Analysis of occupational accidents in Spain using shrinkage regression methods

V. Gallego[a], A.I. Sánchez[b], I. Martón[a] , S. Martorell[a]*

[a]Department of Chemical and Nuclear Engineering. MEDASEGI research group, Universitat Politècnica de València, Valencia, Spain

[b]Department of Statistics and Operational Research. MEDASEGI research group, Universitat Politècnica de València, Valencia, Spain

* Corresponding author: S. Martorell (smartore@iqn.upv.es)

Abstract

This paper analyses the relationship between the evolution of occupational health indicators, i.e. frequency rate (FR), severity rate (SR) and days lost per accident (DLA), and the performance of relevant influencing factors related to the labour market, the productivity structure and the economy using regression shrinkage methods using Lasso, Elastic Net and AdaLasso regression methods. The advantage of applying these methods is that they overcome two common problems faced in this type of analysis: the number of input factors greatly exceed the number of observations (data set) and the multicollinearity of the input factors.

A case study is presented focused on occupational accidents in Spain in the time period 1995-2017. The analysis covers an instable period in Spanish labour market consisting of stages of economic growth and deep recession, as well as changes in the productive structure. The shrinkage methods permitted the identification and selection of the most important factors that significantly affected each occupational health indicator. In addition, the AdaLasso method provided the best results for FR y DLA, based on the BIC measure, and performed slightly worse than Elastic Net for SR. It can be concluded that AdaLasso seems to be the most coherent and robust method, which also explains the different main relationships with the least number of factors (variables).

Thus, based on the results found by the AdaLasso regression method, factor hours worked showed a strong and positive relationship with all three indicators. In addition, factors

employment total females and agriculture sector showed a strong and negative relationship with FR and SR. However, several factors affecting either FR or DLA seemed affect SR in no way, while others were affecting in the same way but with weaker or stronger relationship to each other. The results found not only were coherent with the results concerning health indicators analysed in previous work but also provided additional insights concerning non-covered occupational indicators up to now.

Keywords:

Accidents, regression methods, Lasso, AdaLasso, Elastic Net, occupational health indicators, influencing structural factors

LIST OF ACRONYMS

BIC	Bayesian Information Criterion
DL	Days Lost
DLA	Days Lost per Accident
FAR	Fatal Accident Rate
FR	Frequency Rate
GLM	Generalized Linear Models
ILO	International Labour Organization
NA	Number of Accidents
NH	Number of Hours worked (hours-worker)
OHS	Occupational health and safety
R	Risk
SR	Severity Rate

1. Introduction

Occupational accidents have serious adverse effects on workers and companies, affecting their productivity and widely impacting the economy. Each year 78 million workers die from occupational accidents and 374 million workers suffer from non-fatal occupational accidents according to International Labour Organization (ILO) (ILO, 2019). It is estimated that the economic cost of work-related injuries and illnesses vary between 1.8 and 6% of Gross Domestic Product each year, the average being 4%. So, the analysis and forecasting of the evolution of occupational accidents is a subject of concern and relevance for the society in general, mainly because it can allow for realizing the impact of different factors (economics, labour market, productivity structure, etc) on the increase or decrease of occupational accidents and the need to adjust work safety policies to improve or seek alternative policies.

In the literature, there is a significant number of papers analysing the effects of different factors on occupational health. Some are focused on the effect of economic variables on occupational accidents. Many of these papers use econometric models that attempt to explain the evolution of occupational health indicators in terms of economic performance. For example, (Boone van Ours et al. 2011) and (Livanos & Zangelidis 2013) suggested a relationship between accidents rate and the state of economic activity. The Granger causality test and co-integration test and an impulse response function based on the vector auto-regression model were used by Song (Song et al. 2011) to analyse the relation between economic development and occupational accidents in China. They concluded that the variation of economic speed had an influence on occupational accidents. Therefore, it seems justified that occupational safety policies should be planned in a coherent manner and considering the economic cycle. Additionally, there is a significant number of papers which analyse the effects of regulation on occupational health (Saphiro et al., 2000).

Other studies have analysed the influence of the productivity structure on the occurrence of occupational accidents. For example, a relationship between occupational injuries and firm size in the Italian industry was found by (Fabiano et al. 2004). The authors concluded that it exists an inverse correlation between occupational accidents frequency index and firm size. This relationship was more evident between fatal accidents frequency rate and firm size. A study is presented in (Fernandez Muñiz et al., 2018) that shows that the number of accidents in Spain tends to concentrate in certain sectors. The results of this

study demostrates that occupational accidents happen more frequently in sectors like construction, industry and agriculture, where more fatal accidents took place, especially in small and medium companies. Several studies have focussed on specific productive sectors, like industry (Carrillo-Castrillo et al., 2012, Altunkaynak, 2018), construction (Irumba, 2014, Kanchana et al. 2015¸ Segarra et al. 2017, Benavides et al., 2003) and agriculture (Kumar and Dewangan. 2009).

Other influencing factors analysed in the literature are the ones related with personal and occupational characteristics such as gender, age, education level or occupational status. According to (Villanueva & Garcia, 2011), the risk of a fatal workplace accident increases with age and hours of work, and is higher for male workers and temporary workers. Additionally, (Bhattacherjee et al., 2003, García & Monuenga, 2009, Eurostat. 2019) found similar results. These authors identified the gender, job category and age as the most influencing factors for occupational injuries. Some papers have focused on analysing the effect of personal and occupational characteristics in these sectors (Dimich-Ward, et al., 2004, Gauchard et al., 2003 and Dimichet al., 2004).

The objective of the work presented in this paper is to analyse the relationship between different occupational health indicators and influencing factors related to the labour market, the productivity structure and the economy using regression shrinkage methods. In particular, Lasso, Elastic Net and AdaLasso methods are applied. The advantage of applying these methods as compared to previous work is that they overcome two common problems in this type of analysis, as found in previous studies: the number of input variables (input factors) greatly exceed the number of observations (input data set) and the multicollinearity of the input factors. In addition, these methods make a selection of variables, i.e. factors, identifying the most important ones that significantly affect occupational health indicators. The case study focused on Spain in the time period 1995-2017. The analysis covers an instable period in Spanish labour market consisting of stages of economic growth and deep recession, as well as changes in the productive structure.

The paper is structured as follows. Section 2 presents a brief review of the occupational health indicators considered in this paper and their relationship with the risk components. Section 3 describes the factors considered in this paper and the corresponding dataset used in this work. Section 4 presents the different statistical tools and regression methods applied to each health indicator addressing the whole set of input factors and data. Section

5 reports the application results. Section 6 presents a discussion of the results found and, finally, Section 6 provides the conclusions.

2. Occupational health indicators and risk

Occupational health and safety (OHS) measurement relies heavily on lagging indicators, such as incidents of workplace injury, as these measures provide important feedback information about deficiencies and safety incidents that have occurred. Lagging OHS indicators are a reactive measurement approach to safety management and measure events or outcomes that have already take place.

The ESAW methodology (Eurostat, 2001) considers different indicators to monitor occupational health in accordance with the ILO resolution of 1998 concerning "Statistics of Occupational Injuries resulting from Occupational Accidents", such as incidence rate (*IR*), frequency rate (*FR*), severity rate (*SR*) or fatal accident rate (*FAR*), among others (ILO, 1998).

