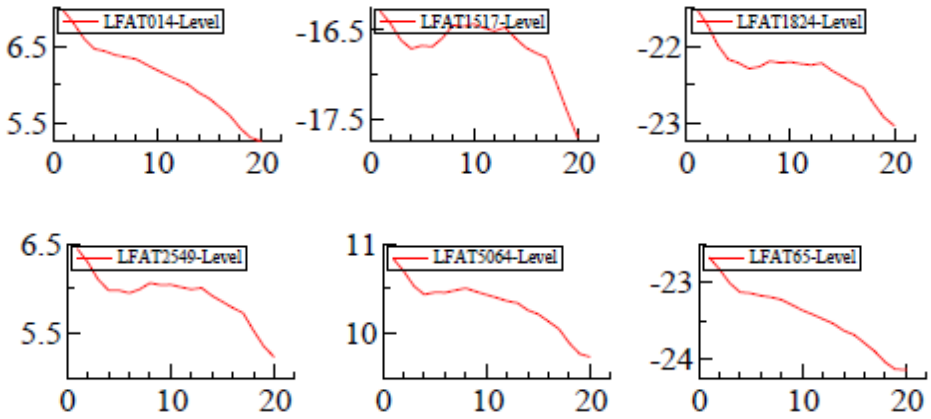

Figure 4: Levels of the mortality trends for the different age categories (log scale).

At the end, the composition of the regression effect of the population size (a kind of deterministic trend) and of the trend due to the stochastic slope leads to the development of the number of fatalities. Some evolutions are more pronounced than others: weak for elderly, strong for the 15-17 (Figure 5).

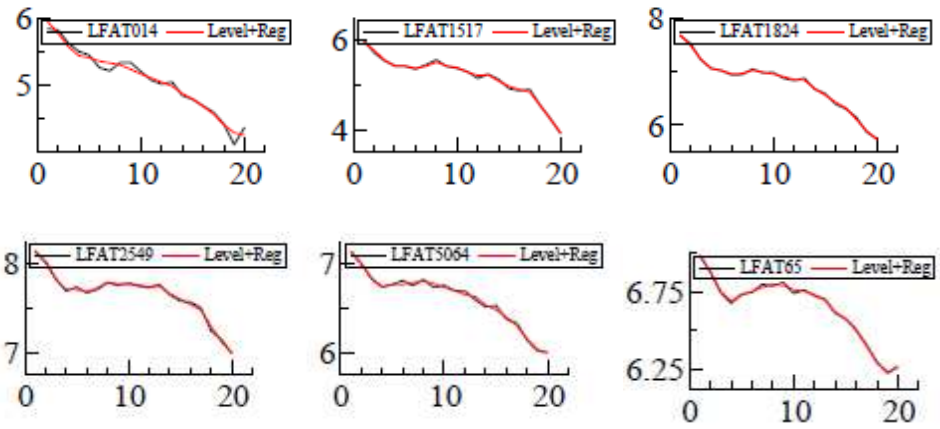

Figure 5: Trends and observed values (log scale).

4.2 Discussion:

The analysis has revealed the presence of two common patterns chosen to be the adult and senior classes of age, which structure the evolution of the number of fatalities. Furthermore only three classes of age are positively sensitive to the evolution of the population: the teenagers, the adults and the seniors. The evolution of mortality is not homogeneous among the classes of age, but not totally heterogeneous and combines the effect of population sizes with two specific stochastic trends.

5 GDP / FATALITIES MODELS

5.1.1 Introduction

Microscopic or occasional changes in economic indicators, interrupting the smooth macroscopic trends, may be associated with road safety changes. Studies on the effect of the global petrol crisis (Tihansky, 1974) in the early seventies on the development of fatality numbers yielded to the conclusion that the reduced speed limits introduced by the authorities, along with more cautious driving by an energy-conscious public have contributed to striking declines in fatalities The economic recession of the early-eighties has been studied by several researchers with respect to its effects on road traffic fatalities (Wagenaar, 1984; Hedlund et al. 1984; Reinfurt et al. 1991). Recently (2008 onwards), road traffic fatalities have exhibited important reductions in several countries. These reductions may not be the sole result of increased efforts in terms of road safety policy efforts, might be partly attributable to the recent global economic recession and its effect on mobility.

5.1.2 Model

If we limit the time period to 20 (or 30, at most) years, the hypothesis that the number of fatalities is I(1) integrated of order 1 (plus a deterministic linear trend) is acceptable. For a longer period, the series is rather I(2) integrated of order 2. The GDP is I(1) integrated of order 1, if we take the real GDP deflated by the price or inflation rate. Otherwise, the nominal GDP is I(2) integrated of order 2. So, we could start the modeling inside the I(1) cointegration framework.

One way to explore the relationship is to associate annual changes in the Gross Domestic Product (GDP) with the related annual changes in the number of road traffic fatalities. It is an exploration of the short term relationship between the number of fatalities and the GDP through their relative rates of change by using the difference of the logarithms (Yannis et al., 2012).

