

Stochastic reserving with a stacked model based on a hybridized Artificial Neural Network

Eduardo Ramos-Pérez⁽¹⁾,
Pablo J. Alonso-González⁽²⁾, José Javier Núñez-Velázquez⁽²⁾

(1) *Ph D Student (Economics and Management Program). Universidad de Alcalá.*

(2) *Economics Department. Universidad de Alcalá. *†*

Abstract

Currently, legal requirements demand that insurance companies increase their emphasis on monitoring the risks linked to the underwriting and asset management activities. Regarding underwriting risks, the main uncertainties that insurers must manage are related to the premium sufficiency to cover future claims and the adequacy of the current reserves to pay outstanding claims. Both risks are calibrated using stochastic models due to their nature. This paper introduces a reserving model based on a set of machine learning techniques such as Gradient Boosting, Random Forest and Artificial Neural Networks. These algorithms and other widely used reserving models are stacked to predict the shape of the runoff. To compute the deviation around a former prediction, a log-normal approach is combined with the suggested model. The empirical results demonstrate that the proposed methodology can be used to improve the performance of the traditional reserving techniques based on Bayesian statistics and a Chain Ladder, leading to a more accurate assessment of the reserving risk.

Keywords: Stochastic reserving, Reserving Risk, Machine Learning, General insurance, Run-off prediction

AMS Subject Classification: 62-07, 62P05, 65C60, 90-08.

1 Introduction

As with any other company, the survival of an insurance firm depends on its ability to obtain a sustainable profit over the years. These entities have to offer their services at an adequate and competitive premium, while the ultimate cost of the claims is subject to uncertainty. Thus, reserving models were developed in order to estimate and monitor the expected ultimate cost of outstanding claims. Although life insurance contracts manifest uncertainty about the claims cost, reserving takes a special relevance in general insurance as that uncertainty tends to be higher, at least in the

Authors' address: (1)&(2) :Economics Department, Universidad de Alcalá, Plaza de la Victoria 2, 28802 Alcalá de Henares, Spain. E-mails: P.J. Alonso-González, pablo.alonsog@uah.es, J.J. Núñez, josej.nunez@uah.es, E. Ramos, ramos.perez.e@gmail.com

Corresponding author: P. Alonso; Date: August 19, 2020. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

short term.

Methods of estimating the level of reserves in non-life insurance have evolved from classical and deterministic methods toward others that take into account the loss reserve uncertainty. The aim of the first type is to estimate the expected level of reserves by taking the historical information into consideration. Chain Ladder is the most frequently used method of this family. When historical data are not stable enough to use the Chain Ladder technique, the Bornhuetter and Ferguson (1972) model tends to be the preferred option to obtain an adequate estimate of the expected ultimate cost.

The increasing interest of investors in the risk profile of financial institutions since the Financial Crisis of 2007-2008 and the implementation of the Solvency II Directive in the European market have fostered the use of stochastic reserving models. As in the case of deterministic approaches, stochastic models based on the Chain Ladder technique are the most commonly used. One of the main techniques within this family is the Overdispersed Poisson (ODP) model developed by Renshaw and Verrall (1998) and its bootstrap implementation suggested by England and Verrall (1999) and England (2002) which assumes that incremental claims follow an ODP distribution where the variance is proportional to the mean.

In this model, incremental claims must be positive, but this limitation can be overcome by using the quasi-likelihood approach introduced by McCullagh and Nelder (1989). In cases where the ODP assumption does not properly fit the data, Kremer (1982), Mack (1991) and Verrall (2000) developed other models assuming log-normal, gamma and negative binomial distributions respectively. In contrast to the methods within this family, Mack (1993) developed a free-distribution model by focusing and limiting the claims reserve distribution analysis to the first two moments.

Thus, the bootstrap implementation of Mack's model allows the analyst to obtain a reserve distribution without the necessity of defining a theoretical distribution for the cumulative or incremental claim cost. If the bootstrapping procedure is to be avoided, England and Verrall (2006) introduced a stochastic Bayesian implementation of the ODP, Negative Binomial and this last free-distribution model. This approach was recently expanded by Meyers (2015), who developed some Bayesian Markov Chain Monte-Carlo (MCMC) models (Levelled Chain-Ladder, Correlated Chain-Ladder, Levelled Incremental Trend, Correlated Incremental Trend and Changing Settlement Rate) for incurred and paid data. Their aim is to improve the performance of ODP and Mack models by using different approaches such as recognizing the correlation between accident years, including a skewed distribution to model negative incremental payments, introducing a trend over the development years and allowing changes in the claim settlement rate.

Another set of models is focused on using several triangles simultaneously in order to take into consideration different characteristics of incurred and paid data. The main models within this family are the Munich Chain Ladder (MCL) method and Double Chain Ladder (DCL) model developed by Quarg and Mack (2004) and Martínez-

Miranda et al. (2012), respectively. By modifying this last method, Margraf et al. (2018) addressed the problem of calculating general insurance reserves when the portfolio is covered by an excess-of-loss reinsurance. In addition to MCL and DCL, Merz and Wüthrich (2010) introduced a Bayesian implementation of the paid-incurred chain (PIC) reserving method (Posthuma et al. 2008) based on using both incurred and paid data. Happ et al. (2012) and Happ and Wüthrich (2013) also investigated and developed models related to the PIC method, while Halliwell (2009) and Venter (2008) introduced regression approaches based on using both data sources. Pigeon et al. (2014), Antonio and Plat (2014), and Martínez-Miranda et al. (2013b) also proposed models by taking into consideration different data sources to estimate the expected ultimate claim cost.

In addition to the different approaches exposed above, it is possible to find models where the information is not organized in an aggregated way, as in the classical triangles, but rather in individual claims data (see Taylor et al. 2008, Jessen et al. 2011, Pigeon et al. 2013, Antonio and Plat 2014, Martínez-Miranda et al. 2015, Charpentier and Pigeon 2016, or Wüthrich 2018b).

Thanks to the increase in computational power, machine learning techniques have turned into an adequate tool for reserving purposes. Artificial Neural Networks (Gabrielli and Wüthrich 2018 and Wüthrich 2018b), regression trees (Wüthrich 2018a), Recurrent Neural Networks (Kuo 2018) or tree-based algorithms (Lopez et al. 2019) have been used to predict claim reserves. Gabrielli et al. (2018) embedded the ODP model into a neural network framework, and Baudry and Robert (2019) introduced a nonparametric reserving model based on extremely randomized trees (Geurts et al. 2006) and individual claims data. In addition to the aforementioned algorithms, other machine learning techniques were used by Martínez-Miranda et al. (2013a) for reserving purposes, and a support vector machine was applied to classify risks prior to the reserve calculation (Duma et al. 2011).

The research carried out in this paper develops a nonparametric reserving model based on the stacking algorithm methodology. The proposed architecture consists of two different levels. Random Forest (RF) (Breiman 2001), Gradient Boosting (GB) with regression trees (Friedman 2000), Artificial Neural Network (ANN) (McCulloch and Pitts 1943), Changing Settlement Rate (CSR) reserving model and the Chain Ladder assumptions are incorporated within the first level, while an ANN is included in the second level of the stacked model (Stacked-ANN) architecture in order to generate the final predictions. Therefore, the aim of this hybrid model is to improve the performance of the individual components by creating an architecture that can learn from the different algorithms and the reserving models included within the first level.

Although the overall methodology is based on that proposed by Ramos-Pérez et al. (2019) for stock volatility forecasting purposes, the model architecture proposed in this study is different. In this research, machine learning algorithms and reserving models are present in the first level, while in the architecture developed by Ramos-Pérez et al. (2019), only machine learning algorithms were included. Therefore, the

most popular models for forecasting volatility such as GARCH or EGARCH were not integrated within the model architecture, while in this case, Chain Ladder and CSR are incorporated. It is also worth mentioning that in contrast to the hybrid model proposed for forecasting volatility purposes, in this research, the second level only receives information already processed by the models within the first level. In addition to the main differences explained above, it should be pointed out that the stacking algorithm methodology has not appeared previously in the actuarial literature related to the valuation of loss reserves. Apart from that, a log-normal approach is combined with the suggested reserving model based on machine learning in order to compute the reserve variability.

As all the different algorithms and reserving models of the first level are incorporated in the ANN of the second level, some of the most important research studies carried out in the context of selecting the optimal ANN architecture will be discussed. There is a significant amount of literature supporting the use of ANNs with just one hidden layer because under mild assumptions on the activation functions, the universal approximation theorem states that a feedforward ANN with a single hidden layer and a finite number of neurons can approximate any continuous function on compact subsets of the Euclidean space.