These indicators are related to risk and its components (Martorell et al., 2016a). Thus, the risk (*R*) of an accident at work can be formulated in terms of two components, the frequency and the damage, as follows:

$$R = F \cdot D \tag{1}$$

where *F* is the frequency of occurrence of the work accident and *D* represents the damage, which is a measure of the severity of the occupational injury resulting from the work accident.

For example, the frequency rate (*FR*) indicator is related with risk component *F*, which can be estimated as the ratio between the number of accidents (NA) and the number of hours worked (hours-worker) (NH) in a year, as follows:

$$FR = \frac{NA}{NH} \cdot 10^6 \tag{2}$$

In addition, the mean number of days lost per accident (*DLA*) indicator is related with component *D*, which can be estimated as the ratio between the days lost (DL) and the number of accidents in a year as follows:

$$DLA = \frac{DL}{NA} \tag{3}$$

Furthermore, the severity rate (SR) indicator is related with risk, i.e. social risk, which can be estimated either directly from data or indirectly using the *FR* and *DLA* indicators as follows:

$$SR = \frac{DL}{NH} \cdot 10^3 = FR \cdot DLA \cdot 10^{-3} \quad (4)$$

In this study, the *FR*, *DLA* and *SR* indicators are considered based on their relationship with risk and its components. The objective is to analyze the relationship between these indicators and different factors related to the labor market, the productive structure and the economy using the data set and the regression methods introduced in the following sections.

3. Output variables, input factors and data sets

The data set used in this work for the different variables (input or output to the regression models) comes from three main sources: Sub-directorate General for Statistics of the Ministry of Employment and Social Security in Spain, statistical office of the European Union (EuroStat) and International Labour Organization (ILO). The data set has an annual periodicity and covers a period of 23 years from 1995.

Table 1 shows the input explanatory variables considered, including factors related to the labour market, the productive structure and the economy, indicating the source from which the different data have been obtained.

Table 1. Input factors (input explanatory variables).

Group	Subgroup	Variable	Description	Source
Labour market	Sex	V1	Employment Total Females (10^3)	Eurostat
		V2	Employment total Males (10^3)	
	Age	V3	From 16 to 24 years (10^3)	Eurostat
		V4	From 25 to 49 years9 (10^3)	
		V5	50 years or over (10^3)	
	Education	V6	Less than primary (levels 0-2) (%)	ILO
		V7	Upper secondary (levels 3-4) (%)	
		V8	Tertiary education (levels 5-8) (%)	
	Work	V9	Part time (%)	ILO
		V10	Self-employment (10^3)	
		V11	Temporary employment (%)	
		V12	Hours worked	
Productive structure	Category	V13	Managers (10^3)	Eurostat
		V14	Professionals (10^3)	
		V15	Technicians and associate professionals (10^3)	
		V16	Service and sales workers (10^3)	
		V17	Skilled agricultural, forestry and fishery workers (10^3)	
		V18	Craft and related trades workers (10^3)	
		V19	Plant and machine operators and assemblers (10^3)	
		V20	Elementary occupations (10^3)	
	Sector	V21	Agriculture	Eurostat
		V22	Industry	
		V23	Construction	
		V24	Service	
Economy	GDP	V25	Gross Domestic Product per capita (10^6)	Eurostat
	Unemployment	V26	Unemployment rate (%)	ILO

(1) Eurostat https://ec.europa.eu/eurostat/data/database (2) ILO (International Labor Organization). https://www.ilo.org/ilostat

The evolution of these explanatory variables in the time period considered in the study are shown in Figures 1, 2 and 3.

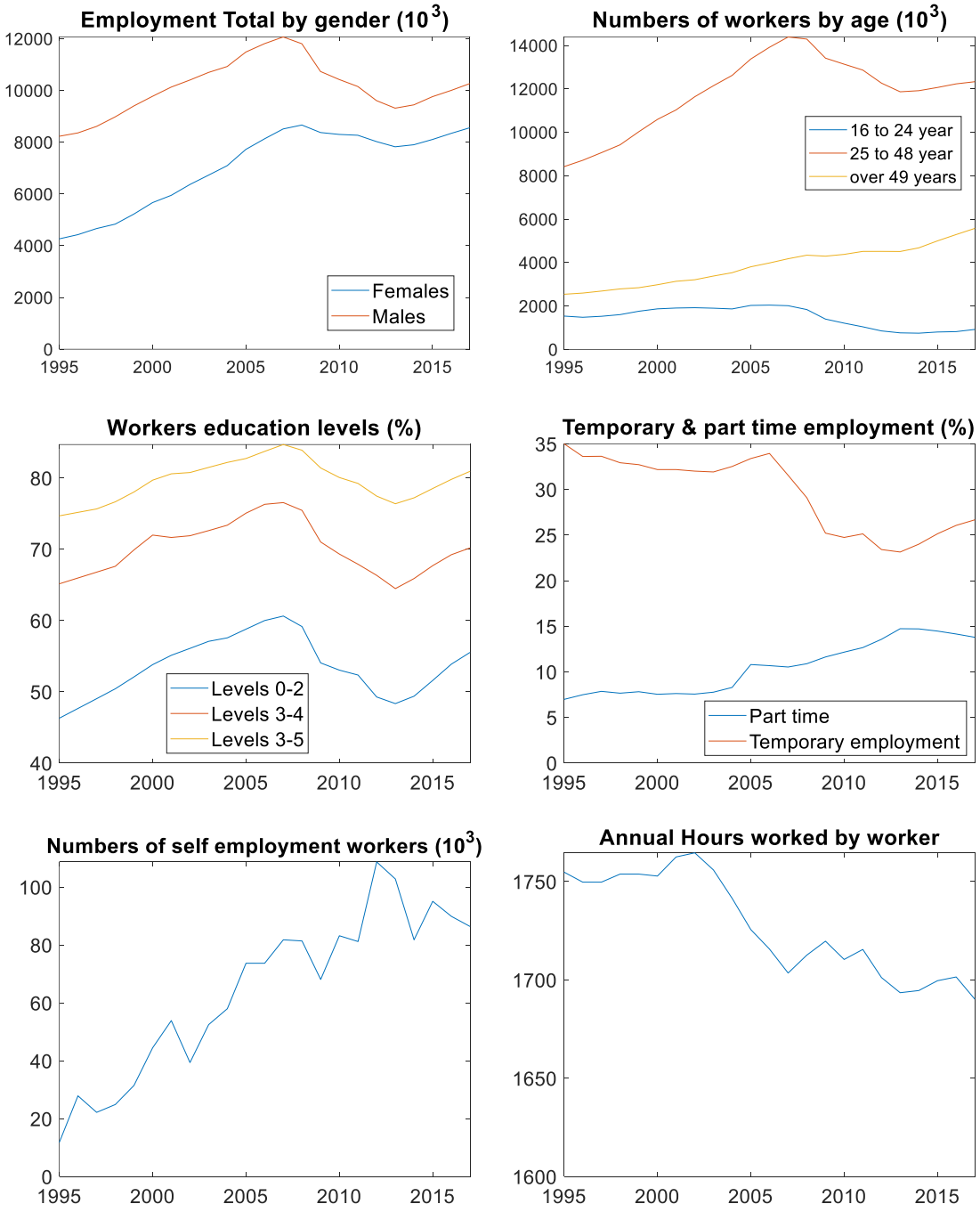
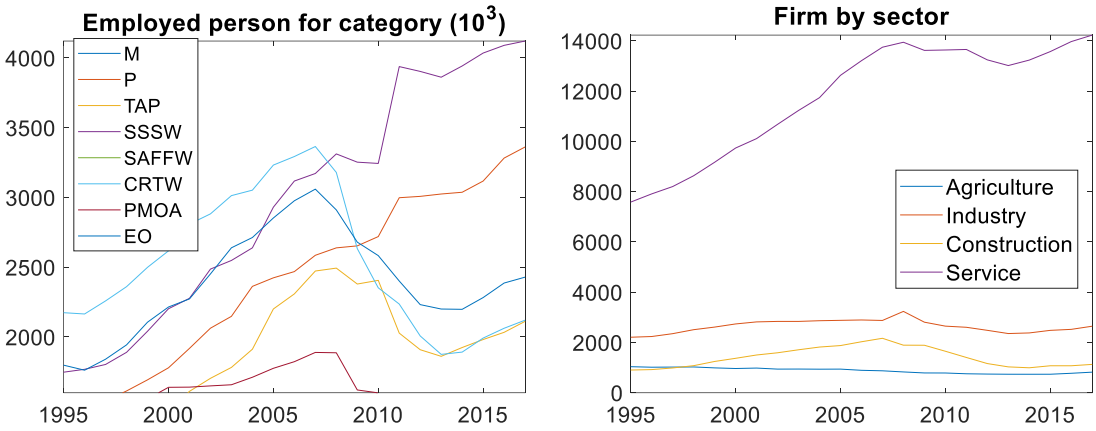


