The regional analysis of road mortality in Europe: A Bayesian ecological regression model

Vojtech EKSLER (1), **Sylvain LASSARRE** (1), **Isabelle THOMAS** (2)

[1]) *INRETS - GARIG, 23, rue Alfred Nobel - Cité Descartes, 77420, France, eksler@inrets.fr, lassarre@inrets.fr*

[2]) *Département de Géographie de l'U.C.L. and C.O.R.E., Place Louis Pasteur 3, B-1348-Louvain-la-Neuve, Belgium, Isabelle@geog.ucl.ac.be*

Summary

This paper aims at analyzing variations in road mortality between and within 24 European countries and seeks to attribute underlying structural factors to them. We introduce a Bayesian spatial multi-level regression model using generalised linear mixed model (GLMM) framework, which allows the inference for the model to be carried. The population density and country affiliation were used as covariates and were fitted into the model at using four levels of spatial aggregation known as NUTS regions – official statistical units in the EU defined by Eurostat. Population density has been found to have a significant influence on road mortality at regional level. For all countries, the elasticity estimate is -0.33, meaning that a 10% increase in population density will lead to a 3.3% decrease in road fatalities. Multi–level model defined at NUTS-3 level, taking into account NUTS-2 aggregation enables to take into account infra-regional variances in road mortality and produce most reliable model parameter estimates. Variations in Bayes relative risk (mortality ratio standardized by population density and country affiliation) is highest at NUTS-3 level, while it decreases for country level and NUTS-2 level, what suggests the existence of other important underlying factors being responsible for the variations among regions. Mapping Bayes relative risk allows identifying those regions, which should be targeted by national and regional policies. Last, not least, the new ranking of countries according to their road mortality risk adjusted for population density is presented.

Keywords: Road mortality, NUTS, spatial analysis, population density

1 Introduction

It is well known that the level of road safety varies in time and space. Although time variations have always been subject of interest of policy makers and researchers, the spatial component has drawn less attention and has mainly developed in the 1980's. Based on the assumption that many factors affecting the occurrence of road accidents operate at a spatial scale (e.g. demographical characteristics, infrastructure structure and quality, or land use policy), the spatial analysis must be an integrated part of any rational road accidents analysis. Potential utility of such investigation is then the identification of the areas to which the resources should be allocated. Empirical spatial data analysis mainly deals with two levels of spatial data aggregation: highly aggregated data such as states or countries and, at the opposite, local data analyses such as road sections. Regional data (intermediate aggregation level) have often been neglected, partly due to data unavailability, and to unavailability of convenient methods. Only a few papers on regional accident data have been found in the scientific literature (Amoros et al. 2003, Fridstrom et al. 1995, Noland and Quddus 2004, or Shaw et al., 2000); in general, they are restricted to one or just a few countries. More generally, regional variations in road mortality risk in Europe have hitherto received little analytical attention, despite the regular publication of atlases on regional variations in risk analysed as a health-problem, hence taking into consideration the place of residence of the victim instead of the accident location (Eurostat 2005, Shaw 2000). The only exception in this field is the recent work of Lassarre and Thomas (2005), who performed the first descriptive spatial analysis of road mortality risk based on regional fatality data coming from 17 European countries.

The choices of spatial units for regional analyses traditionally tended to be dominated by what is available rather than what is the best and not surprisingly only little effort has been done to identify the ideal aggregation level for a spatial analysis of road mortality. This is but essential as the choice and definition of spatial unit might seriously influence statistical results and consequent conclusions. This phenomenon is often termed the Modifiable Areal Unit Problem (MAUP), and it is formally defined as a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns (Heywood, 1998 and Openshaw, 1977). In general, the use of small spatial units has a tendency to provide unreliable rates because the population used to calculate the rate is small. On the other hand, using larger area units will provide more stable rates but may mask meaningful geographic variation evident with smaller areal units ("scale problem") (Nakaya 2000). Using different boundaries for aggregation, e.g. aggregating neighbours, may improve the problem to a small degree, but does not get round the quantity of variations in counts which remains (aggregation problem). In our case, the use of politically defined spatial units makes most sense since we aim to assess the differences in risk resulting from different road safety policies and provisions.

Road fatalities are considered as the consequences of accidents due to deficiencies in the vehicle-road user-infrastructure system and are traditionally treated as random events. The major factors that are used for explaining their different frequency outcomes are traffic structure and density, quality of infrastructure and vehicles and drivers behaviour. National road safety policies aim to modify the prevalence of these risk factors by means of laws and regulations. Hence, comparing accident statistics means evaluating those policies and searching for structural explanations for the observed differences. However, the policy makers are not able to control all relevant factors, contributing to higher risks in traffic, such as human and physical environment of the mobility context, in which the accident take place.

Thus, the human component consists of factors such as the distribution of the population by age and sex, and the urban and economic structures, whereas the physical component includes climate, physical geography, accessibility and land use. When road mortality rates are compared by area, it is thus more informative if structural factors are controlled for. Epidemiologists often standardize by age–sex structure. For road accidents, we know that young male drivers are approximately three times more at risk than adults, and that senior road users are more vulnerable in cases of collision. Moreover, we know that mobility patterns vary greatly with age and gender. This has not be done in this paper because we think that (a) factors such as differential mobility on the motorways or local roads are more important, (b) variations in the age–sex distribution between regions are fairly low and (c) migration and mobility effects, especially among young drivers, are difficult to grasp at the scales of analysis that are used in this paper.