Based on regularization techniques and using just one hidden layer network, Poggio and Girosi (1990) developed a theoretical framework to approximate nonlinear mappings named regularization networks. These authors demonstrated that their architecture can approximate any continuous function on a compact domain if the number of units is high enough. Cybenko (1989) and Hornik et al. (1989) also proved that one hidden layer networks with sigmoidal activation functions can approximate continuous functions on any compact Euclidean space. It was also shown that, under certain conditions, an arbitrarily small error between a single hidden layer ANN and any other continuous function can be obtained by increasing the number of neurons (Barron 1994, Funahashi 1989 and Hornik 1993). Nakama (2011) showed that the range of effective learning rates is wider in the case of ANN with one hidden layer than in architectures with multiple hidden layers.

On the other hand, Hornik (1991) and Leshno et al. (1993) demonstrated that ANNs have the potential of being universal approximators not only due to the choice of a specific activation function but also because of the possibility of using several hidden layers. Limitations of the approximation capabilities of one hidden layer networks were demonstrated by Chui et al. (1994) and Chui et al. (1996). In recent years, multi-hidden layer architectures have improved the state of the art in machine learning.

For example, in the context of natural language processing, the models and architectures created by Devlin et al. (2018) (BERT), Brown et al. (2020) (GPT3) and Vaswani et al. (2017) (Transformer) overcome the performance of other less complex models. In addition, it is worth mentioning that agents trained with multi-hidden layer ANNs have been able to overcome the human performance in specific tasks such as playing chess (Silver et al. 2017) or 'go' (Silver et al. 2016). With respect to the

optimal number of neurons, Celikoglu (2007) analysed this issue in the context of solving the dynamic network loading problem, while Sheela and Deepa (2013) proposed a list of principles to select this number.

Results from recently published papers in the actuarial field support the idea of applying ANNs with multiple hidden layers. Indeed, Richman and Wüthrich (2018) and Nigri et al. (2019) applied this structure to model human mortality, while Castellani et al. (2018) used it for estimating the economic capital of insurance companies under the Solvency II framework. Thus, the ANNs included within the architecture of the Stacked-ANN model have several hidden layers.

The rest of the paper proceeds as follows: Section 2 presents the set of models used for comparison purposes. Additionally, the error and risk measures taken to validate the stochastic reserves, payments and ultimate losses are discussed. In Section 3, the theoretical background and architecture of the reserving model based on stacking algorithms (Stacked-ANN) are explained. Details about the log-normal approach proposed for obtaining a stochastic distribution are also given in this section. The empirical results, error and risk measures of the different reserving models are shown in Section 4. Finally, Section 5 presents the main conclusions derived from the results and comparisons presented in Section 4.

2 Benchmark models and validation

As previously stated, this section explains the benchmark models and the different measures used to assess their performance. Thus, the first paragraphs are dedicated to ODP, Mack's model, CSR and a nonparametric approach based on ANNs, while the end of this section presents the indicators used to compare and validate the reserve distribution functions estimated by the benchmark models with those simulated by the model presented in Section 3.

The first benchmark model is ODP (Renshaw and Verrall 1998 and England and Verrall 1999). Denoting the origin year as i and the development year as j , this reserving model based on the Chain Ladder technique assumes that incremental payments, C_{ij} , follow an overdispersed Poisson distribution with a variance proportional to the mean:

$$E[C_{ij}] = \mu_{ij} \qquad \text{Var}[C_{ij}] = \phi\mu_{ij} \qquad (1)$$

where ϕ is the parameter that determines the level of overdispersion. Even though this model assumes C_{ij} to be a positive integer, the quasi-likelihood (McCullagh and Nelder 1989) approach allows fits the model to non-integer data, which can be either positive or negative. The bootstrapping procedure used in this study to compute a reserve distribution function with the ODP model was introduced by England and Verrall (1999) and England (2002).

Mack (1993) model, which is also based on the Chain Ladder technique, is the second benchmark. The main characteristic of this reserving model is the lack of assumptions

about the underlying distribution of the payments. This is achieved by using only the first two moments:

$$E[D_{ij}] = \lambda_j D_{i,j-1} \qquad \text{Var}[D_{ij}] = \sigma_j^2 D_{i,j-1} \qquad (2)$$

where λ_j and σ_j^2 refer to the parameters to be estimated, and D_{ij} is the cumulative payment. As with the ODP model, a bootstrapping procedure is used to calculate the reserve distribution function with Mack's model.

The third benchmark model is CSR, a Bayesian approach introduced by Meyers (2015). The default calibration and prior distributions suggested by this author will be used in this study:

- $\alpha_i \sim N(\ln P_i + \text{logelr}, \sqrt{10})$, where $\text{logelr} \sim U(-1, 0.5)$ and P_i are the premiums by accident year.
- $\beta_j \sim U(-5, 5)$ for $j = 1, \dots, J - 1$. In the last development year, $\beta_J = 0$.
- $\mu_{i,j} = \alpha_i + \beta_j(1 - \gamma)^{i-1}$, where $\gamma \sim N(0, 0.025)$.
- Each $\sigma_j = \sum_{i=j}^J a_i$, where $a_i \sim U(0, 1)$.

Taking into consideration the aforementioned distributions and parameters, the cumulative payments simulated by the CSR model follow a log-normal distribution, $D_{i,j} \sim LN(\mu_{i,j}, \sigma_j)$, subject to the constraint $\sigma_1 > \sigma_2 > \dots > \sigma_J$.

To analyse the improvement in the performance due to the stacking procedure that is presented in Section 3, the last benchmark model to be introduced is an individual ANN. The inputs and characteristics (hidden layers, activation functions, etc.) of this algorithm will be the same as those used for the ANN included within the first level of the Stacked-ANN. Additionally, the log-normal procedure to obtain the reserve variability is the same as that for the Stacked-ANN model. To avoid repeating content, refer to Section 3 for further details about the characteristics of the ANN used as a benchmark.

Once the four benchmark models are explained, the different measures selected to compare the performance of the Stacked-ANN with the aforementioned reserving models are presented. Insurance regulations such as the Solvency II Directive and Swiss Solvency Test ask the general insurance companies to evaluate their expected reserves and potential deviations from these central scenarios. Thus, the error of the estimated reserves will be computed in order to compare the performance of the different models. As several triangles with different levels of payments are used during this study, the measure for evaluating the reserves is

$$\%RMSE(R^t) = \frac{\sqrt{\sum_{k=1}^K (\hat{R}_{k,\mu}^t - R_k^t)^2 / K}}{\sum_{k=1}^K R_k^t} * 100 = \frac{RMSE(R^t)}{\sum_{k=1}^K R_k^t} * 100 \qquad (3)$$

where K is the total number of triangles, t is the calendar year when the reserves are evaluated, $\hat{R}_{k,\mu}^t$ is the reserve predicted by the reserving model using the triangle

k and R_k^t the reserves that were actually observed for that triangle. As it can be derived from the former expression, the aim of this error measure is the evaluation of the weight of the root mean squared error over the total reserves. To understand the model's performance, this error measure will also be calculated for the next year's payments ($\%RMSE(P^{t+1})$) and the ultimate loss cost ($\%RMSE(U^t)$).

In addition to the aforementioned error measures, the reserving risk (RR) per unit of reserve derived from the use of the different stochastic reserving models will be analysed. As previously stated, the models are going to be fitted to several triangles, so the average of the former ratio is taken as a risk measure:

$$Ratio(RR_{1-\alpha}^t) = \frac{\sum_{k=1}^K (\hat{R}_{k,1-\alpha}^t - \hat{R}_{k,\mu}^t) / \hat{R}_{k,\mu}^t}{K} = \frac{\sum_{k=1}^K RR_{1-\alpha}^t / \hat{R}_{k,\mu}^t}{K} \quad (4)$$

where $\hat{R}_{k,\mu}^t$ is the mean and $\hat{R}_{k,1-\alpha}^t$ is the percentile $1 - \alpha$ of the estimated reserve distribution function of the company k . A deeper evaluation of the variation estimated by the different stochastic models is carried out by calculating the standard deviation per unit of reserve:

$$Ratio(\sigma) = \frac{\sum_{k=1}^K \sigma(\hat{R}_k^t) / \hat{R}_{k,\mu}^t}{K} \quad (5)$$

Finally, in order to check the adequacy of the reserving risk calculated for the different companies, the Kupiec (1995) test is applied in order to verify if the number of excesses is aligned with the selected confidence level. The empirical results of the test and measures are collected in Section 4.