Fig. 1. Time evolution of variables belonging to the labour market group.



M: Managers, P=Professionals, TAP: Technicians and associate professionals, SSW: Service and sales workers, SAFFW: Skilled agricultural forestry and fishery workers, CRTW: Craft and related trades workers, PMOA: Plant and machine operators and assemblers, EC: Elementary occupations.

Fig. 2. Time evolution of variables belonging to the productive structure group.

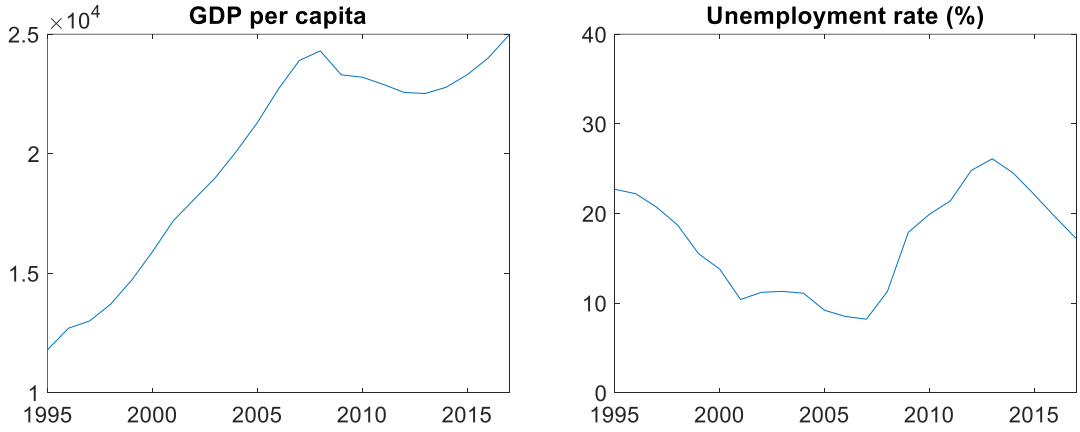


Fig. 3. Time evolution of variables belonging to the economy group.

The output variables considered for each regression model are the occupational health indicators introduced in section 2: The frequency rate, the number of days lost per accident and the severity rate. Annual data related to these indicators have been obtained from the Sub-Directorate General for Statistics of the Ministry of Employment and Social Security in Spain. Figure 4 shows the evolution of the three indicators in the period 1995-2017 in Spain.

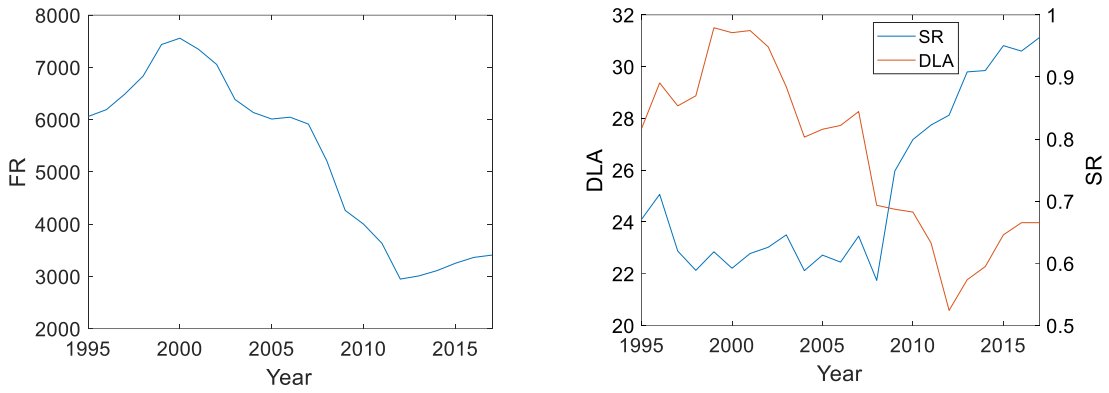


Figure 4. Frequency Rate (*FR)*, mean number of Days Lost per Accident (*DLA)* and Severity Rate (*SR)* evolution in Spain (1995-2017).

Figure 4 shows that the three occupational health indicators present fluctuations in the period analyzed. The evolution of the FR begins with an increase until year 2000 where a maximum is reached. Subsequently, the indicator decreases with a slight increase in 2006. The minimum in the time series is observed in 2012, the year in which an increase in the indicator begins until 2017.

Regarding the DLA, two stages are distinguished, the first one, until 2008, in which the indicator fluctuates around a value of 21 days lost per case of occupational injury, and a later stage, in which the indicator grows to reach a maximum value in 2017 with a value of 31 DLA.

The SR presents different stages of growth and decrease, reaching a maximum of 0.98 days lost per thousand hours worked in 1999. Next, from 2001 the indicator decreases until 2004 when a growth stage begins which remains until 2007 when the SR begins to decrease reaching a minimum in 2012 with a value of 0.59 days lost per thousand hours worked. Finally, from 2012 to 2017 the indicator shows continuous growth.

4. Statistical methods

4.1. Pre-processed data

A specific feature of regression models is that they can only work with stationary data. If the process is stationary the statistical properties (mean, variance and covariance) does not change over time.

Time series can be non-stationary (see, for example, Figure 4) and if neglected can cause the problem of spurious regression. To avoid this, prior to modelling, the annual growth rates (i.e., the change in the value of a measurement over the period of a year) were calculated for the explanatory and the output variables.

4.2. Shrinkage regression methods

In the literature, Generalized Linear Models (GLM) have been commonly used to analyse the effect of different explanatory variables in occupational health indicators (Martorell et al. 2016b). GLMs can present problems of bias and over-dispersion specially when the explanatory variable number (p) is greater than the sample size (n). In this situation, the estimator is unstable in the sense that there is unacceptable growth in the variance .