In this paper, we use the population density as a synthetic indicator (explanatory proxy) to “explain” major differences in road mortality across European regions, as it was used by Lassarre and Thomas, 2005 in their exploratory paper. Population density takes into account many factors which are often not available by region (such as the exposure in traffic for different road types, road network or urbanization). We can assume that regions with higher population density have developed more sophisticated and safe road network, together with the other optional transport choice, partly as a response to higher mobility demands of the population and partly due to higher economical performance allowing for infrastructural improvements. Often the distinction is made between urban and rural areas as two different environments having different accident outcomes. In particular, mortality in rural areas is much higher than in urban partly due to the availability and efficiency of emergency services (see e.g. Clark and Cushing, 1999, 2004), generally higher driving speed, drinking and driving prevalence, lower use of protective systems and older vehicle fleet. Less populated areas are also often characterized by a different age structure and social deprivation leading to additional risk factors (Baker et al. 1997, Clark, 2003). Rural residents travel more kilometres by road and are, thus, more likely to be involved in a serious collisions (Muelleman and Mueller, 1996).

In this paper, we have replaced single spatial level analysis by an analysis combining several aggregation levels. Instead of having one model for each level, we introduced a multilevel model, which allows producing non-biased estimates in order to correct the biases due to the ecological effects of single level models. To do this, we used Bayesian modelling techniques using Markov Chain Monte Carlo fitting procedure to estimate the distribution of all model parameters. It allows determining a confidence interval of a country effect, which improves significantly the interpretation of the ranking of the countries according to their standardized road mortality ratio. It further give us a possibility to map Bayes relative risk at NUTS-2 and NUTS-3 level in a coherent way taking into account an influence of regional and infra-regional heterogeneities.

2 Methods

2.1 The data

The analysis covers 25 EU member states and is geographically based on the EUROSTAT NUTS 2003 regional classification (Eurostat, 2005). Data are analysed at four aggregation levels, which correspond to the first four “Nomenclature of statistical territorial units” levels (NUTS-0, NUTS-1, NUTS-2 and NUTS-3). The classification is hierarchical and subdivides each country (NUTS-0) into NUTS-1 territorial units, each of which is subdivided into

NUTS-2 territorial units, these in turn are each being subdivided into NUTS-3 units. Nevertheless, some territorial units are classified at several NUTS levels.

Regional accident fatalities data refer to 2002 and were exclusively collected for this analysis by personally contacting the national road administrations and statistical offices. The following areas were excluded from the analysis due to data unavailability or unreliability: 31 NUTS-3 regions of Poland and 3 NUTS-3 regions of Scotland (UKM41-UKM43). Furthermore 4 NUTS-2(3) French overseas regions (DOM-TOM) were excluded due to their outlying location. The 30-days fatality definition originally adopted by the Vienna Convention in 1968 is used here: a road death is deemed to have occurred when a person injured dies within 30 days of the crash. We did apply IRTAD/CARE correction factors to standardize fatalities data in the countries using different definitions as follows: France (1.057), Italy (1.078), Latvia (1.08) and Slovakia (1.25).

Level	N	AREA (square km)			POPULATION (thousands)			ROAD FATALITIES		
		Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.
NUTS-0	25	155708	316.0	543965.0	18159	396.0	82482.0	2004	16	7651
NUTS-1	88	44741	161.4	410934.2	5118	26.1	18062.9	554	1	1771
NUTS-2	249	15562	12.0	154312.0	1793	26.1	11106.7	195	1	1024
NUTS-3	1211	4601	10.0	98910.7	375	19.2	5499.8	34	0	506

Table 1: Basic indicators of NUTS regions for 2002

Basic indicators of NUTS regions summarized in Table 1 demonstrate a great heterogeneity in terms of area and population at all aggregation levels. Despite the NUTS classification was originally strictly based on population criteria, a lack of integrity can nowadays be observed in population size and area coverage of the regions at the same NUTS level.

The level of road safety is measured in this paper in terms of mortality rate. Elsewhere presented also as a health risk, this is the one of the most common indicator for the evaluation of road safety level and the most common indicator used by the health sector to prioritise diseases and other causes of death. The major reason for choosing this indicator is the unavailability of exposure data on different levels of spatial aggregation applied in this study. The mortality rate is the number of fatalities in a region i (Y_i) during 1 year divided by the total number of inhabitants residing in i (N_i) in the middle of that year ($T_{mi}=100\ 000\ Y_i/N_i$). Once divided by the mean mortality rate observed for all EU25 countries, a mortality ratio can be obtained: $T_{mi}^{*}=T_{mi}/11.03$.

2.2 Introduction of explanatory variables

Road safety researchers have been traditionally using different explanatory variables in their statistical models in order to provide a link between observed accident or fatality counts and different infrastructure-vehicle-driver risk factors. As an example the work of Friedstrom et al., 1995 can be mentioned here. When analyzing large sets of data covering large number of different countries, the number of available explanatory factors is significantly reduced to a limited number of rather general factors describing the conditions of road traffic. One of such factors is population density, which can be understood as a synthetic indicator for mobility demands, demography structure, urbanization, road user mix, infrastructure structure and quality, and response time of emergency service. (e.g. Baker et al., 1987, Muelleman et Mueller, 1996, Yang et al.,1997)

We use Pearson correlation test for measuring the relationship between mortality rate (T_m) and population density at different aggregation level. At all four level of spatial aggregation, the Pearson product-moment correlation coefficients (r) are significant confirming a very strong negative correlation between the two variables. For example for NUTS-3 regions, the tests parameters are: r =-0.69, p <0.001. (Table 2) When considering subsets of very densely populated regions (population density higher than 500 inhabitants per sq. km), the relationship is relatively strong (r =-0.45, p <0.001) while it becomes weaker when only scarcely populated regions are considered (typically with the population density under 50 inhabitants per square kilometre). (r =-0.25, p <0.01) The identified relationship between mortality rate and population density at NUTS-3 level for all countries is strongly negative, meaning that the increase in population density leads to a decrease in mortality rate (T_m).