5.1.3 Results

Data for 27 European Union countries have been extracted from the IRTAD database (1975-2010). The dependent variable is the annual percentage change in the fatality rate, the main explanatory variable is the annual percentage change of GDP per capita. A mixed effect modelling technique has been applied with a logarithmic form of the model with fixed effects by groups of countries (Northern, central, southern) and a random effect with an autoregressive covariance structure to capture the time series residual effect. A distinction has been made between the effect of increases and of decreases in the GDP. These effects are supposed to be common to all countries. Both elasticities are significant +0,2 for an increase and -0,34 for a decrease (Table 6). A statistically significant relationship between annual GDP increase and fatality rate increase was established. The relationship between between annual GDP decrease and fatality rate decrease was also significant. Particularly in Northern / Western European countries, a decrease of GDP is associated with a decrease of the fatality rate for the year the GDP decrease occurred, but also one year later.

Fixed effects	Estimate	T	p-value
Intercept	-1,244	-5,983	0,000
[COUNTRYg=Central/Eastern]	0,782	2,530	0,012
[COUNTRYg=Southern]	0,333	0,850	0,396
[COUNTRYg=Northern/Western]	0[a]	.	.
GDPincrease	0,207	2,979	0,003
GDPdecrease	-0,336	-2,970	0,003
[COUNTRYg=Central/Eastern] * GDPincrease	-0,144	-1,704	0,089
[COUNTRYg=Southern] * GDPincrease	-0,015	-0,110	0,912
[COUNTRYg=Northern/Western] * GDPincrease	0[a]	.	.
[COUNTRYg=Central/Eastern] * GDPdecrease	0,230	1,931	0,054
[COUNTRYg=Southern] * GDPdecrease	0,062	0,268	0,789
[COUNTRYg= Northern/Western] * GDPdecrease	0[a]	.	.
Random effects	Estimate	Wald Z	Sig.
AR-1	4,826	19,786	0,000

Table 6: Estimates and t values for the short model based on the differences of logarithms.

We can go one step further by combining the short-term relationships between the first differences

$$\log FAT_{it} - \log FAT_{it-1} = \%FAT_{it}$$
$$\log GDP_{it} - \log GDT_{it-1} = \%GDP_{it}$$
$$\%FAT_{it} = a_i + b\%GDP_{it}$$

with a long-term relationship between the levels (cointegration)

$$\log FAT_{it} = a_i + b_i t + \beta \log GDP_{it}$$

by means of an Error correction model (ECM) for example, to grasp the total dynamics between the fatalities and GDP time-series.

$$\log FAT_{it} - \log FAT_{it-1} = a_i + b(\log GDP_{it} - \log GDT_{it-1})$$
$$+ C(\log FAT_{it} - a_i + b_i t + \beta \log GDP_{it})$$

In fact, we start to explore the long-term relationships with some classical models, such as the Augmented Mean Group model, which introduce two variables in the long-term regression: the mean values of the logarithms of the number of fatalities and GDP. These variables play the role of possible common factors. Furthermore, we suppose that the beta coefficients, the elasticities, take different values according to countries, which is a more admissible hypothesis.

MG		
(lFAT)	**coef**	**z-test**
lGDP	0.74	23.07

CCEMG Pesaran		
(lFAT)	**coef**	**z-test**
intercept	1.266	1.4
lGDP	0.458	2.23
t	0.018	0.015
lFAT2	0.928	5.57
lGDP2	-0.971	-2.51

Table 7: Parameters estimates for the long-term models.

The mean elasticity is significant and equal to 0.46. The mean value of the coefficient of the linear trend is null. The added mean variables are significant (Table 7). The distribution of the elasticities shows that for most countries the elasticity does not differ from zero, except for a dozen of countries (UK, FR, NL, DK, PL …) for which it is significantly positive. The elasticity is negative and significant for CZ only (Figure 6).