An alternative to GLM are shrinkage methods which have the advantage of exhibiting less variance than least squares estimates. In addition, some shrinkage methods also reduce the number of covariates included in the regression model by yielding coefficient estimates of exactly zero, facilitating the model selection process. In this paper different shrinkage methods are used: Lasso, Elastic Net regression and AdaLasso regression (Tibshirani, 1996, Zou 2006).

Lasso regression was introduced by Tibshirani (Tibshirani, 1996). Lasso is a shrinkage and variable selection method for linear regression models. Lasso estimates are obtained by minimizing the sum of squared residuals subject to a penalty of L_1 norm of the coefficients:

$$\min_{\beta_0,\beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \quad (5)$$

subject to

$$\|\beta\|_1 \leq t \quad (6)$$

where y_i are the values of the independent variable, x_{ij} represents the covariates, β_j are the corresponding coefficients, $\| \quad \|_1$ the L_1 norm and $t \geq 0$ is the tuning parameter which controls how much penalty is applied on the set of coefficients.

The Lagrangian formulation of the Lasso regression is given by:

$$\min_{\beta_0 \beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \|\beta\|_1 \qquad (7)$$

where $\lambda \geq 0$ is called the regularization parameter which is function of the parameter t. Problem solution depend on the value selected for λ. When $\lambda = 0$ the problem is simplified to the ordinary least square estimator. The increase of the λ value causes the contraction of the estimator of the β parameter. Lasso results in a model where some coefficient estimates are equal to zero if λ is sufficiently large, this has a positive effect on model interpretability since it makes a selection of variables.

Lasso regression can have very poor performance when there are highly correlated variables in the predictor set. Collinearity can degrade the performance of the Lasso (Zou and Hastie, 2005). (Zou & Zhang 2009) proposed an improved version of the Lasso named Elastic Net which is a technique uses to perform variable selection and coefficient estimation that combines the penalties of the Lasso and Ridge methods. Ridge method was introduced by Hoerl (Hoerl, 1962). The penalty on the coefficient vector β_j imposed by Ridge is slightly different from Lasso. In case of Ridge, the λ parameter is multiplied by the L_2 norm of the vector β. Based on ridge and Lasso methods (Zou & Hastie, 2005) proposed the Elastic Net regression whose Lagrangian formulation is given by:

$$\min_{\beta_0 \beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda_2 \|\beta\|_2 + \lambda_1 \|\beta\|_1 \qquad (8)$$

Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ then the Elastic Net penalty can be written as $\alpha\|\beta\|_2 + (1-\alpha)\|\beta\|_1$, which is a combination of the Lasso and ridge penalty. When $\alpha = 1$ the Elastic Net becomes ridge regression.

Lasso selects groups of variables according to their correlation, therefore, from which a pair of correlated variables will choose one of the two, regardless of whether the rejected

variable represents the data set better; Elastic Net will keep the two, implicitly choosing the variable with the most contribution so that the model fits the data better.

Other drawback of the Lasso method is its instability with high-dimensional data. So, this method could be inconsistent for model selection unless the predictor matrix satisfies a rather strong condition (Zou, 2006). In this context, Zou proposed the adaptive Lasso (AdaLasso) estimator:

$$\min_{\beta_0 \beta_j} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \widehat{w_j} \left|\beta_j\right| \quad (9)$$

where $\widehat{w_j}$ are the adaptive data-driven weights. The weights can be estimated as $w_j = \frac{1}{\widehat{\beta_j}}$ being $\widehat{\beta_j}$ some initial estimates of the β using, for example, Lasso.

An important issue in the above methods is the selection of a proper λ values, that is, to find a parameter value that ensures a proper balance between bias and variance (or flexibility and interpretability). The λ value can be determined by data-driven techniques such as cross-validation or information criteria. In a time-series context, the Bayesian Information criterion (BIC) is a reliable alternative to select the parameter λ (Medeiros and Mendes, 2015).

Bayesian Information Criterion is defined as:

$$BIC = -2 \cdot \log(L) + p \cdot log(n) \quad (10)$$

where L is the likelihood and p the number of covariates. This metric considers the adjustment and it counteracts the excess of parameters. The objective is to find a model with a small BIC value in this case the likelihood is big and the number of parameters small.

The optimization problems showed in equations (7) to (9) are standard quadratic program with convex constraints, and there are a variety of numerical methods to solve it.

5. Results

This section presents the results obtained from the application of Lasso, Elastic Net and AdaLasso regression methods to build the corresponding regression models for the three occupational health indexes FR, DLA and SR as a function of the input explanatory variables showed in Table 1 using the approaches introduced in section 4. The estimation process of the coefficients of inputs variables for each occupational health index has been developed using the R package glmnet (Hastie & Tibshirani, 2008).

In this application, the number of potential covariates, p, is equal to twenty six excluding the intercept, and the number of observations, n, is twenty two.

To build the Elastic Net model a range of α-values (0.25, 0.5,0.75) has been screened in order to select the best model. Finally, $\alpha = 0.5$ was chosen, which gives equal weight to Ridge and Lasso regressions and is associated with the lowest error in this case. So, the final Elastic Net model built benefits from both the selection capability of Lasso regression and the ability of Ridge regression to handle multicollinearity in the dataset. The optimum λ parameter was determined based on BIC.

Table 2 shows the estimated coefficients in the different regression models obtained. Figures 5, 6 and 7 show the same information where it can be observed more clearly the coefficients with a higher absolute value and if the relationship with occupational health indicators is positive or negative.

Table 3 shows the BIC values obtained for each regression model. Based on the BIC, the best model is the one with the smallest value. Thus, in the case of FR and DLA the best model is the AdaLasso while for SR the best model is the Elastic Net.

An analysis of the residuals was carried out to verify their normality and homoscedasticity. In all cases it was observed that the residuals are distributed approximately normally and the variance is constant.

Table 2. Lasso, Elastic Net and AdaLasso coefficients estimated for each occupational health index and input factor based on occupational accidents occurred in Spain in the time period 1995-2017.