NUTS level	r	CI (95%)	t-value	p
NUTS-0	-0,419	-0.029, -0,699	-2,214	<0,1
NUTS-1	-0,655	-0.517, -0.760	-8,085	<0.001
NUTS-2	-0,616	-0.532, -0,687	-12,312	<0.001
NUTS-3	-0,682	-0,649, -0,712	-31,648	<0.001

Table 2: Pearson correlation coefficients for Log(PopDens) vs. Log(Mortality)

Figure 1 shows for each country the distribution of the NUTS-3 regions in terms of mortality rate (log) versus population density (log). Three countries were excluded because they do not have a NUTS-3 regional division Luxemburg (LU), Malta (MT) and Cyprus (CY). The relationship is most significant (more pronounced) in case of Austria (AT), Belgium (BE), Germany (DE) and Portugal (PT), while there is no relationship to be observed in case of Slovakia (SK), Finland (FI) and Lithuania (LT). In some countries, the capital city region is characterized by a very low mortality rate and a very large population density. This is partly to be explained by the limits of the NUTS regions for those cities: let us take the example of Brussels. Brussels is at the NUTS-3 level represented by a region that corresponds to the "Brussels Capital Region", an administrative reality in Belgium. It is well-known that the city of Brussels sprawls much beyond the limits of this entity and that hence, the NUTS-3 region does not include the suburbs (see e.g. Thomas et al, 2003 and Vanderhaegen et al., 1996). In other cases, the capital city can be fully included in a very large administrative unit such as Madrid, or Paris. This is a typical problem in geography which disallows to relate directly different road environments to the spatial units.

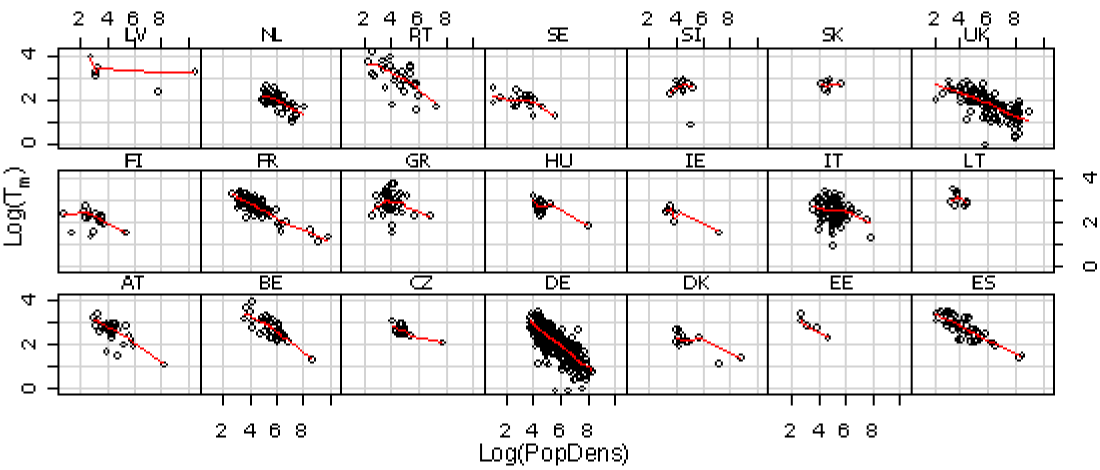

Figure 1: Smooth curve for the relationship between log-Mortality rate (LogMort) and log-Population density (LogPop) by country for NUTS-3 regions

Let us now to further statistically describe the relationship between road mortality (log) and population density (log). We use six Poisson (generalised linear additive) regression models (McCullagh and Nelder, 1989). Model 1 depends on the logarithm of the density and model 2 on a spline function of the logarithm of the density. Models 3 and 4 include an intercept, which varies by country (NUTS-0 level). Model 3 allows for a separate intercept for each country but imposes a straight-line relationship (with a common slope) with log-density. Model 4 relaxes this assumption and instead allows for a common non-linear spline relationship after allowing for different country intercepts. Mortality is indeed affected to some extent by national policies towards drivers and cars. Issues such as speed and alcohol limits, safety features, safety enforcement and driving tests are all determined nationally. The model 5 is the model 4 for a broken-line relationship. This model assumes two segments line, the first segment with a slightly positive slope of a regression line 0.24 [CI (95%)=0.19,0.28] for all Log(PopDens)=2.394 (11 hab/km^2) and the second with a negative slope -0.49 [CI (95%)=-0.54,-0.45] for higher values of Log(PopDens). In model 6, we add the interaction between the NUTS-0 level (country effect) and the logarithm of the population density.

(a) model 1, $\log(\lambda_i)=\alpha+\beta$ log(PopDens)
(b) model 2, $\log(\lambda_i)/=\alpha+\beta$-spline log(PopDens)
(c) model 3, $\log(\lambda_i)=\alpha+\beta_c+\beta_2$ log(PopDens)
(d) model 4, $\log(\lambda_i)/=\alpha+\beta_c+\beta_2$-spline log(PopDens)
(e) model 5, $\log(\lambda_i)=\alpha+\beta_c+\beta_{2c}$ log(PopDens) - segmented
(f) model 6, $\log(\lambda_i)=\alpha+\beta_c*\beta_{2c}$ log(PopDens)

Considering the analysis of deviance, as the observed Pearson value of model 6 is high (χ^2= 6761 with 1116 degrees of freedom), it is preferable to take into account the overdispersion by inflating the variance by a factor Φ. The estimation of Φ can be done by dividing Pearson Chi-square by the difference in degrees of freedom as follows: $\chi^2/(n-p)=6.006$. The statistic for comparing the models is equal to the difference in deviance divided by Φ multiplied by the difference in degrees of freedom and should be compared with the appropriate *F*-distribution (Firth, 1991). The spline function is fitted by means of a generalized additive model. The spline fit accounts for 3 degrees of freedom. The interaction term relative to the different slopes by country (Model 6) improves the model to a lesser extent (1029/(6.006×21)=8.16 which is large compared with the *F*(18.1116) distribution (1.60) and should not therefore be neglected. For all countries, the slope is equal to -0.29, it is less pronounced for Italy (-0.13) and Sweden (-0.12) and more pronounced for Belgium (-0.45), Austria (-0.43) and Portugal (-0.39). Generally, a 10 % increase in population density leads to a 2.94 % decrease in the number of fatalities, on average. Considering the statistic 549/(6.006×3)=30.47 which is large compared with the F(3.1134) distribution (3.84), the best model must include a country effect (NUTS-0 level) and the logarithm of the population density with a non-linear (spline) rather than a linear form.