3 Stochastic reserving model based on the stacking algorithm approach

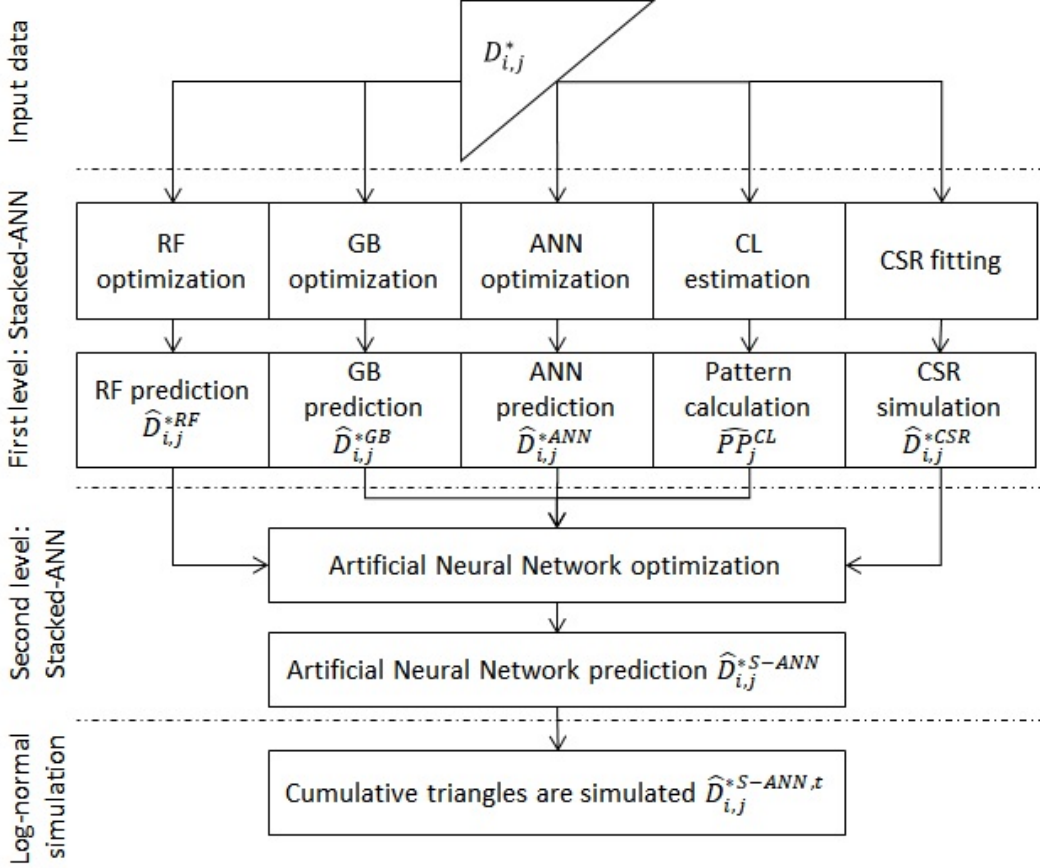
This section is divided into several subsections in order to sequentially explain the proposed reserving model. In addition, Figure 1 presents the model architecture in order to support the explanation.

3.1 Model inputs

Before estimating the different reserving models within the first level of the Stacked-ANN model, the database used, as well as the response and explicative variables for fitting the algorithms within this level, need to be defined.

The lower and upper triangles needed to fit and validate the models are obtained from Schedule P of the NAIC Annual Statement. This database (available on the CAS website) was collected from property and casualty insurers that underwrite business in the US, and it contains both paid and incurred losses (net of reinsurance) of the accident years from 1988 to 1997. Ten development years are available for every accident year. In addition to loss data, gross and net premiums by accident year are also reported in the database.

Figure 1: Stacked-ANN model structure



In this paper, the different reserving models will be fitted to 200 loss triangles from NAIC Schedule P, 50 from each of the following lines of business: Commercial Auto (CA), Private Passenger Auto Liability (PA), Workers' Compensation (WC) and Other Liability (OL). As pointed out by Meyers (2015), selecting triangles from insurers who made significant changes in business operations is one of the main mistakes that could be made with NAIC Schedule P data. The coefficient of variation of the net premiums and the net/gross premium ratio should be appropriate indicators of changes in business operations, so this author selected insurers that minimize the aforementioned metrics. The triangles selected by Meyers (2015) are used in this research in order to avoid the former issue and ensure comparability with other studies.

With regard to the explanatory variables, as with other nonparametric reserving models based on Generalized Additive Models (Hastie and Tibshirani 1986 and England and Verrall 2002) or RNN (Kuo 2018), accident i and development j years were selected to be the inputs of the first-level algorithms. Both were initialized as one and then scaled to range $[0, 1]$ (hereinafter AY_i^* and DY_j^*) in order to facilitate the fitting of the algorithms (Hastie et al. 2009).

The response variable of these algorithms is the scaled cumulative payments D_{ij}^* . Depending on the data availability and the characteristics of the portfolio to be modelled, different exposure measures can be selected to scale D_{ij} . In this paper, net premiums P_i play the role of exposure measure, as this is the most relevant option between the variables available in the database.

Figure 2: Train and test sets



Loss triangles are a representation of payments over time by accident or underwriting year. Thus, the training and optimization of the deep learning algorithms within the Stacked-ANN model architecture need to take into consideration that loss triangles are composed of temporal series. Accordingly, the last diagonal is selected as a test set because it contains the most updated information, while the rest of the triangle is used for fitting the algorithms (Figure 2).

During the optimization process, different configurations of the algorithms are fitted with the training data. To obtain the best configuration, the test set is predicted, and the root mean squared error of every option is computed. Finally, the configuration that minimizes the former test error is selected.

3.2 First level: Individual models

The first level of the Stacked-ANN model consists of a Chain Ladder, CSR, and three algorithms whose inputs were described in Section 3.1. It is worth mentioning that as ODP and Mack's model are based on the Chain Ladder technique, the Stacked ANN model incorporates the core rationale behind these stochastic reserving models. The machine learning algorithms (RF, GB and ANN) fitted at this step are explained in the following paragraphs and will be optimized by applying a grid search to some hyperparameters and by measuring the test error. Additionally, at the end of this subsection, the Chain Ladder and CSR hypothesis are integrated within the Stacked-ANN model architecture.

The Random Forest (RF) algorithm introduced by Breiman (2001) averages B different regression trees. In every fitted tree, the explanatory variables and data points used during the training are randomly selected. Therefore, the formal expression to

predict the scaled cumulative payments is:

$$\hat{D}_{ij}^{*RF} = \frac{\hat{D}_{ij}^{RF}}{P_i} = \frac{\sum_{b=1}^B T_b(X)}{B} \quad (6)$$

T_b represents the b -th regression tree fitted and X the selected subset of AY_i^* and DY_j^* to fit T_b . During the estimation process, the hyper-parameters optimized are the number of variables randomly selected, N , and the minimum number of observations to be kept in the terminal nodes of every fitted tree, Obs_{RF} .

The second algorithm within the first level is Gradient Boosting (GB) with regression trees (Friedman 2000). In this case, the gradient is minimized by sequentially fitting B regression trees. The subset of data to be used during the estimation process of every tree is also randomly selected. The expression to obtain the predicted scaled cumulative payments is

$$\hat{D}_{ij}^{*GB} = \frac{\hat{D}_{ij}^{GB}}{P_i} = \hat{f}_{B-1}(X) + \delta_{GB} T_B(X) \quad (7)$$

$\hat{f}_{B-1}(X)$ represents the function obtained after adding sequentially $B - 1$ regression tree models and, δ_{GB} is the learning rate. The hyperparameter selected to be optimized during the training process is the minimum number of observations to be kept in the terminal nodes of every fitted tree, Obs_{GB} . Regarding the hyperparameters, it is worth mentioning that the learning rate, δ_{GB} , is set to 0.01.

The last algorithm of the first layer is an Artificial Neural Network (ANN) (McCulloch and Pitts 1943). Following the notation provided by Bishop (2006) and taking into consideration that the feed-forward ANN used in this paper is composed of 2 hidden layers with 5 neurons each, the formal expression to obtain the predictions can be defined as follows:

$$\begin{aligned} \hat{D}_{ij}^{*ANN} &= \hat{D}_{ij}^{ANN} / P_i = \\ &= h^{(3)} \left(\sum_{k=1}^5 w_{1,k}^{(3)} h^{(2)} \left(\sum_{j=1}^5 w_{k,j}^{(2)} h^{(1)} \left(\sum_{i=1}^2 w_{j,i}^{(1)} x_i + w_{j,0}^{(1)} \right) + w_{k,0}^{(2)} \right) + w_{1,0}^{(3)} \right) \end{aligned} \quad (8)$$

where $h^{(n)}$ is the activation function associated with layer n , $w_{z,v}^{(n)}$ is the v -th weight associated with the neuron z inside layer n , and x_i refers to the i -th input variable of the database composed of two explanatory variables, the scaled accident (AY_i^*) and development year (DY_j^*). The percentage of dropout regularization θ is the hyperparameter to be optimized by applying a grid search and measuring the test error. As with the other algorithms, upper triangle predictions will be used as input within the second level of the architecture.