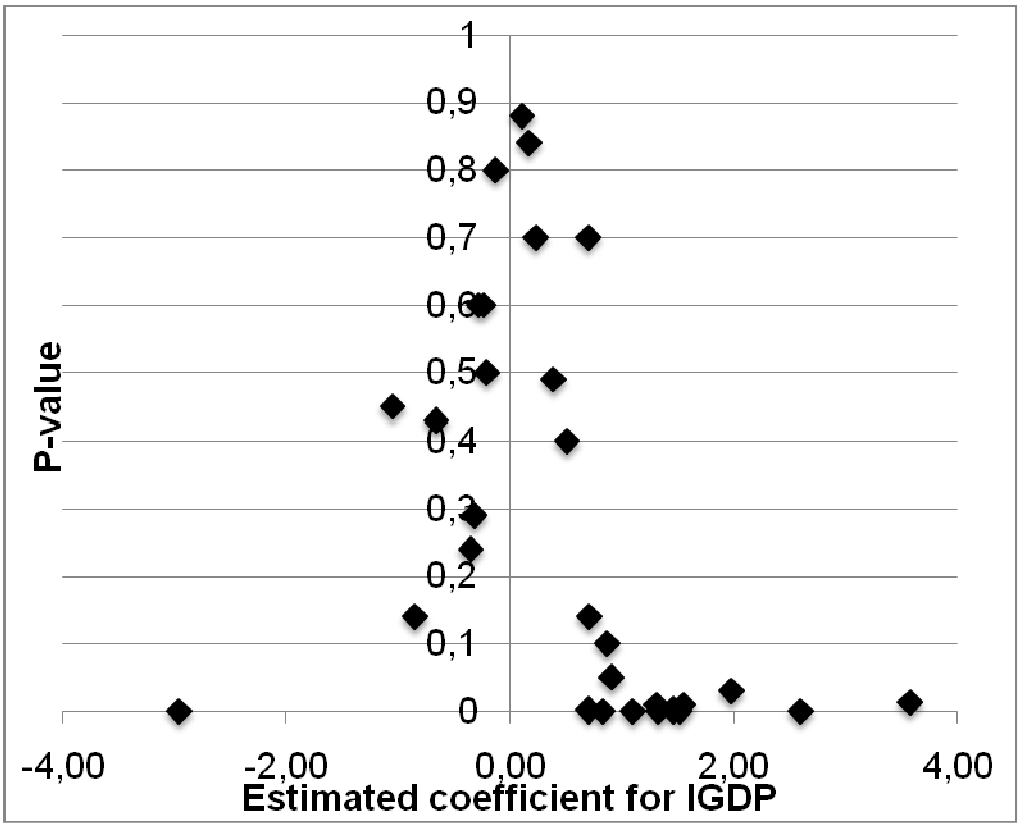

Figure 6: The distribution of the elasticities according to the P-value.

5.1.4 Discussion:

Short term elasticity of GDP to the number of fatalities is different and smaller from long term elasticity. Both elasticities are positive and less than 0.5. In fact there is a strong heterogeneity between countries. The majority of countries does not show any relationship between GDP and the number of fatalities. Having a relationship appears only for a set of a dozen of countries with an elasticity around 1. This phenomenon needs further examination.

6 CONCLUSION

This is an exploratory analysis on the possibilities of multivariate analysis of the number of fatalities over a panel of countries. Statistical techniques appropriate for such macro panel data have been presented and tested successfully on three cases: a disaggregation of the number of fatalities by age on six classes on a specific country, a cointegration study of risk and exposure on two countries, and an exploration of the relationships between fatalities and GDP on 30 countries.

The next step is to improve the multivariate model relating GDP to the number of fatalities. Some interventions related to road safety national measures have to be introduced and the deterministic trends have to be monitored on the basis of the results of univariate models coming from earlier results in DaCoTA. The question of sensitivity of road mortality to GDP has to be studied in relationships with the influence of GDP on mobility on one part and on the risk in other part. The second step is to introduce some identified common factors in the models instead of one or two "proxy" common factors by making some principal components analysis to detect some common patterns. It will pave the way for the development of a multivariate model of the number of fatalities for Europe that will provide clusters of countries according to their specific dynamics and development of risks and exposures through the development of a multivariate structural model with common factors.

REFERENCES

Bijleveld F., Commandeur J., Gould P., Koopman S. J. (2008),. Model-based measurement of latent risk in time series with applications. Journal of the Royal Statistical Society, Series A. *171*(1), 265-277.

Breitung, J., and Pasaran, H. (2005) Unit roots and cointegration in panels. IEPR working paper 05.32

Commandeur, J.J.F. & Koopman, S.J. (2007) An Introduction to State Space Time Series Analysis. Oxford University Press.

Dupont, E., and Martensen, H. (2012). Forecasting road traffic fatalities in European countries. Deliverable 4.4 of the EC FP7 project DaCoTA.

Eberhardt, M. (2011) Panel time-series modeling: new tools for analyzing xt data. UK Stata group users meeting, London.

Hedlund, J. (1984). Comments on Hauer's approach to statistical inference. Accident Analysis and Prevention, 16:163-165.

Hendry, D.F. & Juselius, K. (2000). Explaining cointegration analysis: Part 1. Energy Journal, 21, 1-42.

Koopman S.J., Harvey, A.C.., Doornik, J.A. and Shepard, N. (2007) STAMP 8 Structural Time Series Analyser, Modeller and Predictor. London, Timberlake Consultants.

Koornstra, M. (1973), A model for estimation of collective exposure and proneness from accident data, Accident analysis & prevention, 5(2), 157-173.

Lassare, S. (2012) – Reference in p. 17

Martensen & Dupont (Eds.) 2010. Forecasting road traffic fatalities in European countries: model and first results. Deliverable 4.2 of the EC FP7 project DaCoTA.

Reinfurt, D.W., J. R. Stewart, and N.L. Weaver (1991). The economy as a factor in motor vehicle fatalities, suicides and homicides. Accident, Analysis and Prevention, 23(5):453-462.

Tihansky, Dennis P., 1974, “Impact of the Energy Crisis on Traffic Accidents”, Transportation Research, 8, 481-492.

Wagenaar, A.C. (1984). Effects of macroeconomic conditions on the incidence of motor vehicle accidents. Accident Analysis and Prevention, 16:191-205.

Yannis, G., Papadimitriou, E. and Folla, K. (2012). Effects of GDP changes on road traffic fatalities. IRTAD Meeting, Amsterdam, 18-19 October 20.