Model	Deviance	Degrees of freedom	Difference in degrees of freedom	Test statistic
Null (NUTS-3)	62114	1162	-	-
1, logDens	27145	1161	1162-1161=1	5822.34 vs. F(1,1161)
2, spline (logDens)	23329	1158	1161-1158=3	211.79 vs. F(3,1158)
3, Country+logDens	7790	1137	1161-1137=24	134.28 vs. F(24,1137)
4, Country+ spline(logDens)	7241	1134	1137-1134=3	30.47 vs. F(3,1134)
5 , Country+ (logDens)	7238	1134	1137-1134=3	30.47 vs. F(3,1134)
*6, Country*logDens*	6761	1116	1134-1116=18	4.44 vs. F(18,1116)

Table 3: Analysis of deviance for the Poisson regression

As mortality rate decreases with population density, it is recommended to take this factor into account when comparing countries, while both linear and spline form can be suggested. We hence select the model 5 with population density and country effect as covariates.

2.3 The Bayesian ecological regression model

We here propose an ecological Bayesian regression model incorporating the two chosen effects influencing road fatality counts in European regions – population density and country affectation to analyse the heterogeneity in road mortality across European regions. Bayesian methods have recently gained increased popularity and recognition especially among epidemiologists as they allow handling context data where unmeasured confounders and spatial autocorrelation are evident, what is the weakness of traditional methods such as Gaussian and Poisson regression models (see e.g. Elliott et al., 1992). The most important feature of this kind of model is that it allows incorporating random spatial effects into the modelling of potential associations between fatality occurrence and covariate effects, at an ecological (i.e. regional) level. This model framework makes it possible to ascertain whether residual variation remains after accounting for known and measured covariate effects and whether the residual effects suggest spatial patterns or clusters. This method also facilitates spatial smoothing when regions under investigation involve small-population areas (areas not necessarily small in geographical size, but rather with small at risk population, since they are subject to high chance of variation). Moreover, it enables data sharing, i.e. risk smoothing, over space, which often results in more reliable risk prediction (MacNab, 2004).

Markov Chain Monte Carlo (MCMC) is a generic name given to a whole set of model fitting procedures that has revolutionized the use of this set of statistical procedures by providing a very convenient method for fitting ecological models which cannot be solved analytically. A tractable introduction to the approach is available in a road safety context (Tunaru, 1999). Briefly explained, MCMC enables an estimate of the posterior distribution, something that cannot be estimated analytically for some models. However, unlike many numerical algorithms, this is a simulation approach which is not guaranteed to converge around the correct solution, hence in addition to model fitting diagnostics it is important to carry out convergence diagnostics with these models (Cowles and Carlin, 1996). Our model has been fitted with the WinBUGs software (Win referring to Windows and BUGs referring to Bayesian Updating by Gibbs Sampling -one of a number of MCMC methods) as this software is particularly capable of fitting these spatial models (Spiegelhalter et al., 1998).

Road mortality counts can be conveniently modeled assuming the data to be Poisson distributed:

$$Yi \sim \text{Poisson}\ (\lambda_i N_i)$$

where λ_i is the Poisson parameter, often referred to in this context as the mortality rate, and N_i is the number of road fatalities in region i. In the generalized linear model, the λ_i for each region contains covariates and parameters which indicate the association between the covariates and the mortality rate. This model can be extended to form a generalized mixed linear model incorporating a random effect within λ_i. Such model combining fixed and random effects is sometimes called also as a mixture model. The Bayesian modelling may be implemented for rates (Y_i/N_i) and counts (Y_i), assuming that the response Y_i's, given a vector $\boldsymbol{\nu}$ of random effects, are conditionally independent, $Y_i|\nu \sim Poisson(\lambda_i^\nu)$, while λ_i^ν represents the expectation of Y_i given all the random effects. By other words, while conditioning on random effects, the corresponding fatality counts are assumed to follow Poisson distribution. In the analysis of mortality rates, one assumes the following form of the model:

$$log(\lambda_i^{\nu})=log(N_i)+\beta_c+\beta_2.(logDens)+\nu_i$$

where $log(\lambda_i^{\nu})$ represents the expectation of Y_i conditioning on random spatial effects ν_i, $log(N_i)$ is an offset population level intercept, β_c is a fixed effect measuring the strength of the association between the country and mortality rate in a region i (country effect) and β_2 is a fixed regression parameter for population density. Both β_c and β_2 are believed to be random. Random effect ν captures all the uncertainty regarding the differential mortality rate in each geographical unit, such as that arising due to reporting error, missing covariates, over-dispersion or even genuine underlying differences in mortality rate. The $exp(\nu_i)$'s represent local area mortality ratios adjusted for country and population density (also called relative risk). If ν_i are assumed to be zero mean normal variables $\nu \sim N(0,\sigma^2)$, the model takes on the fairly standard mixed form and only σ has to be estimated during the model fitting procedure. Our model can be easily extended by spatial interaction term in form of an additional random effect (Aguero-Valverde and Jovanis, 2005), however we decided to do not do so, since it is very difficult to rightly interpret all existing interaction between neighbouring regions.