In addition to the three aforementioned algorithms, Chain Ladder assumptions are incorporated in the model architecture. To do so, the development factors of the Chain Ladder technique are used as an input in the second level of the Stacked-ANN

model:

$$\hat{\lambda}_j^{*CL} = \frac{\sum_{i=1}^{n-j-1} D_{ij}^*}{\sum_{i=1}^{n-j-1} D_{ij-1}^*} \quad (9)$$

where $\{\hat{\lambda}_j^{*CL} : j = (2, 3, \dots, J)\}$. Although the Chain Ladder methodology does not produce any parameters for $j = 1$, the second-level algorithm needs a value for $j = 1$. Thus, within the Stacked-ANN methodology, it is assumed that $\hat{\lambda}_1^{*CL} = 1$.

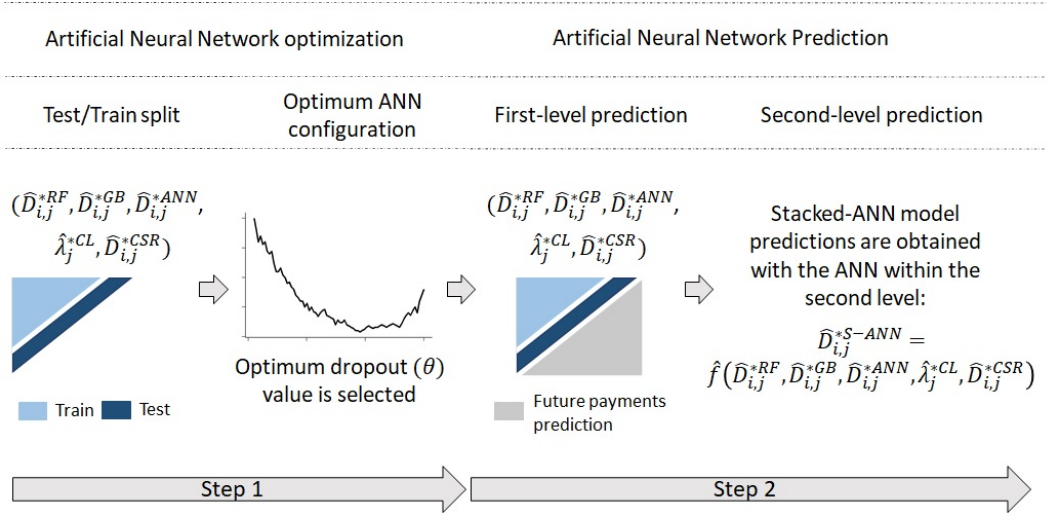
Finally, CSR methodology (Meyers 2015) is integrated. To achieve this, 10,000 MCMC simulations are produced within the first level of the Stacked-ANN model. Then, the expected scaled cumulative payments of the upper triangle arising from the aforementioned simulations are used as input in the algorithm within the second level of the Stacked-ANN model:

$$\hat{D}_{ij}^{*CSR} = \frac{\sum_{k=1}^{10,000} \hat{D}_{ijk}^{CSR} / P_i}{10,000} \quad (10)$$

3.3 Second level: Stacking algorithm

As previously stated, the inputs of this level are the scaled cumulative payments predicted by the algorithms named in Section 3.2 (RF, GB and ANN), the development factors based on the Chain Ladder technique and the expected scaled cumulative payments simulated by the CSR model. On the other hand, the output of the ANN within the second level and the Stacked-ANN are the cumulative payments \hat{D}_{ij}^{*S-ANN} by accident and development year.

Figure 3: Second-level structure



Similar to the first-level algorithms, the training and optimization processes of the ANN within this level need to recognize that loss triangles are composed of a set of

time series. The most recent information of the loss triangles is the last diagonal; thus, the explicative and response variables of this diagonal are selected as a test set, while the rest of the upper triangle data is used as a training set.

Once the test and training sets are defined, the optimum configuration of the ANN needs to be obtained. To do so, the training data are used to fit ANNs with different levels of dropout regularization θ . Then, the root mean squared error is computed by taking into consideration the predictions made by every ANN configuration. The θ that minimizes the test error is selected.

Due to the Stacked-ANN architecture, two substeps need to be carried out in order to make the final predictions. First, the lower triangle of the first-level models need to be predicted. Second, the data predicted in the previous step are used as input of the ANN within the second layer to make the final predictions. Thus, the Stacked-ANN model tries to obtain more accurate predictions by combining different reserving models and algorithms.

Figure 1 shows the overall Stacked-ANN architecture, and Figure 3 provides a detailed summary of the process defined in the previous paragraphs. Technical details about the feedforward ANN fitted within this level of the Stacked-ANN model are presented below:

- It contains two hidden layers with 5 neurons each. The sigmoid activation function was selected for all neurons within the hidden layers while the linear activation function was used in the output layer, which is composed of one neuron.
- The selected optimization algorithm is Adaptive Moment Estimation (ADAM), which was created by Kingma and Ba (2014). This method consists of a progressive adaptation of the initial learning rate, taking into consideration current and previous gradients. The default calibration proposed by the authors for the ADAM parameters is applied as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Thus, the ANN parameters are updated as follows:

$$\omega_t = \omega_{t-1} - \delta_{ANN} \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \quad (11)$$

$$\hat{m}_t = \frac{\beta_1 m_{t-1} + (1 - \beta_1) g_t}{1 - \beta_1^t} \quad (12)$$

$$\hat{v}_t = \frac{\beta_2 v_{t-1} + (1 - \beta_2) g_t^2}{1 - \beta_2^t} \quad (13)$$

where ω is the parameter to be updated and g_t the gradient in the epoch t . The initial learning rate is set to $\delta_{ANN} = 0.01$.

- The number of epochs is 10,000, and the batch size is equal to the length of the data used for training the ANN.
- The backward pass calculations are done according to the selection of the root mean squared error as a loss function.

- As previously stated, the percentage of dropout regularization θ is the hyper-parameter to be optimized by applying a grid search and measuring the test error.

Taking the abovementioned details into consideration, the scaled cumulative payments predicted by the Stacked-ANN model are obtained by means of the following expression:

$$\begin{aligned}\hat{D}_{ij}^{*S-ANN} &= \frac{\hat{D}_{ij}^{*S-ANN}}{P_i} = \hat{f}(\hat{D}_{ij}^{*RF}, \hat{D}_{ij}^{*GB}, \hat{D}_{ij}^{*ANN}, \hat{\lambda}_j^{*CL}, \hat{D}_{ij}^{*CSR}) = \\ &= h^{(3)} \left(\sum_{k=1}^5 w_{1,k}^{(3)} h^{(2)} \left(\sum_{j=1}^5 w_{k,j}^{(2)} h^{(1)} \left(\sum_{i=1}^5 w_{j,i}^{(1)} x_i + w_{j,0}^{(1)} \right) + w_{k,0}^{(2)} \right) + w_{1,0}^{(3)} \right)\end{aligned}\quad (14)$$

3.4 Log-normal simulation

To compute the Kupiec test and the measures related to reserve variability (Section 2), the deviation around the central scenario predicted by the Stacked-ANN model needs to be obtained. Due to its right skewness and long tail, log-normal distribution is widely used within reserving models to derive the variability of the claims cost. Many papers used the lognormal distribution to compute this variability (see, among others, Kremer (1982), Antonio et al. (2006), Rehman and Klugman (2009), Weke and Ratemo (2013), Meyers (2015) or more recently, Omari et al. (2018)).

In this study, a log-normal distribution is used to compute the reserve variability around the central scenario predicted by the Stacked-ANN. To do so, the parameters of this distribution are obtained using the aforementioned predictions and the moments method. Therefore, regardless of the distribution selected, the central scenario is that predicted by the Stacked-ANN, and thus, changing the distribution has no effect on the error measures described in Section 2. Nevertheless, changes to the log-normal hypothesis will modify the variability and, consequently, the risk measures ($Ratio(RR_{1-\alpha}^t)$ and $Ratio(\sigma)$) and the results of the Kupiec test. Below, the steps of the procedure are described:

1. Starting with the scaled cumulative payments predicted by the Stacked-ANN (\hat{D}_{ij}^{*S-ANN}), the variance by development year is computed as follows:

$$Var[\hat{D}_j^{*S-ANN}] = \frac{\sum_{i=1}^n \left(\hat{D}_{ij}^{*S-ANN} - E[\hat{D}_j^{*S-ANN}] \right)^2}{n-1}\quad (15)$$

where n refers to the total number of accident years and $E[\hat{D}_j^{*S-ANN}]$ is the mean of the scaled cumulative payments by development year.