		DLA			Frequency Rate			Severity Rate		
N	Variable*	Lasso	Elastic Net	AdaLasso	Lasso	Elastic Net	AdaLasso	Lasso	Elastic Net	AdaLasso
	Constant	3.08E-02	1.67E-02	3.38E-02	8.70E-03	1.69E-02	3.07E-02	2.15E-02	3.06E-02	4.15E-02
1	Employment Total Females				-8.57E-01	-8.40E-01	-1.33E+00	-1.17E+00	-1.21E+00	-2.47E+00
2	Employment Total Males									
3	From 15 to 24 years	1.92E-01		2.13E-01		1.49E-01		5.57E-01	5.39E-01	1.10E+00
4	From 25 to 49 years				-5.86E-01	-2.99E-01		-2.36E-01	-5.07E-01	
5	50 years or over				-1.05E+00	-1.09E+00	-1.21E+00	-2.50E-01	-6.10E-01	
6	Employment by education (0-2)				5.03E-02	2.93E-02	4.22E-02	7.58E-02	4.84E-02	5.20E-02
7	Employment by education (3-4)		-1.30E-03			3.89E-04		-7.92E-03		
8	Employment by education (5-8)					1.19E-02			9.74E-03	
9	Part time(%)				3.19E-02	2.11E-02	2.75E-02	6.40E-02	5.00E-02	4.97E-02
10	Self-employment (Thousand)				-9.50E-04	-8.73E-04	-7.28E-04	-6.52E-04	-5.06E-04	-2.06E-04
11	Temporary employment	-1.17E-02	-3.78E-03	-9.36E-03	-4.07E-03			-3.46E-02	-2.02E-02	-2.30E-02
12	Hours worked	2.39E+00		3.25E+00	1.92E-01	2.52E-01	4.72E-01	5.37E+00	4.68E+00	6.95E+00
13	Managers				-2.84E-02	-1.83E-02			-7.75E-02	
14	Professionals	-1.24E-02				8.75E-02			1.77E-02	2.27E-01
15	Technicians and professionals				1.70E-01	1.46E-01	1.71E-01			
16	Service and sales workers					-1.15E-02		-3.33E-01	-2.47E-01	-1.57E-01
17	Skilled agricultural, forestry and fishery workers	1.61E-01		1.57E-01	-9.90E-02	-2.83E-01		6.46E-01		
18	Craft and related trades workers				-3.99E-01	-2.35E-01		-7.28E-01	-4.91E-01	-3.98E-01
19	Plant and machine operators and assemblers		-4.98E-02		4.22E-01	3.42E-01	4.69E-01	3.50E-01	3.41E-01	5.60E-01
20	Elementary occupations									
21	Agriculture				-5.86E-01	-3.65E-01	-4.57E-01	-1.87E+00	-1.08E+00	-1.46E+00
22	Industry	-1.02E+00	-4.70E-01	-1.11E+00				-1.48E+00	-1.13E+00	-1.63E+00
23	Construction				2.39E-01	1.04E-01	1.86E-01	7.13E-02	1.59E-01	
24	Service									
25	GDP per capita nama_10_GDP				-2.77E-02	-9.29E-02			2.63E-01	3.92E-01
26	Unemployment rates (%)				-1.37E-02	-1.21E-02	-5.18E-03	-2.47E-02	-1.51E-02	-4.61E-03

* Annual Growth Rates t-1

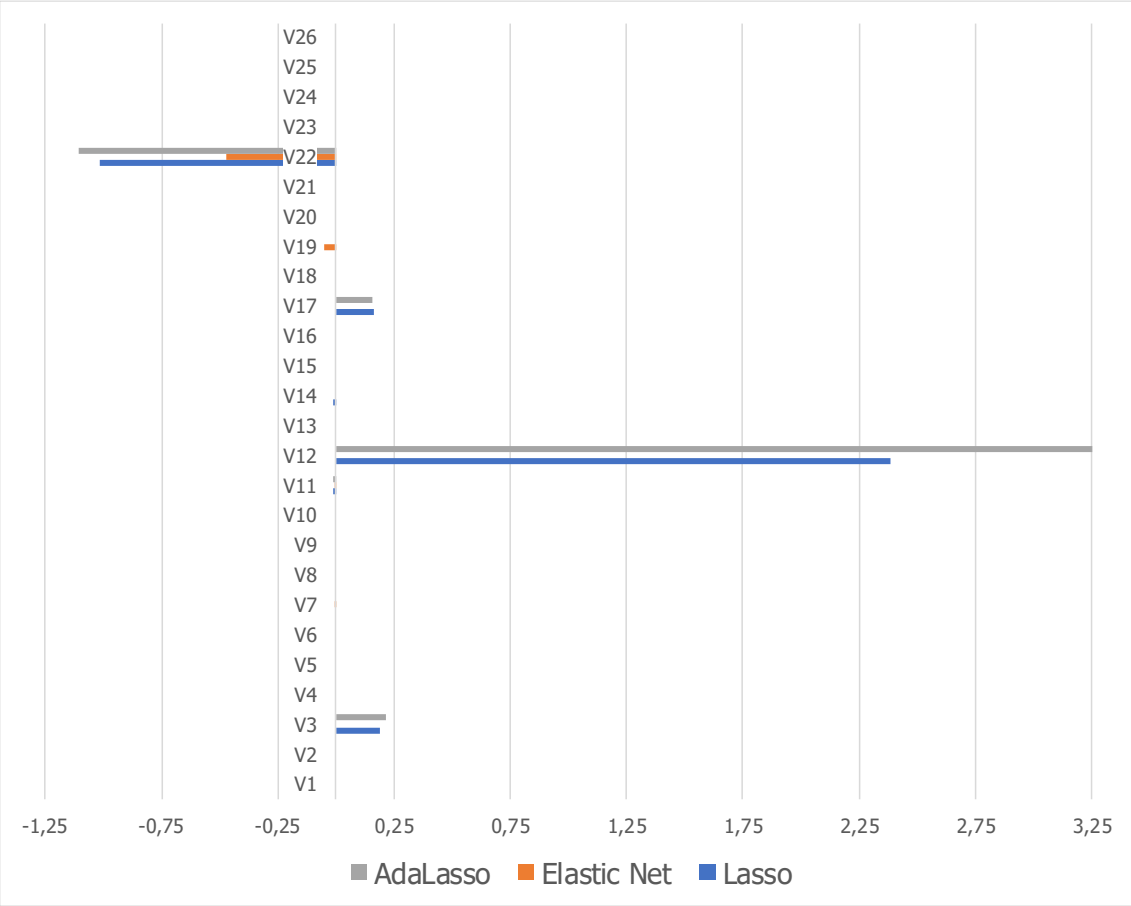


Figure 5. Estimated coefficients using Lasso, Elastic Net and AdaLasso regression methods for mean number of days lost per accident based on occupational accidents occurred in Spain in the time period 1995-2017.