Within a generalised linear mixed model (GLMM) framework, inference for the models can be carried out using full Bayesian methods described above. The Bayesian methodology requires prior assumptions about the model to be made. The parameter β_c has been assumed to follow a uniform distribution (-100,100) and β_2 a uniform $(-\infty,\infty)$ distribution. The precision (inverse value of variance) $\tau(\nu)$ is assumed to follow a gamma distribution. $\tau.\nu=1/\sigma^2\sim$ gamma (0.001,0.001). We run 50.000 iterations separately for each data set. Such high number of iterations is a fundamental condition to reach a convergence of all monitored parameters in our models, especially if the prior variables are suffering from random variation. As the result, standard deviation, MC error, Deviance Information Criterion (DIC) proposed by Spiegelhalter *et al*. (2002), and 5% and 95% confidence intervals of most important parameters as the average value from 25.000-50.000 iterations taken as a reference, are considered. (After 25.000 running 25.000 iterations all model parameters are likely stabilised and converge to the true values of relevant posterior distributions.) This choice was justified by the comparison of Bayesian ecological regression model parameters with GLMM parameters assuming Poisson and negative binominal probability distribution.

2.4 Choropleth maps

Choropleth maps provide an easy way to visualize how a measurement varies across a geographic area and allow identification of areas with extreme values. In our paper, the values of Bayes relative risk are of a primarily interest rather then values of mortality rates and are later mapped using choropleth map techniques. (Bayes relative risk refers here to the random effects representing an extra quantity variation estimable within the map and which can be ascribed a defined probabilistic structure.) The histogram of the Bayes relative risk $exp(\nu)$ follows a bell-shaped distribution curve; it is more or less symmetrical at all NUTS levels, but becomes slightly positively skewed at NUTS-2 and NUTS-3 level of spatial aggregation. In order to produce reliable maps, the choice of the number of classes is essential, as too few classes will tend to emphasize broad regional patterns and some details can be lost. Conversely, too many classes can make the map difficult to interpret (see e.g. Campbell, 2001). We used the equal interval counts technique for defining the class limits because this method ensures, that the ranges are well represented by their averages, and that the data within each range are fairly close together. We adjusted all values of Bayes relative risk by dividing them by the average value of Bayes relative risk at each level of spatial aggregation. Class limits identified at NUTS-2 level were slightly modified assuring that the limits of the middle class are symmetrically set around 1 so it has the same interval length for positive and negative values (0.93-1.07). Identified class limits were then employed to map the Bayes relative risk coming from the NUTS-3(2) model at NUTS-3 spatial aggregation level.

3 Results

Running the specified model for the four sets of data each of which corresponding to one level of spatial aggregation allows to study in details the changes of key model parameters at different aggregation levels. (Please note, that in order to study the effects of spatial aggregation on model parameter, within Bayesian modelling we had to exclude Poland from all datasets.) Regression coefficient β_2 representing a fixed effect of population density on the occurrence of road fatalities is changing significantly with different aggregation levels as it decreases from -1.245 (NUTS-0) to -0.293 at NUTS-2 level. It becomes however relatively stable at NUTS-2 and NUTS-3 level of spatial aggregation, whereas it can be assumed, that the best estimate of the parameter is close to the -0.326 [CI (95%)=-0.307,-0.326] estimated from the model at NUTS-3 level. By the way, with spatial aggregation, both standard deviation and MC error estimates increase. (Table 4)

β_2	mean	SD	MC error	5.0%	median	95.0%
NUTS-0	-1.245	0.1019	0.0081	-1.457	-1.209	-1.126
NUTS-1	-0.4782	0.0237	0.00181	-0.5171	-0.481	-0.4347
NUTS-2	-0.2925	0.0143	0.001029	-0.3155	-0.2921	-0.27
NUTS-3	-0.3257	0.01107	0.000752	-0.344	-0.3257	-0.3069

Table 4: Regression coefficient β_2 at four levels of spatial aggregation

Let us now to analyse the effect of spatial aggregation on standard deviation of random effect (relative risk) $\sigma(\nu)$. The standard deviation of relative risk decreases from 1.153 at NUTS-0 level to 0.294 at NUTS-3 level, but the decrease is not linear (or proportional) over all levels of spatial aggregation employed. Not surprisingly, the MC error decreases with the increase in the number of regions in the model (spatial disaggregation). (Table 5)

$\sigma(\nu)$	mean	SD	MC error	5.0%	median	95.0%
NUTS-0	1.1530	0.4793	0.03459	0.3166	1.1690	1.9140
NUTS-1	0.3398	0.0384	0.00118	0.2819	0.3373	0.4065
NUTS-2	0.2423	0.01445	0.00019	0.2193	0.2418	0.2670
NUTS-3	0.2943	0.009112	0.00011	0.2795	0.2941	0.3096

Table 5: Standard deviation of random effects $\sigma(\nu)$

The model run at both NUTS-2 and NUTS-3 level allows reliable estimation of all covariates in the model and of random effects ν, but we might aim to further improve the estimates. We can do so by the integration of the random effects coming from the models run at NUTS-3 aggregation level into the new model run at NUTS-2 level taking this variation into account. The two random variation parameters are defined as follows: $\nu_i = U_j + V_i$ (i=1-1139, j=1-238 for NUTS-3 and NUTS-2 regions.) This model, later called NUTS-3(2), takes into account infra-regional variations of road mortality between NUTS-2 and NUTS-3 level. The sum of the variance for the two parameters is now almost equal to the variance of ν_i at NUTS-3 level: $(0.2943)^2 \cong (0.2457)^2 + (0.1682)^2$. The resulting values for all variables are as follows: The coefficient β_2 has now a value close to the β_2 at NUTS-3 level and its standard error of its estimate is now significantly lower than at any other level of spatial disaggregation. This is likely the most reliable value of this coefficient (-0.321 [CI (95%)=-0.305,-0.321]). Note also that the standard deviation of V_i (NUTS-3 vs. NUTS-0) is much higher than the standard deviation of U_i (NUTS-2 vs. NUTS-3 regions). This confirm that the hypothesis that the best estimates of model parameters come from the models counting for infra-regional variation in mortality rates. By the way, the DIC for this model (7748) is slightly lower than in case of the model run at NUTS-3 level (7784).