2. By using the method of the moments and values calculated in the previous step, the parameters of the log-normal distribution are obtained:

$$\hat{\mu}_{ij}[\hat{D}_j^{*S-ANN}] = \ln \left(\frac{E[\hat{D}_j^{*S-ANN}]^2}{\sqrt{Var[\hat{D}_j^{*S-ANN}] + E[\hat{D}_j^{*S-ANN}]^2}} \right)\quad (16)$$

$$\hat{\sigma}_j^2[\hat{D}^{*S-ANN}] = \ln \left(1 + \frac{\text{Var}[\hat{D}_j^{*S-ANN}]}{E[\hat{D}_j^{*S-ANN}]} \right) \quad (17)$$

3. For $t = (1, 2, \dots, T)$:

- (a) A triangle is generated by sampling random values from the following distribution function: $\hat{C}_{ij}^{*S-ANN,k} \sim LN(\hat{\mu}_{ij}[\hat{D}^{*S-ANN}], \hat{\sigma}_j^2[\hat{D}^{*S-ANN}])$.
- (b) The final simulated values, $\hat{C}_{ij}^{S-ANN,k}$, are obtained by removing the scaling. Hence, the scaled payments obtained in the previous step are multiplied by P_i .

4 Results

In this section, the data used, the fitting process and a final comparison between the Stacked-ANN and the benchmark models are shown.

4.1 Data and fitting of the Stacked-ANN

As stated in Section 3.1, the upper and lower triangles required to fit and validate the models are obtained from Schedule P of the NAIC Annual Statement. This database contains the losses, reserves and premiums from 1988 until 1997 of different property and casualty insurers that underwrite business in the United States.

Meyers (2015) indicated that one of the main mistakes with the NAIC Schedule P data is selecting triangles from insurers that made significant changes in their businesses. Meyers used the coefficient of variation of the net premiums and the net-on-gross ratio to select 50 triangles of each of the following lines of business: Commercial Auto (CA), Private Passenger Auto Liability (PA), Workers' Compensation (WC) and Other Liability (OL). This triangle selection was also used in this paper in order to ensure comparability with other studies that take as a reference the selection made by Meyers (2015). For further details about the data used to fit the Stacked-ANN, refer to Section 3.1.

Once the data have been presented, the subsection focuses on the fitting of the Stacked-ANN. The first level of the proposed model is composed of three individual algorithms (RF, GB and ANN), the CSR reserving model and the development factors derived from the use of the Chain Ladder technique. The second level is composed of an ANN. As pointed out in Sections 3.2 and 3.3, the optimum hyperparameters of the algorithms within the first and second levels are obtained for each triangle using a grid search. Table 1 lists the minor differences across the lines of business in the means of the 50 optimum hyperparameters obtained for each algorithm.

Table 1: Mean of the optimum hyperparameters by line of business

Line of Business	RF first level	GB first level	ANN first level	ANN second level
CA	$Obs_{RF} = 2.04; N = 1.94$	$Obs_{GB} = 4.38$	$\theta = 0.10$	$\theta = 0.09$
PA	$Obs_{RF} = 2.66; N = 1.86$	$Obs_{GB} = 4.22$	$\theta = 0.07$	$\theta = 0.06$
WC	$Obs_{RF} = 1.78; N = 1.68$	$Obs_{GB} = 3.66$	$\theta = 0.13$	$\theta = 0.12$
OL	$Obs_{RF} = 1.50; N = 1.82$	$Obs_{GB} = 4.24$	$\theta = 0.12$	$\theta = 0.12$

Source: own elaboration

As previously stated, the development factors ($\widehat{\lambda}_j^{*CL}$) obtained by applying the Chain Ladder technique to D_{ij}^* are used as input for the ANN included within the second level of the Stacked-ANN model. These values are calculated for each triangle. Table 2 presents the means of the development factors by line of business.

With regard to the three algorithms of the first layer and the Chain Ladder technique, the CSR model is also incorporated in the Stacked-ANN architecture by means of inputting \widehat{D}_{ij}^{*CSR} in the second-level algorithm. This Bayesian reserving model is fitted to every single triangle. Tables 3 and 4 list the means of the CSR parameters by line of business.

Table 2: Mean of the development factors by line of business

Development factors	CA	PA	WC	OL
$\widehat{\lambda}_1^{*CL}$	1.89	1.77	2.21	6.66
$\widehat{\lambda}_2^{*CL}$	1.35	1.22	1.29	1.90
$\widehat{\lambda}_3^{*CL}$	1.16	1.10	1.13	1.33
$\widehat{\lambda}_4^{*CL}$	1.08	1.06	1.07	1.18
$\widehat{\lambda}_5^{*CL}$	1.04	1.03	1.04	1.10
$\widehat{\lambda}_6^{*CL}$	1.02	1.01	1.02	1.04
$\widehat{\lambda}_7^{*CL}$	1.00	1.01	1.02	1.02
$\widehat{\lambda}_8^{*CL}$	1.01	1.00	1.01	1.02
$\widehat{\lambda}_9^{*CL}$	1.00	1.00	1.01	1.01
$\widehat{\lambda}_{10}^{*CL}$	1.00	1.00	1.00	1.00

Source: own elaboration

Table 3, which is focused on the parameters needed to calculate the mean of the cumulative payments, presents positive γ and negative β_j for every line of business with the unique exception of CA, where $\beta_6, \beta_7, \beta_8$ and β_9 are positive. According to the model definition, the claims settlement speed increases when $\beta_j < 0$ and $\gamma > 0$. This common trend across the different lines of business about the claim settlement rate of the NAIC Schedule P data was already observed by Meyers (2015).

Table 3: CSR parameters by line of business: $D_{i,j}$ mean

CSR parameter	CA	PA	WC	OL	CSR parameter	CA	PA	WC	OL
α_1	7.094	8.959	8.423	6.162	β_1	-1.235	-0.987	-1.447	-2.446
α_2	7.166	9.047	8.612	6.269	β_2	-0.514	-0.400	-0.626	-1.332
α_3	7.171	9.148	8.779	6.330	β_3	-0.229	-0.198	-0.322	-0.709
α_4	7.280	9.143	8.666	6.309	β_4	-0.085	-0.097	-0.178	-0.363
α_5	7.348	9.201	8.644	6.334	β_5	-0.003	-0.042	-0.089	-0.173
α_6	7.347	9.282	8.537	6.515	β_6	0.039	-0.017	-0.057	-0.079
α_7	7.554	9.374	8.595	6.480	β_7	0.060	-0.008	-0.041	-0.045
α_8	7.540	9.389	8.514	6.327	β_8	0.028	-0.001	-0.029	-0.030
α_9	7.494	9.464	8.543	6.543	β_9	0.012	-0.001	-0.013	-0.014
α_{10}	7.556	9.492	8.500	6.327	γ	0.021	0.008	0.016	0.028

Source: own elaboration

The comparison of the CSR deviation by development year presented in Table 4 reveals that OL is the most volatile portfolio, while PA has the most stable reserves. For CA and WC, the reserve variability estimated by this Bayesian reserving model is quite similar, and it is located at an intermediate point between the OL and PA lines of business.

Table 4: CSR parameters by line of business: $D_{i,j}$ STD

CSR Parameter	CA	PA	WC	OL
σ_1	0.303	0.028	0.236	0.771
σ_2	0.176	0.011	0.164	0.488
σ_3	0.109	0.007	0.117	0.327
σ_4	0.079	0.004	0.090	0.229
σ_5	0.063	0.003	0.069	0.164
σ_6	0.052	0.002	0.052	0.120
σ_7	0.043	0.002	0.037	0.087
σ_8	0.035	0.001	0.025	0.061
σ_9	0.026	0.001	0.016	0.038
σ_{10}	0.014	0.001	0.008	0.019

Source: own elaboration

4.2 Comparison against benchmark models

Once the Stacked-ANN reserving model is fitted, its performance is compared with the benchmark models explained in Section 2 (ODP, Mack, CSR and an individual ANN).