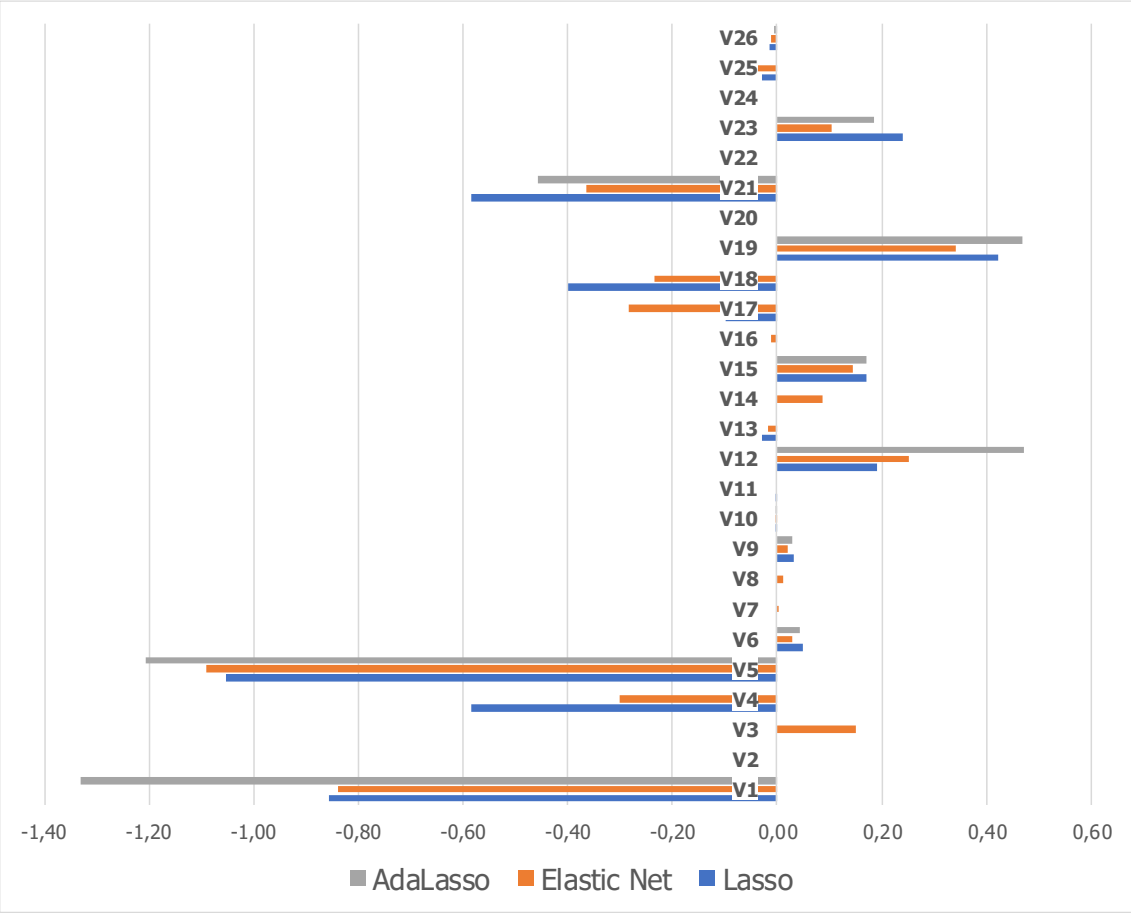


Figure 6. Estimated coefficients using Lasso, Elastic Net and AdaLasso regression methods for frequency rate based on occupational accidents occurred in Spain in the time period 1995-2017.

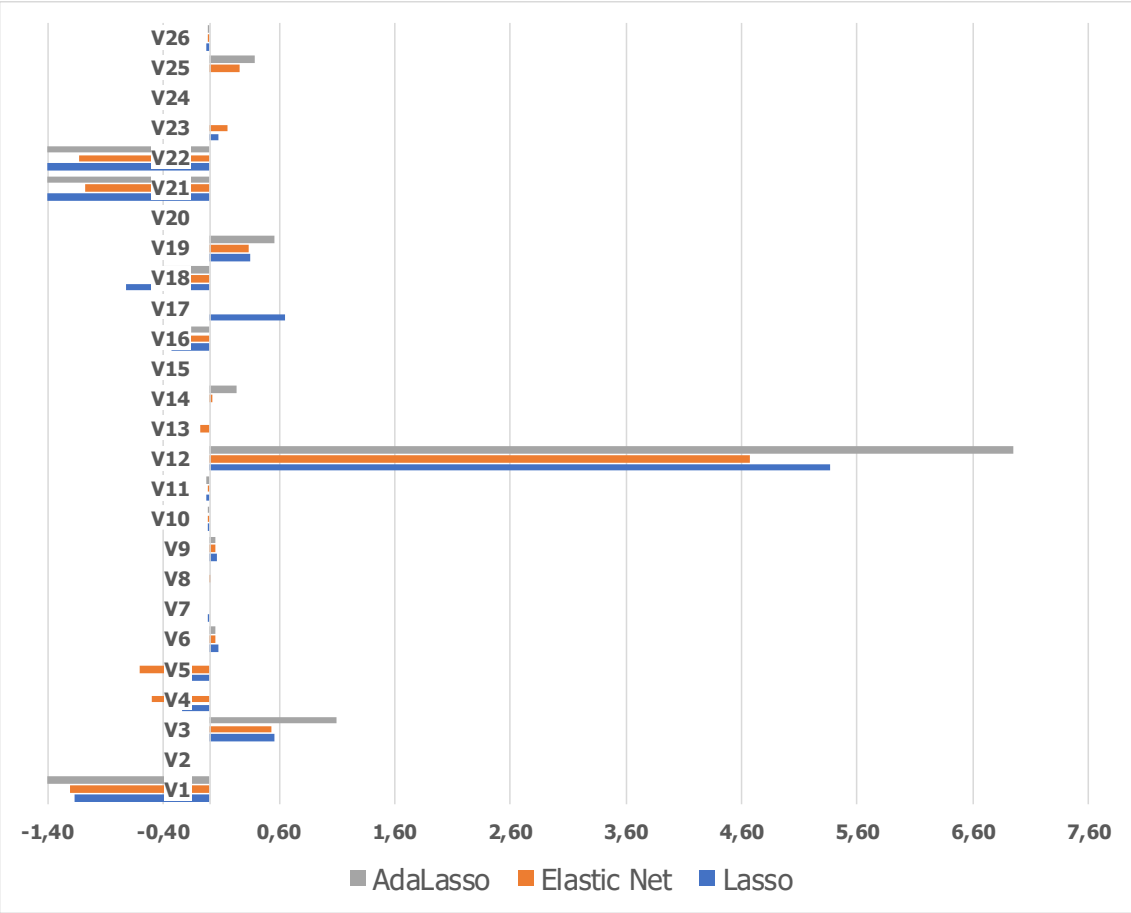


Figure 7. Estimated coefficients using Lasso, Elastic Net and AdaLasso regression methods for severity rate based on occupational accidents occurred in Spain in the time period 1995-2017.

Table 3. Bayesian Information Criteria (BIC) obtained for the different regression models.

	Frequency Rate (*FR*)	Mean number of days lost per accident (*DLA*)	Severity Rate (*SR*)
Lasso	-45.80	-52.39	-43.20
Elastic Net	-142.56	-141.26	-149.34
AdaLasso	-162.20	-148.90	-138.12

6. Discussion

Table 2 shows that all three Lasso, Elastic net and AdaLasso methods reduce the dimensionality of the problem significantly in all cases analyzed. This reduction in dimensionality is observed, mainly, in the regression models of the annual growth rate of the DLA, in which no more than six variables of the original set of twenty-six are significant.

In the following subsections the results obtained for the variables that show a significant correlation with the annual growth rate of the different occupational health indicators analyzed are discussed.

6.1 Annual growth rate of days lost per accident

In “Labour market” group, within the subgroup “Age”, there is a modest but positive relationship between the annual growth rate of the *DLA* and the annual growth rate of number of employees in the age range from 15 to 24 years (V3) according to the Lasso and AdaLasso regression models. Within the “Education” subgroup, only the Elastic Net method shows a negative but weak relationship with the annual growth rate of employment with education levels 3-4 (V7). Within the subgroup “Work”, all three regression methods detect a negative but weak relationship with the annual growth rate of temporary employment (V11). In addition, within the same subgroup, the Lasso and AdaLasso methods show a positive and strong relationship with the annual growth rate of the number of hours worked by year (V12).

In the “Productive structure” group, within the “Category” subgroup, there is a positive but modest relationship between the annual growth rate of the *DLA* and the annual growth rate of skilled agricultural, forestry and fishery workers (V17) according to the Lasso and AdaLasso regressions. The Lasso regression shows a negative but weak relationship with the growth rate of professionals (V14). In addition, the Elastic Net regression shows a negative but weak relationship with the growth rate of "Plant and machine operators and assemblers" (V19). Within the “Sector” subgroup, all three regression methods show a negative and strong relationship between the annual growth rate of the *DLA* and the annual growth rate of companies in the “Industry” sector (V22).