NUTS-3(2)	mean	SD	MC error	5.0%	median	95.0%
β_2	-0.3207	0.0102	0.000667	-0.3377	-0.3207	-0.3047
$\sigma(V)$	0.2457	0.009338	0.000155	0.2306	0.2456	0.2613
$\sigma(U)$	0.1682	0.01615	0.000432	0.1427	0.1678	0.1955

Table 6: Regression coefficient β_2 and standard deviation of random effects $\sigma(\nu)$ for NUTS-3 (2) model

We can now study the effect of spatial aggregation on the regression coefficients β_c standing for country effects in the above presented models. The variance of 24 parameters β_c decreases with spatial disaggregation in general, while there is only little difference between the two models results for NUTS-3(2) level as shown in Table 6. Comparing the standard deviation of regression parameters $\sigma(\beta_c)$ with random effects parameters $\sigma(\nu)$ the following conclusions can be drawn: The variance of β_c parameters as well as the variance of ν is more or less the same at different aggregation levels, meaning that the same degree of heterogeneity in terms of road mortality can be found at all aggregation levels. The strongest variation in term of relative risk exists at NUTS-3 level, followed by Country level and NUTS-2 aggregation level, what point to the fact, that the national road safety provisions have a similar weight as the regional differences in terms of traffic conditions, mobility or others.

Level	NUTS-1	NUTS-2	NUTS-3	NUTS-3(2)
$\sigma^2(\beta_c)$	0.186	0.028	0.055	0.049
$\sigma(\beta_c)$	0.432	0.169	0.234	0.222
$\sigma(\nu)$	0.339	0.242	0.294	0.294

Table 6: Standard deviation of random effects $\sigma(\nu)$ and regression coefficient β_2 for NUTS-3 (2) model compared to previous results

We can now estimate the mortality ratios standardized by the population density for NUTS-0 regions as a result of the NUTS-3(2) modelling procedure, i.e. the mortality ratios of EU24 countries involved in the analysis. The exponential of the β_c parameters, as a result of MCMC Bayesian modelling approach after having been centred to 1, is shown in Figure 2. Each line of the plot features a line indicating the 5^{th} and 95^{th} percentiles for the posterior distribution of the mortality ratio, with a point denoting the mean. The interval is naturally large for the small countries, having only limited number of NUTS-3 regions such as Malta, or Luxemburg and for those countries where significant variation in road mortality among regions exists, while relatively narrow for the countries with a large number of NUTS-3 regions. This ranking is fairly different from the simple ranking of countries according to their mortality ratio. Countries with relatively low population density such as Finland or Sweden now occupy better position in ranking, while the countries with relatively high population density (Belgium, Luxemburg) belongs now to the worst performing countries. By the way, the ranking based on NUTS-2, respectively NUTS-1 model is fairly similar to the one presented here. A wider confidence interval is not surprising. Some less relevant changes can be however traced among the countries with a high mortality ratio. E.g. for NUTS-2 regions based ranking, the worst performing country is Portugal.

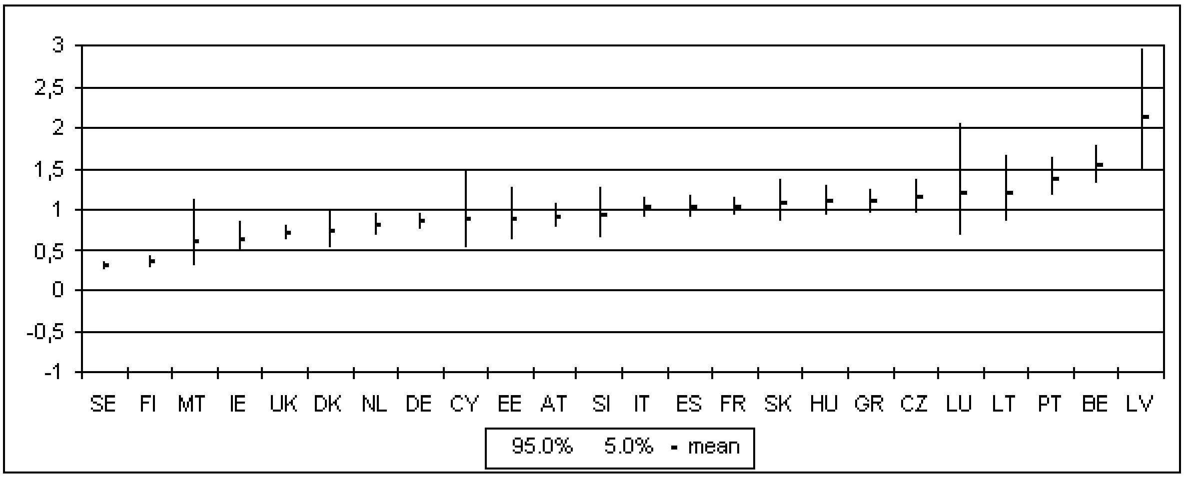

Figure 2: Mortality ratios of 24 EU countries standardized by population density (NUTS-3)

In regards to the 95% confident intervals towards 1, the four groups of countries can be identified: two countries with extremely low standardized mortality ratio (Sweden, Finland), 6 countries, whose confidence interval does not contain 1 (Malta, Ireland, United Kingdom, Denmark, Netherlands and Germany), three countries with extremely high mortality ratio (Portugal, Belgium and Latvia) and then all other countries, whose confidence interval contains 1. Comparing countries among each other, the three groups can be distinguished: The two Scandinavian countries with ratio under 0.5, two countries (Belgium, Latvia) with the ratios above 1.5 and all other countries.

The highest relative difference between the mortality ratio standardized by population density and crude mortality ratio (expressed as a rate between the both) can be found highly populated countries such as United Kingdom, Belgium, Netherlands and Malta, while the smallest one for scarcely populated countries such as Sweden, Finland, Estonia and Lithuania.