Table 5 lists the %RMSEs associated with reserves R^t , next year payments P^{t+1} and ultimate losses U^t by line of business and reserving model. For further details about the measures presented in the table, refer to Section 2.

Table 5: %RMSE by line of business and reserving model

Error Measure	Line of Business	ODP	Mack's Model	CSR	ANN	Stacked ANN
$\%RMSE(R^t)$	CA	0.896%	0.896%	0.534%	1.768%	0.739%
$\%RMSE(P^{t+1})$	CA	0.668%	0.669%	0.573%	1.775%	0.876%
$\%RMSE(U^t)$	CA	0.170%	0.171%	0.102%	0.337%	0.141%
$\%RMSE(R^t)$	PA	1.012%	1.004%	0.823%	5.006%	0.254%
$\%RMSE(P^{t+1})$	PA	1.290%	1.286%	0.258%	1.900%	0.320%
$\%RMSE(U^t)$	PA	0.131%	0.131%	0.107%	0.651%	0.033%
$\%RMSE(R^t)$	WC	1.295%	1.286%	1.751%	1.943%	1.058%
$\%RMSE(P^{t+1})$	WC	0.887%	0.880%	1.531%	1.525%	0.676%
$\%RMSE(U^t)$	WC	0.222%	0.221%	0.301%	0.333%	0.182%
$\%RMSE(R^t)$	OL	5.274%	5.086%	3.153%	5.725%	0.722%
$\%RMSE(P^{t+1})$	OL	2.216%	2.102%	5.528%	0.268%	1.095%
$\%RMSE(U^t)$	OL	1.760%	1.709%	1.056%	1.918%	0.242%

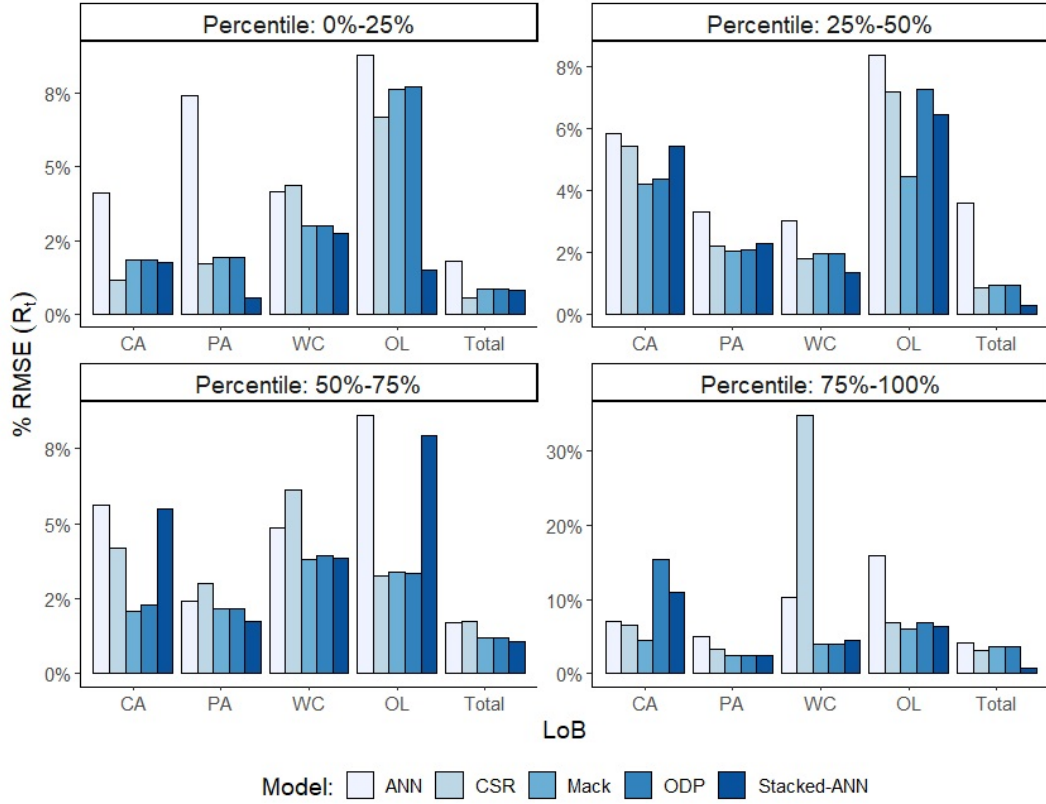
Source: own elaboration

The results obtained by using the different reserving models are summarized as follows:

- The Stacked-ANN model outperforms the individual ANN. The proposed architecture is empirically more accurate because it can learn from the reserving models (Chain Ladder and CSR) and machine learning algorithms (RF, GB and ANN) included within the first level of the Stacked-ANN, while the individual ANN must base its training only on the origin data (AY_i^* and DY_j^*) without taking advantage of other models that are able to capture different patterns and characteristics.
- As they are based on Chain Ladder assumptions, the mean of the distributions generated by ODP and Mack's model should converge to the values obtained by applying the deterministic approach of the Chain Ladder technique. Consequently, the error measures observed in Table 5 for these two stochastic reserving models are almost the same. The table also reveals that ODP and Mack are less accurate than the Stacked-ANN model in most cases. $\%RMSE(P^{t+1})$ of CA is a unique category in which the benchmark models based on the Chain Ladder technique are more accurate than the proposed methodology.
- Regarding the comparison between Stacked-ANN and CSR, R^t and U^t of PA, WC and OL estimated by the proposed model are significantly more accurate than those obtained when using the Bayesian model. Additionally, $\%RMSE(P^{t+1})$ of the Stacked-ANN model is lower in WC and OL. Thus, in the majority of cases, the CSR model is outperformed by the proposed methodology.

To enhance the analysis of the results presented in Table 5, Figure 4 shows the $\%RMSE(R^t)$ by line of business and volume of reserves. First, the companies were

Figure 4: $\%RMSE(R^t)$ by line of business and volume of reserves.



classified in four different groups taking into consideration the volume of reserves and the quartiles associated with the distribution. Then, the error rate of each reserving model was computed by line of business. The former calculation was carried out without making any distinction between lines of business.

The results of the aforementioned figure reveal that when no distinction between lines of business is made, the Stacked-ANN architecture outperforms the rest of the reserving models regardless of the company size. This difference is especially relevant for those companies with a higher level of reserves, whose results are collected in the graph labelled ‘Percentile: 75%-100%’. As expected, some fluctuations in the performance of the models are observed when the results are analysed by line of business and volume of reserves. Nonetheless, the error rate of the Stacked-ANN tends to be lower than the rest of the benchmark models.

In accordance with the reasons explained within the former paragraphs, it can be concluded that the Stacked-ANN model takes advantage of the different characteristics of several reserving models and machine learning algorithms, leading to a more flexible and precise architecture in most of the cases.

In addition to the error analysis, the risk measures ($Ratio(RR_{1-\alpha}^t)$ and $Ratio(\sigma)$)

and the p-values of the Kupiec test obtained by using each reserving model are compared. Before examining the results, it is important to point out that Mack's model does not make any assumptions about the payment distribution, ODP assumes that incremental payments follow an overdispersed Poisson distribution, and CSR, ANN and Stacked-ANN presume that cumulative payments are log-normally distributed. The hypothesis taken regarding the payments impact the distribution shape and consequently the risk measures. Therefore, in this case, ODP and Mack's model are not going to converge like they did in the central scenario.