In the “Economy” group, none of the three methods shows that there is a significant relationship between economic factors and *DLA*.

6.2 Annual growth rate of the frequency rate

What concerns the annual growth rate of the frequency rate (FR), it is observed that, in general, in all groups and in most subgroups, there is at least one variable that participates in the regression models, with a positive or negative correlation. This means that, apparently, in all three regression models all subgroups are significant to explain the evolution of the annual growth rate of the frequency rate. This general conclusion needs to be analyzed in more detail.

In the “Labour Market” group, within the “sex” subgroup, there is a negative and strong relationship between the annual growth rate of the *FR* and the annual growth rate of the employed female (V1) according to all three regression models. In the same group, in the “Age” subgroup, there is a negative and strong relationship with the annual growth rate of the number of workers in the group 50 years or over (V5) according to all three models too, being similar for the subgroup 25-49 years (V4) according to Lasso and Elastic Net models only. On the contrary, the Elastic Net method is the only one that shows a positive but modest relationship with the annual growth rate of the number of employees of the age group from 15 to 24 years (V3). Within the “Education” subgroup, there is a positive and modest relationship with the annual growth rate of the number of employees with education less than primary (levels 0-2) (V6) according to all the regression models. Within the subgroup “Work”, all the methods show a positive and strong relationship with the annual growth rate of hours worked (V12) and a positive but modest relationship with the annual growth rate of part-time employment (V9). In addition, all three models show a negative but weak relationship with the annual growth rate of self-employment (V10). Only the Lasso model shows a negative but weak relationship with the annual growth rate of temporary employment (V11).

In the “Productive structure” group, within the subgroup “Category”, all three regression models show a positive and strong relationship with the annual growth rate of the number of contracts for technicians and associate professionals (V15) and a positive and strong relationship with the annual growth rate of plant and machine and assembly operators (V19). In addition, the Lasso and Elastic Net regression models show a negative and

strong relationship with the annual growth rate of the number of Skilled agricultural, forestry and fishery workers (V17) and a negative and strong relationship with the annual growth rate of craft and related trades workers (V18). Both also shows a negative but weak relationship with the annual growth rate of managers (V13). The Elastic Net model is the only one that shows a positive but modest relationship with the annual growth rate of professionals (V14). Within the " Sector" subgroup, all three regression models show a positive and strong relationship with the annual growth rate of the number of contracts in the construction sector (V23) and a negative and strong relationship with the annual growth rate of contracts in the agriculture sector (V21).

In the "Economy" group, all three regression models show a negative and modest relationship with the annual growth rate of the unemployment (V26). The Lasso and Elastic Net regression models show a similarly negative but weak relationship with the annual growth rate of GDP per capita (V25).

6.3 Annual growth rate of the severity rate

What concerns the annual growth rate of the severity rate (SR), it is observed that, in general, this occupational health indicator performs qualitatively in a similar way than the FR indicator for all groups and in most subgroups. This means that, apparently, in all three regression models all subgroups are significant to explain the evolution of the annual growth rate of the severity rate. This conclusion is analyzed in more detail in a quantitative way in the following paragraphs.

In the "Labour Market" group, within the "sex" subgroup, there is a negative and strong relationship between the annual growth rate of the *SR* and the annual growth rate of the employed female (V1) according to all three regression models. In the same group, in the "Age" subgroup, there is a negative but modest relationship with the annual growth rate of the number of workers in the subgroup 50 years or over (V5) and with the subgroup 25-49 years (V4) according to Lasso and Elastic Net models only. On the contrary, the all three regression models show a positive and strong relationship with the annual growth rate of the number of employees of the age group from 15 to 24 years (V3). Within the "Education" subgroup, there is a positive and weak relationship with the annual growth rate of the number of employees with education less than primary (levels 0-2) (V6) according to all the regression models. Within the subgroup "Work", all the models show

a positive and strong relationship with the annual growth rate of hours worked (V12) and a positive but weak relationship with the annual growth rate of part-time employment (V9). On the contrary, all three models show a negative but weak relationship with the annual growth rate of self-employment (V10) and the annual growth rate of temporary employment (V11).

In the “Productive structure” group, within the subgroup “Category”, all three regression models show a positive and modest relationship with the annual growth rate of the number of contracts for plant and machine and assembly operators (V19). In addition, all three models show a negative but modest relationship with the annual growth rate of craft and related trades workers (V18) and with the annual growth rate of service and sales workers (V16). The AdaLasso model is the only one that shows a positive but modest relationship with the annual growth rate of professional workers (V14). Within the " Sector" subgroup, all three regression models show a negative but strong relationship with the annual growth rate of the number of contracts in the agriculture sector (V21) and the industry (V22). On the contrary, the Lasso and Elastic Net models show a positive and modest relationship with the annual growth rate of the number of contracts in the construction sector (V23).

In the “Economy” group, all three regression models show a negative and weak relationship with the annual growth rate of the unemployment (V26). The Elastic Net and AdaLasso regression models show a similarly positive but modest relationship with the annual growth rate of GDP per capita (V25).

7. Conclusions

This paper analysed the relationship between different occupational health indicators (FR, DLA, SR) and factors related to the labour market, the productive structure and the economy using Lasso, Elastic Net and AdaLasso regression methods, which allow to overcome two common problems in this type of analysis when the number of input variables (input factors) greatly exceed the number of observations (input data set) and the multi-collinearity between input variables. In addition, these methods allowed to identify the most important factors that significantly affect health indicators as they made a selection of variables.

The case of application of such regression methods focused on occupational accidents in Spain during the time period 1995-2017 and it has been shown the AdaLasso method provided the best results for FR and DLA, based on the BIC measure, and performed slightly worse than Elastic Net for SR. It can be concluded that AdaLasso seems to be the most coherent and robust method, which also explains the different relationships with the least number of factors (variables). It not only allowed to explain the evolution of health indicators in the past but also it would allow the prognosis of their future evolution based on the monitoring of the effect of the most significant factors. This way, compensatory measures could be analyzed in the light of such principal relationships with an aim to study changes within occupational health policies that minimize negative and maximize positive impacts of relevant structural factors on the evolution of occupational indicators.

Thus, based on the results found adopting the AdaLasso regression method, in two of the three main groups, i.e. labour market, productive structure, at least one factor was found with strong and either positive (hours worked, technicians and machine operators, construction sector) or negative (female sex, age group older than 50, agriculture sector) relationship with the annual growth rate of the FR indicator. In the economy group, unemployment factor showed negative, but moderate relationship with FR. The factors related to female sex, hours worked and agriculture sector showed similar relationship with the annual growth rate of the SR. An age group older than 50 (negative) and machine operators (positive) showed a moderate but a similar relationship with SR. In addition, an age group between 15-24 and industry sector showed a strong and positive relationship with SR. Also, the GDP showed a moderate but positive relationship with SR. In addition, SR indicator showed weak but meaningful relationships with many others factors, like those belonging to the type of job subgroup and the unemployment factor. However, only

factors hours worked (positive) and industry sector (negative) showed a strong relationship with the annual growth rate of the DLA. In addition, age group between 15-24 and agriculture sector showed a positive but moderate relationship with DLA.