We can further map the Bayes relative risk (the exponential of random effect v_i) having the effect of shrinking on the estimate of the mortality across regions. Map of Bayes relative risk provides a smooth representation of the road mortality risk in Europe, adjusted for the population density and country effect (Figure 3). Surprisingly, some very densely populated regions (usually capital regions) still appear on the maps as light spots (see e.g. Ile-de-France, Madrid, etc.). In general, smoothing provided by the use of Bayes approach is clearly tractable, likely diminishing heterogeneity in mortality risk. Nevertheless, all significant variations remain clearly visible on the maps presented. (e.g. Italy or Sweden) By the way, the heterogeneity in Bayes relative risk for NUTS-2 and NUTS-3 regions is tractable also in the countries in which the mortality rates shows rather homogenous pattern (e.g. Sweden, Czech Republic, Hungary). The highest variance in Bayes relative risk across regions is traceable in Portugal, Latvia and Finland. There is a clear north-south division of Bayes risk across some countries like Italy, Sweden, or Finland. It can be assumed that this is due to geographic, demographic, weather-conditions or economic differences between the northern and southern part of each country. In Italy, for example, significant differences in land use, or mobility demands can be identified between the North and the South. To attribute these differences to the observed road fatality counts remains, however a challenging task for future research.

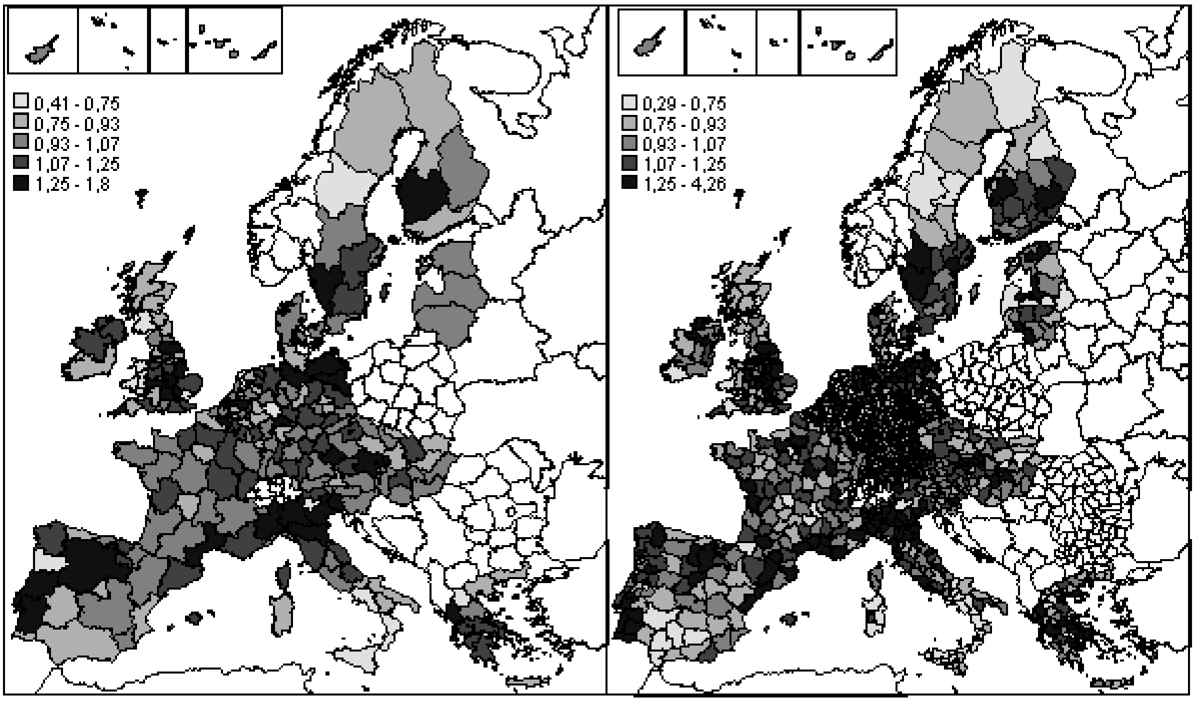

Figure 3: Bayesian relative risk in Europe $\exp^{v_i}$ at the NUTS-2 (left) and NUTS-3(right) aggregation level for 24 European countries

4 Discussion, conclusions

The use of NUTS classification allows an easy access to regional data, however their use should be always taken with certain precaution, since there are significant differences between the regions belonging to the same aggregation level. For example, French counties (NUTS-3) correspond to Belgian provinces (NUTS-2) in terms of area and population making the comparison rather difficult. A solution might be mixing regions from several aggregation levels, nevertheless, no one can guarantee that this will lead to a better description and understanding of studied problems.

Regional disparities in terms of road mortality are higher than disparities identified among countries. It's likely that the existing regional disparities existing within one country comes from structural differences (represented here by population density), but there are still significant differences coming from other, road safety policy related issues.

Standardizing road mortality by population density allows taking into account a large number of underlying structural factors having a significant influence on road safety of countries and regions, which cannot be addressed empirically. It should be however highlighted, that it cannot take into account all existing structural differences. While intra-national differences can partly be explained by the different general conditions of road traffic, such as level of motorization, or road safety related measures and partly by other differences, the intra-national heterogeneities in road mortality cannot be unambiguously attributes to many of these structural factors. Generally higher level or road safety in densely populated regions can be due to lower travelling speed, availability of public transport services lowering traffic performance of individuals, better access and quality of emergency services, more developed infrastructure and higher standards of vehicles in traffic. Mortality ratios of countries standardized by population density (as the product of modelling procedure) allow ranking 24 European countries according their road safety level. Such ranking allows identifying those countries with significantly better and worse road safety records then the others. Similarly maps of Bayes relative risk allow identifying those regions having significantly different mortality rates than country's average.