Table 6: Risk measures by line of business and reserving model

Risk measures	Line of Business	ODP	Mack's Model	CSR	ANN	Stacked ANN
$Ratio(RR_{0.995}^t)$	CA	1.936	1.460	2.776	1.387	1.884
$Ratio(\sigma)$	CA	2.561	0.461	0.681	0.456	0.642
Kupiec p-value	CA	≥ 0.05	≥ 0.05	≥ 0.05	< 0.05	≥ 0.05
$Ratio(RR_{0.995}^t)$	PA	0.544	0.373	0.918	0.783	0.888
$Ratio(\sigma)$	PA	0.277	0.135	0.270	0.279	0.332
Kupiec p-value	PA	≥ 0.05	≥ 0.05	≥ 0.05	≥ 0.05	≥ 0.05
$Ratio(RR_{0.995}^t)$	WC	2.525	0.691	1.797	2.194	2.149
$Ratio(\sigma)$	WC	1.273	0.245	0.474	0.682	0.717
Kupiec p-value	WC	< 0.05	< 0.05	< 0.05	< 0.05	≥ 0.05
$Ratio(RR_{0.995}^t)$	OL	7.506	3.287	4.843	2.315	3.522
$Ratio(\sigma)$	OL	6.275	1.217	1.119	0.690	1.099
Kupiec p-value	OL	≥ 0.05	≥ 0.05	≥ 0.05	≥ 0.05	≥ 0.05

Source: own elaboration

The $Ratio(RR_{1-\alpha}^t)$ and the p-values collected in Table 6 evaluate the 99.5 percentile ($\alpha = 0.005$) of the reserve distribution function, which is the confidence level set up by Solvency II to calculate the risk of the insurance companies. The results of this table are summarized below:

- According to the results of the Kupiec test, the Stacked-ANN generates an adequate risk assessment for every line of business. It is worth mentioning that when compared with the individual ANN, the empirical results show that the stacking process not only improves the error rate but also allows for the generation of more appropriate distribution functions using the same simulation approach (presented in Section 3.4). With regard to the comparison between the Stacked-ANN and the rest of benchmark models, the Kupiec test reveals that CSR, ODP and Mack's model do not produce appropriate risk measures for WC, while the proposed methodology passes the test.
- Intuitively, the duration of the liabilities should have a close relation with $Ratio(RR_{0.995}^t)$ and $Ratio(\sigma)$: the longer the duration, the higher is the uncertainty around each economic unit of reserve. The development factors based on the Chain Ladder technique measure the claim settlement speed. Therefore,

they can be considered a good indicator of the duration of liabilities. A development factor at year t , λ_t , means that the $t + 1$ cumulative payment is λ_t times the cumulative claims settled at t . Consequently, high development factors indicate a long duration, while low values reflect a high settlement speed. According to Table 2, OL is the line of business with the highest duration, while PA has the lowest. CA and WC, whose durations are in a similar range, are located at an intermediate point between PA and OL. As can be observed in Table 6, this intuition about the relation between the duration and reserve uncertainty is followed by the Stacked-ANN and benchmark models.

- In general, the $Ratio(RR_{0.995}^t)$ and $Ratio(\sigma)$ by line of business are similar across the different reserving models. The two main exceptions are the risk measures of ODP for OL and Mack for WC. The high values observed in the ODP estimations for OL are due to two companies whose $RR_{1-\alpha}^t/\hat{R}_{k,\mu}^t$ ratios are higher than 60, while in the second case, Mack’s model systematically underestimates the variability of the payments, leading to lower values compared with the rest of the models and an inadequate risk assessment according to the results of the Kupiec test. The $Ratio(RR_{0.995}^t)$ and $Ratio(\sigma)$ of the Stacked-ANN are in line with the majority of the benchmark models, and no extremely high/low risk measures are observed in Table 6.

4.3 Sensitivity analysis of the number of hidden layers

As explained in Section 3, the ANNs included within the proposed Stacked-ANN architecture are composed of two hidden layers, each with five neurons. To analyse the impact of the ANN complexity (Cybenko 1989, Hornik et al. 1989, Hornik 1991 and Leshno et al. 1993, among others, introduced the theoretical framework to analyse the approximation capabilities of neural networks) on the predictive power of the Stacked-ANN model, a sensitivity analysis of the number of hidden layers was carried out. Thus, Table 7 compares the configuration selected for the Stacked-ANN model in this paper with two alternative configurations: ANNs composed of one and three hidden layers with five neurons each.

Two main conclusions can be drawn from the results obtained. First, the high level of error of the one hidden layer model demonstrates that more complexity is needed in order to properly predict general insurance reserves. The structure proposed during this study for the Stacked-ANN model (two hidden layers) performs significantly better than this first alternative in every single line of business.

Second, the performance of the three hidden layers alternative is similar to that of the suggested architecture. As no significant differences are observed, the two hidden layer structure is considered more appropriate because the three hidden layer structure adds complexity to the model without a significant improvement in the error rate.

Table 7: Sensitivity analysis of the number of hidden layers

Hidden layers	Line of Business	$\%RMSE(R^t)$	$\%RMSE(P^{t+1})$	$\%RMSE(U^t)$
1	CA	0.840%	0.766%	0.160%
2	CA	0.739%	0.876%	0.141%
3	CA	0.780%	0.643%	0.149%
1	PA	2.512%	3.507%	0.326%
2	PA	0.254%	0.320%	0.033%
3	PA	0.231%	0.115%	0.030%
1	WC	1.398%	1.296%	0.240%
2	WC	1.058%	0.676%	0.182%
3	WC	1.140%	1.030%	0.196%
1	OL	1.419%	1.145%	0.475%
2	OL	0.722%	1.095%	0.242%
3	OL	0.613%	0.794%	0.205%

Source: own elaboration

5 Conclusions

This paper introduced a stochastic reserving model based on stacking different machine learning algorithms (RF, GB and ANN) and reserving models (Chain Ladder and CSR). The predictive power and reserve volatility of the proposed approach, named Stacked-ANN, were compared with stochastic reserving models based on the Chain Ladder technique (ODP and Mack’s model), an individual ANN and CSR, which is a Bayesian loss reserving model.

Three main conclusions were drawn. First, a comparison of the Stacked-ANN with the individual ANN revealed that the predictions of the reserves R^t , next year payments P^{t+1} and ultimate losses U^t made by machine learning algorithms were improved by applying the proposed stacking procedure. The hybrid architecture learns patterns and characteristics from several algorithms and reserving models, resulting in a more flexible and accurate model than an individual ANN, whose inputs for training are limited to the original data.

Second, the empirical results indicated that the Stacked-ANN model is more precise than CSR and the most widely used stochastic reserving models based on the Chain Ladder technique (ODP and Mack’s model). In particular, the R^t and U^t predictions made by the Stacked-ANN were more precise than those of ODP and Mack’s model in all the lines of business analysed, while the Bayesian model (CSR) was outperformed by the proposed architecture in three out of four lines of business. It is important to remark that in Other Liability (OL), which is a line of business with a longer duration and therefore a portfolio where the importance of an accurate reserves estimation is especially relevant, the error of the models based on Chain Ladder or Bayesian statistics was more than four times the error of the Stacked-ANN. Therefore, it can be concluded that machine or deep learning techniques can be used to improve the per-

formance of the traditional reserving techniques based on Bayesian statistics or the Chain Ladder.

With regard to accuracy, it is worth mentioning that the proposed structure of the ANNs (two hidden layers) within the Stacked-ANN model seems to be the optimal configuration according to the empirical results. On the one hand, the error increased significantly when the number of hidden layers is reduced to one. On the other hand, the results demonstrated that increasing the number of hidden layers does not have an impact on the accuracy. Thus, increasing the complexity of the ANNs by up to three hidden layers will extend the training phase without making any significant improvement in the error.

Third, the results of a Kupiec test revealed that the risk estimation made by the Stacked-ANN can be considered as appropriate in all lines of business analysed, while the rest of the benchmark models failed the test at least once. In particular, CSR, ODP and Mack's model were unable to produce an appropriate p-value for the Kupiec test in the Workers' Compensation (WC) business, while the individual ANN failed the test in Commercial Auto (CA) and, as with the previous models, in Workers' Compensation. Taking into consideration that the same log-normal approach was used to obtain the reserves variability of the individual ANN and the Stacked-ANN, it must be mentioned that the stacking procedure not only increases the accuracy but also allows for the simulation of more adequate distribution functions.

The aforementioned robustness and predictive power of the Stacked-ANN compared with other reserving models suggest that further investigation should be conducted about the possible application of this model within the *actuary in the box* approach. The generation of outliers is one of the main problems when using the former methodology with Chain Ladder models. Therefore, the robustness of the Stacked-ANN can be exploited in order to improve the *actuary in the box* methodology, which is widely used to assess the fact that reserves can be insufficient to cover their runoff over a 12-month time horizon.

References

- Antonio, K., J. Beirlant, T. Holdemakers, and R. Verlaak (2006). Log-normal mixed Models for Reported Claims Reserves. *North American Actuarial Journal* 7, 1223–1237.
- Antonio, K. and R. Plat (2014). Micro-level stochastic loss reserving in general insurance. *Scandinavian Actuarial Journal* 2014, 649–669.
- Barron, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 115–133.
- Baudry, M. and C. Robert (2019). A Machine Learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 1–29.