In summary, factor hours worked showed a strong and positive relationship with all three indicators. In addition, factors employment total females and agriculture sector showed a strong and negative relationship with FR and SR. However, although SR depends on the combination of FR and DLA, it is not clear from the results that, in general, factors affecting FR and/or DLA are only the ones affecting the SR. On the contrary, several factors affecting either FR or DLA seemed affect SR in no way, while others were affecting in the same way but with weaker or stronger relationship to each other. Therefore, providing compensatory measures to mitigate or enhance the negative or positive effect, respectively, of the different factors one must account particularly the occupational health indicator being considered.

In is worthy to compare the conclusions found in this work with results reported in previous studies, but taking into account that most of previous work in this area focuses more on FR or FAR (e.g. Boone van Ours et al. 2011, Livanos & Zangelidis 2013, Fabiano et al. 2004) than in SR and DLA, the later often focussing more on particular sectors like construction and mining (Soltanzadeh, et a., 2017, Mirzaei M. et. 2019). However, in this paper the results found for FR and SR indicators are similar while different results both qualitatively and quantitatively were found for DLA indicator.

Thus, it is observed that the conclusions obtained for the FR indicator are similar to other previous works in what refers to the factors related to sex, the age of the workers or the sector (Fernandez Muñiz et al., 2018, Bhattacherjee et al., 2003.), which show these factors as relevant in the behavior of the FR. Likewise, the results are consistent with respect to the influence of the economic cycle on FR. In addition, also for the FR, this work showed that the "Self employment" factor, which did not appear in previous studies, also seems important for the FR, presenting a positive impact on the indicator. Regarding SR, other studies show, similar to the results of this study, that age and sex can be crucial factors in the severity rate (Soltanzadeh, et a., 2017) in mining and construction.

8. References

Altunkaynak, B. A statistical study of occupational accidents in the manufacturing industry in Turkey. International Journal of Industrial Ergonomics 66 (2018) 101-109.

Benavides, F.G., Giráldez, M.T., Castejón, E., Catot, N., Zaplana, M., Delclós, J., Benach, J., & Gimeno, D. (2003). Análisis de los mecanismos de producción de las lesiones leves por accidentes de trabajo en la construcción en España. Gaceta Sanitaria, 17(5), 353-359.

Bhattacherjee A., Chau N., Sierra C.O. et al. Relationships of job and some individual characteristics to occupational injuries in employed people: a community-based study. J Occup Health. 2003 Nov;45(6):382-91

De Leeuw, J. 1992. Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle. Pages 599-609 In: Kotz, S., and N.L. Johnson, editors. Breakthroughs in Statistics Volume 1. Foundations and Basic Theory. Springer Series in Statistics, Perspectives in Statistics. Springer-Verlag: New York.

Dimich-Ward, H., Guernsey, J.R., Pickett, W., Rennie, D., Hartling, L., Brison, R.J., 2004. Gender differences in the occurrence of farm related injuries. Occup. Environ. Med. 61, 52–56.

Eurostat, 2001. European Statistics on Accidents at Work (ESAW)- Methodology-2001 Edition. Directorate General for Employment and Social Affairs (DG EMPL), Statistical Office of the European Union (EUROSTAT), Luxemburg.

García, I. and Montuenga, V. (2009). "Causas de los accidentes de trabajo en Espaa analisis longitudinal con datos de panel". En: Revista aceta anitaria. Barcelona. Sociedad Española de Salud Pública y Administración Sanitaria (SESPAS). No. 3 (23). Pp. 174-178.

Gauchard, G.C., Chau, N., Touron, C., Benamghar, L., Dehaene, D., Perrin, Ph.P., et al.,

2003. Individual characteristics in occupational accidents due to imbalance: a case-control study of the employees of a railway company. Occup. Environ. Med. 60, 330–335.

Friedman,J.,Hastie,T.andTibshirani,R.(2008). Regularization Paths for Generalized Linear Models via Coordinate Descent, https://web.stanford.edu/~hastie/Papers/glmnet.pdf Journal of Statistical Software, Vol. 33(1), 1-22.

Gallego, V. Martorell, S. & Sánchez A.I. Analysis of economic and structural factors on occupational accidents using a generalized linear model: Example of application to Spain. "Risk and Reliability:Innovation theory and practice" CRC Press. (2016b), pp. 1668-1674.

Hoerl, A.E. (1962) Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress, 58, 54-59.

ILO, 1998. Resolution concerning statistics of occupational injuries (resulting from occupational accidents), adopted by the Sixteenth International Conference of Labour Statisticians. 6-15 October, Geneva, Switzerland.

Irumba R. Spatial analysis of construction accidents in Kampala, Uganda, Saf. Sci., 64 (2014), pp. 109-120.

Mirzaei Aliabadi, M., Aghaei, H., Kalatpuor, O., Soltanian, A. R., & Nikravesh, A. (2019). Analysis of the severity of occupational injuries in the mining industry using a Bayesian network. Epidemiology and health, 41, e2019017. doi:10.4178/epih.e2019017

Kanchana, S., Sivaprakash, P. and Joseph, S. (2015). Studies on Labour Safety in Construction Sites. The Scientific World Journal. 2015. 1-6. http://dx.doi.org/10.1155/2015/590810

Kumar G.V. and Dewangan K.N. Agricultural accidents in north eastern region of India. Saf. Sci., 47 (2009), pp. 199-205

Martorell, S., Gallego , V. & Sánchez A.I. On the use of accident indicators in risk-based management of occupational safety and health: Example of application to Spain. Occupational Safety and Hygiene IV. CRC Press. (2016a), pp. 327-331.

Medeiros, M. C. & Vasconcelos, G. F. R. (2015). Forecasting inflation with highdimensional time-series models. Working Paper.

Robert, K. , Elisabeth, Q. and Josef B. Analysis of occupational accidents with agricultural machinery in the period 2008–2010 in Austria. Saf. Sci., 72 (2015), pp. 319-328.

Shapiro SA, Rabinowitz R (2000) Voluntary regulatory compliance in theory and practice: the case of OSHA. Admin L Rev 52, 97–135.

Segarra M., Villena B.M., González M.N. et al. Occupational risk-prevention diagnosis: A study of construction SMEs in Spain. Safety Science 92 (2017) 104–115

Schwarz, G. (1978), Estimating the dimension of a model, The Annals of Statistics, 6(2), 461-464.

Soltanzadeh, A., Mohammadfam, I., Moghimbeygi, A. et al. Exploring Causal Factors on the Severity Rate of Occupational Accidents in Construction Worksites. Int J Civ Eng 15, 959–965 (2017). https://doi.org/10.1007/s40999-017-0184-9

Song Li, He Xueqiu B,C, and Li Chengwub. Longitudinal relationship between economic development and occupational. Analysis and Prevention 43 (2011) 82–86 accidents in China

Safety and health at the heart of the future of work. International Labour Organization. (2019).

Tibshirani, R. (1996), Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58: 267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

Villanueva V. and Garcia A.M. Individual and occupational factors related to fatal occupational injuries:A case-control study. Accident Analysis and Prevention 43 (2011) 123–127.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67: 301-320. doi:10.1111/j.1467-9868.2005.00503.x

Zou H. (2006) The adaptive Lasso and its oracle properties. Journal of the American Statistical Association. 2006;101:1418–1429.