The use of Bayesian modelling allows mixing fixed and multi-level random effects what is a clear advantage comparing to commonly used general linear models. Mapping random effects allows identifying at regional and infra-regional level those areas with significantly higher mortality rates, being standardized by fixed effects (country effect and population density here). Spatial disaggregation decreases, in general, the value and variance of all model parameters, while the confidence interval narrows at the same time. NUTS-2 aggregation level represents a satisfactory aggregation level, while NUTS-3 level of spatial aggregation allows for most reliable estimates of model parameters but the combination of the two levels leads to a further precision of model parameter and his robustness.

This analysis is limited on the influence of some basic structural factors on observed level of road un/safety in EU25 countries due to data unavailability. Analyses made at national level suggest a wide range of additional explanatory factors, which can be used to explain the heterogeneity of road mortality within the EU. Extending a recent model by a time-series data in a hierarchical model should allow to further precise the estimates of Bayesian relative risk, as it will take into account variations of road fatality counts in time. Similarly, the model does not take into account the effects of spatial autocorrelation (i.e. the effect of the spatial dependence among observations) which produces higher variance of the estimates.

References

Amoros, E., Martin, J.L., and Laumon, B., 2003. Comparison of road crashes incident and severity between some French counties, Accid. Anal. Prev. 35 (2003), pp. 537–547.

Armhein, C., 1995. Searching for the elusive aggregation effect: Evidence from statistical simulations. Environment & Planning A, Jan95, Vol. 27 Issue 1, p105

Aguero-Valverde, J., Jovanis, P.P., 2005. Spatial analysis of fatal and injury crashes in Pennsylvania, Accid. Anal. Prev. 38 (2005), 3, pp. 618–625.

Baker, S., Whitfield, R. A., O'Neill, B., 1987. Geographic variations in mortality from motor vehicle crashes, New Engl. J. Med. 1987, 316, 1384–1387

Clark, D., Cushing, B., 1999. Predicting regional variations in mortality from motor vehicle crashes. Acad. Emergency Med. 1999, 6, 125–130

Clark, D., 2003. Effect of population density on mortality after motor vehicle collisions, Accid. Anal. Prev. 2003, 35, 965–971

Clark, D.E., Cushing, B.M., 2004. Rural and urban traffic fatalities, vehicle miles, and population density, Accid. Anal. Prev., Volume 36, Issue 6, Pages 967-972

Campbell, J., 2001. Map use and Analysis, Boston: McGraw-Hill

Clayton, D., Kaldor, J., 1987. Empirical Bayes estimates of age-standardised relative risks for use in disease mapping, Biometrics 1987, 43, 671–681

Cowles, M.K., Carlin, B.P., 1996. Markov Chain Monte Carlo convergence diagnostics: a comparative review, J. Am. Statist. Ass. 91, pp. 883–904

Elliott, P., Cuzick, J., English, D., Stern, R. (Eds.), 1992. Geographical and Environmental Epidemiology: Methods for Small-Area Studies. Oxford University Press, London.

Eurostat, 2005. Regions - Statistical Yearbook 2004, EC Brussels

Firth, D., 1991. Generalized linear models. In Statistical Theory and Modelling: in Honour of Sir David Cox. FRS (V. Hinkley, N. Reid and E. J. Snell), pp. 66–93. London: Chapman and Hall 1991

Fridstrom, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accid. Anal. Prev. 1995, 27, 1-20

Heywood, 1998. Introduction to Geographical Information Systems, New York: Addison Wesley Longman

Lassarre, S., Thomas, I., 2005. Exploring road mortality ratios in Europe: National versus regional realities, Journal of Royal Statistical Society 2005, 168, Part I, pp. 127-144

MacNab, Y.C., 2003. A Bayesian hierarchical model for accident and injury surveillance, Accid. Anal. Prev., Volume 35, Issue 1, January 2003, Pages 91-102

MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis, Accid. Anal. Prev., Volume 36, Issue 6, November 2004, Pages 1019-1028

McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models, 2nd edition, London: Chapman and Hall

Muelleman, R.L., Mueller, K., 1996. Fatal Motor Vehicle Crashes: Variations of Crash Characteristics within Rural Regions of Different Population Densities. Journal of Trauma-Injury Infection & Critical Care. 41(2):315-320

Nakaya, T., 2000. An information statistical approach to the modifiable areal unit problem in incidence rate maps. Environment & Planning A, Vol. 32 Issue 1

Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road causalities in England, Accident analysis and Prevention 2004, 36, 973-984

Openshaw, S., 1977. A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitionong, and Spatial Modelling, Transactions of the Institute of British

Geographers, New Series 2, pp. 459-475.
Shaw, M., Orford, S., Brimblecome N., Dorling, D., 2000. Widening inequality in mortality between 160 regions of 15 European countries in the early 1990s. Social Science&Medicine 2000, Volume 50, Issues 7-8 , Pages 1047-1058
Spiegelhalter, D., Thomas, A., Best, N., 1998. WinBUGS version 1.1.1 user manual. Technical report, MRC Biostatistics Unit, Cambridge
Spiegelhalter D.J., Best N.G., Carlin B.P. and Van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. Roy. Statist. Soc. B . 64 , 583-640.
Thomas I., Tulkens H., Berquin P. 2003. Quelles frontières pour Bruxelles? La réponse d'un exercice statistique, géographique et économique.
Eggerickx T., Gourbin C., Schoumaker B., Vandershrick C, Vilquin E., Populations et défis urbains. Louvain-la-Neuve, L'Harmattan, pp. 79-97
Tunaru, R., 1999. Hierarchical Bayesian models for road accident data, Traffic Eng. Contr. 40 pp. 318–324.
Van der Hagen H., Van Hecke E., Juchtmans G., 1996, « Les régions urbaines belges en 1991 », Etudes Statistiques, 104.
Yang, C.Y., Chiu, H.F., Cheng, M.F., 1997. Geographic Variations in Mortality from Motor Vehicle Crashes in Taiwan, Journal of Trauma-Injury Infection & Critical Care. 43(1): 74-77