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.
- Bornhuetter, R. L. and R. E. Ferguson (1972). The actuary and IBNR. *Proceedings of the Casualty Actuarial Society*, 181–195.
- Breiman, L. (2001, Oct). Random Forests. *Machine Learning* 45(1), 5–32.
- Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners.
- Castellani, G., U. Fiore, Z. Marino, L. Passalacqua, F. Perla, S. Scognamiglio, and P. Zanetti (2018). An Investigation of Machine Learning Approaches in the Solvency II Valuation Framework. Available at SSRN: <https://ssrn.com/abstract=3303296>.
- Celikoglu, H. (2007, 10). A dynamic network loading process with explicit delay modelling. *Transportation Research Part C: Emerging Technologies* 15, 279–299.
- Charpentier, A. and M. Pigeon (2016). Macro vs Micro Methods in Non-Life Claims Reserving: An Econometric Perspective. *Risks* 4, 12.
- Chui, C., X. Li, and H. Mhaskar (1994). Neural networks for localized approximation. *Mathematics of computation*.
- Chui, C., X. Li, and H. Mhaskar (1996). Limitations of the approximation capabilities of neural networks with one hidden layer. *Advances in Computational Mathematics*.
- Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems* 2, 303–314.
- Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.
- Duma, M., B. Twala, T. Marwala, and F. Nelwamondo (2011). Improving the Performance of the Support Vector Machine in Insurance Risk Classification: A Comparative Study. *Proceedings of the International Conference on Neural Computation Theory and Applications (NCTA-2011)*, 340–346.
- England, P. D. (2002). Addendum to analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics* 31.
- England, P. D. and R. J. Verrall (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics* 25(3), 281–293.
- England, P. D. and R. J. Verrall (2002, 01). Stochastic Claims Reserving in General Insurance. *Br. Actuar. J.* 8, 443–544.
- England, P. D. and R. J. Verrall (2006). Predictive Distributions of Outstanding Liabilities in General Insurance. *Annals of Actuarial Science*, 221270.

- Friedman, J. H. (2000). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29, 1189–1232.
- Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks* 2.
- Gabrielli, A., R. Richman, and M. Wüthrich (2018, 11). Neural Network Embedding of the Over-Dispersed Poisson Reserving Model. Available at SSRN, <https://ssrn.com/abstract=3288454>.
- Gabrielli, A. and M. Wüthrich (2018). An Individual Claims History Simulation Machine. *Risks* 6, 29.
- Geurts, P., D. Ernst, and L. Wehenkel (2006). Extremely Randomized Trees. *Machine Learning* 63, 3–42.
- Halliwell, L. (2009). Modeling Paid and Incurred Losses Together. *CAS E-Forum, Spring*, 1–40.
- Happ, S., M. Merz, and M. Wüthrich (2012). Claims development result in the paid-incurred chain reserving method. *Insurance: Mathematics and Economics* 51, 66–72.
- Happ, S. and M. Wüthrich (2013). Paid-incurred chain reserving method with dependence modeling. *ASTIN Bulletin* 43, 1–20.
- Hastie, T. and R. Tibshirani (1986, 08). Generalized Additive Models. *Statist. Sci.* 1(3), 297–310.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.
- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4, 251–257.
- Hornik, K. (1993). Some new results on neural network approximation. *Neural networks* 6, 1069–1072.
- Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* 2(5), 359 – 366.
- Jessen, A., T. Mikosch, and G. Samorodnitsky (2011). Prediction of outstanding payments in a Poisson cluster model. *Scandinavian Actuarial Journal* 2011(3), 214–237.
- Kingma, D. P. and J. Ba (2014). Adam: A Method for Stochastic Optimization. *CoRR*.
- Kremer, E. (1982, 01). IBNR claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal* 1982.
- Kuo, K. (2018). DeepTriangle: A Deep Learning Approach to Loss Reserving. *CoRR abs/1804.09253*.
- Kupiec, P. H. (1995, January). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives* 3(2), 73–84.

- Leshno, M., V. Lin, A. Pinkus, and S. Schocken (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks* 6, 861–867.
- Lopez, O., X. Milhaud, and P. Thérond (2019). A Tree-Based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*.
- Mack, T. (1991). A Simple Parametric Model for Rating Automobile Insurance or Estimating IBNR Claims Reserves. *ASTIN Bulletin* 21(1), 93109.
- Mack, T. (1993). Distribution-free Calculation of the Standard Error of Chain Ladder Reserve Estimates. *ASTIN Bulletin: The Journal of the International Actuarial Association* 23(02), 213–225.
- Margraf, C., V. Elpidorou, and R. Verrall (2018). Claims reserving in the presence of excess-of-loss reinsurance using micro models based on aggregate data. *Insurance: Mathematics and Economics* 80, 54–65.
- Martínez-Miranda, M., B. Nielsen, and R. Verrall (2012). Double Chain Ladder. *ASTIN Bulletin* 42(1), 59–76.
- Martínez-Miranda, M., B. Nielsen, and R. Verrall (2013a). Continuous Chain Ladder: Reformulating and generalizing a classical insurance problem. *Expert Systems with Applications* 40, 5588–5603.
- Martínez-Miranda, M., B. Nielsen, and R. Verrall (2013b). Double Chain Ladder and Bornhuetter-Ferguson. *North American Actuarial Journal* 17, 101–113.
- Martínez-Miranda, M., B. Nielsen, R. Verrall, and Wüthrich (2015). The Link Between Classical Reserving and Granular Reserving Through Double Chain Ladder and its Extensions. *Scandinavian Actuarial Journal* 2015, 383–405.
- McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.
- McCulloch, W. and W. Pitts (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5, 127–147.
- Merz, M. and M. V. Wüthrich (2010). Paid-incurred chain claims reserving method. *Insurance: Mathematics and Economics* 46, 568–579.
- Meyers, G. (2015). *Stochastic Loss Reserving Using Bayesian MCMC Models*. CAS Monograph Series, number 1. Casualty Actuarial Society.
- Nakama, T. (2011). Comparisons of Single and Multiple Hidden Layer Neural Networks. pp. 270–279.
- Nigri, A., S. Levantesi, S. Marino, M. Scognamiglio, and F. Perla (2019). A deep learning integrated leecarter model. *Risk* 7, 33.
- Omari, C., S. Nyambura, and J. Wairimu (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. *Journal of Mathematical Finance* 8, 137–160.
- Pigeon, M., K. Antonio, and M. Denuit (2013). Individual loss reserving with the Multivariate Skew Normal framework. *ASTIN Bulletin* 43, 399–428.

- Pigeon, M., K. Antonio, and M. Denuit (2014). Individual loss reserving using paid-incurred data. *Insurance: Mathematics and Economics* 58, 121–131.
- Poggio, T. and F. Girosi (1990). Networks for approximation and learning. *Proceedings of the IEEE* 78, 1481–1497.
- Posthuma, B., E. Cator, W. Veerkamp, and E. Van Zwet (2008). Combined Analysis of Paid and Incurred Losses. *CAS E-Forum Fall*, 272–293.
- Quarg, G. and T. Mack (2004). Munich Chain Ladder. *Blätter der Deutschen Gesellschaft für Versicherungen und Finanzmathematik XXVI*, 597–630.
- Ramos-Pérez, E., P. Alonso-González, and J. Núñez-Velázquez (2019). Forecasting volatility with a stacked model based on a hybridized Artificial Neural Network. *Expert Systems with Applications* 129, 1–9.
- Rehman, Z. and S. Klugman (2009). Quantifying Uncertainty in Reserve Estimates. *Variance Journal* 4, 30–46.
- Renshaw, A. E. and R. J. Verrall (1998). A Stochastic Model Underlying the Chain-Ladder Technique. *British Actuarial Journal*.
- Richman, R. and M. Wüthrich (2018). A Neural Network Extension of the Lee-Carter Model to Multiple Populations. *SSRN*. Available at SSRN, <https://ssrn.com/abstract=3270877>.
- Sheela, D. and S. Deepa (2013). Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Mathematical Problems in Engineering*, 1–11.
- Silver, D., A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis (2016, 01). Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489.
- Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. P. Lillicrap, K. Simonyan, and D. Hassabis (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR abs/1712.01815*.
- Taylor, G., G. McGuire, and J. Sullivan (2008). Individual Claim Loss Reserving Conditioned by Case Estimates. *Annals of Actuarial Science* 3(1-2), 215–256.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. *CoRR abs/1706.03762*.
- Venter, G. (2008). Distribution and Value of Reserves Using Paid and Incurred Triangles. *CAS E-Forum, Fall*, 348–375.
- Verrall, R. J. (2000). An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: Mathematics and Economics* 26(1), 91–99.
- Weke, P. and C. Ratemo (2013). Estimating IBNR Claims Reserves for General Insurance Using Archimedean Copulas. *Applied Mathematical Sciences* 7, 1223–1237.

- Wüthrich, M. (2018a). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal* 2018, 465–480.
- Wüthrich, M. (2018b). Neural networks applied to chain-ladder reserving. *European Actuarial Journal* 8, 407–436.