


Neural networks for quantile claim amount estimation: a quantile regression approach

Alessandro G. Laporta¹ , Susanna Levantesi¹ and Lea Petrella²

¹Department of Statistics, Sapienza University of Rome, Roma, Italy; and ²MEMOTEF Department, Sapienza University of Rome, Roma, Italy

Corresponding author: Alessandro G. Laporta; Email: alelaporta93@gmail.com

(Received 21 July 2022; revised 16 April 2023; accepted 19 April 2023)

Abstract

In this paper, we discuss the estimation of conditional quantiles of aggregate claim amounts for non-life insurance embedding the problem in a quantile regression framework using the neural network approach. As the first step, we consider the quantile regression neural networks (QRNN) procedure to compute quantiles for the insurance ratemaking framework. As the second step, we propose a new quantile regression combined actuarial neural network (Quantile-CANN) combining the traditional quantile regression approach with a QRNN. In both cases, we adopt a two-part model scheme where we fit a logistic regression to estimate the probability of positive claims and the QRNN model or the Quantile-CANN for the positive outcomes. Through a case study based on a health insurance dataset, we highlight the overall better performances of the proposed models with respect to the classical quantile regression one. We then use the estimated quantiles to calculate a loaded premium following the quantile premium principle, showing that the proposed models provide a better risk differentiation.

Keywords: Quantile regression; Neural networks; Insurance pricing

1. Introduction

The use of machine learning in insurance pricing has flourished in recent years and the related literature is increasing accordingly. For example, Guelman (2012) adopts gradient boosting for auto insurance cost modelling; Spedicato *et al.* (2018) performs insurance pricing optimisation using several machine learning models; Henckaerts *et al.* (2021) use tree-based techniques such as GBM in order to produce car insurance tariffs; Schelldorfer & Wüthrich (2019) employ neural networks to enhance GLMs performances in non-life insurance and Wüthrich (2020) proposes two different techniques to overcome the unbiasedness of neural network models for insurance portfolios. Machine learning techniques have also found extensive application in the context of insurance claim reserving: Gabrielli *et al.* (2020) improve the performances of over-dispersed Poisson model for general insurance claims reserving by means of neural networks embedding, and Wüthrich (2018) propose several machine learning algorithms for individual claim reserving.

However, such techniques only give an estimation of the expected value of the chosen variable (claim frequency or claim severity), since they are designed to return the pure premium of a specific policy, where the expected value of the total claim amount is given by the product between the expected values of claim frequency and claim severity. Hence, these models, even if they offer insight into the average loss of a policy, are unable to provide the modeler with some valuable information about its potential riskiness, e.g. the quantile of the total claim amount. To overcome this problem a quantile regression approach, originally introduced by Koenker & Bassett

(1978), may be considered since it provides information on the whole distribution of a given phenomenon. The quantile regression technique represents a robust distribution-free methodology that has been widely used in the financial literature to compute risk measures like the value-at-risk (see for example Taylor, 2007; White *et al.*, 2015; Laporta *et al.*, 2018; Adrian & Brunnermeier, 2016; Petrella & Raponi, 2019; Taylor, 2020, and Merlo *et al.*, 2021b).

The quantile regression approach appears particularly suitable in the insurance context, when assessing the Solvency II capital requirements and calculating the premium safety loadings. Furthermore, it enables the insurer to soundly gauge the portfolio riskiness (i.e. computing the Value-at-Risk of a given portfolio). Modeling the quantile claim amount through the quantile regression (QR) has already been discussed by a handful of authors: Kudryavtsev (2009) was the first introducing the use of the two-stage QR model to estimate the quantile of the total claim amount; Heras *et al.* (2018) propose a refinement of the previous model since they take into account heterogeneous claim probabilities, whereas Kudryavtsev (2009) only considers a single probability of having claims for each type of policyholder; Baione & Biancalana (2019) propose an alternative two-stage approach, where the risk margin considered in the ratemaking is calibrated on the claim's severity for each risk class in the portfolio, avoiding some of the drawbacks that characterise the technique proposed by Heras *et al.* (2018).

The common QR method requires the specification of a predetermined dependence structure between the dependent variable and the covariates or to elaborate its complex functional form to account for non-linearity or interactions among regressors. Unfortunately, the structural form of the dependence is often unknown to the modeler. So, a different approach should be pursued in this context. Neural networks appear to be an interesting modeling technique overcoming these limitations since they can fit a complex data structure without any a priori assumption on the relation among variables.

In this paper, we propose two innovative methods to estimate the conditional quantile of the total claim amount for a group of health policies. The first one uses the quantile regression neural network (QRNN), a particular specification of a feed-forward neural network originally introduced by Taylor (2000), able to estimate conditional quantiles. This model, up to our knowledge, has never been used in the context of insurance ratemaking.

The second model we propose considers a new extension of the combined actuarial neural network (CANN) proposed by Schelldorfer & Wüthrich (2019) in a quantile regression framework. The original CANN formulation is devoted to claim frequency estimation and combines a Poisson GLM with a neural network. In our model, since we are interested in the conditional quantile of the total claim amount, we nest the QR model into the structure of a neural network (Quantile-CANN henceforth). This approach is able to represent additional information incorporated in the data and not captured by the simple QR model.

The structure of the approach here considered is based on a two-part model. A simple, but common and useful version of such model involves a model for a binary indicator variable and a model for the response variable given that the binary indicator takes the value one. Following this approach, we fit a logistic regression for the binary variable to estimate the claim probability while we use a QRNN or a Quantile-CANN to model the quantile of the positive outcome. Using the estimated quantiles of the claim amount, we finally calculate a loaded premium following the quantile premium principle considered in Heras *et al.* (2018).

To verify the predictive ability of the proposed models, we conduct an empirical analysis using an Italian health insurance dataset, where we compare the performances of the QR, QRNN, and Quantile-CANN. The results highlight that QRNN and Quantile-CANN exhibit better performance in terms of the quantile loss function compared to the classical QR. In particular, we notice that the Quantile-CANN architecture is always able to enhance the estimates given by the QR model. We then exploit the estimates provided by the different models in the calculation of a quantile based insurance premium using the quantile premium principle (QPP) aforementioned.

The analysis shows that premiums produced by the Quantile-CANN and QRNN provide a better risk diversification for the portfolio.

The remainder of the paper is structured as follows: Section 2 explores the two-part model considered in this work. Section 3 introduces the use of QRNN and of Quantile-CANN for the estimation of the quantile of the total claim amount. Section 4 presents the empirical application carried out on an Italian health insurance claim dataset. Section 5 concludes.

2. The Two-Part Quantile Model

In this Section, we discuss the two-part quantile regression framework, first considered by Heras *et al.* (2018), devoted to conditional quantile estimation of the aggregate claim amount S_i at level τ and for a specific policyholder i . Commonly, two-part models involve a mixture distribution consisting in mixing a discrete point mass, with all mass at zero, and a continuous random variable. In particular, they are described by two equations: a binary choice model is fitted for the probability of observing a positive-versus-zero outcome. Then, conditional on a positive outcome, an appropriate regression model is fitted for the continuous outcome. The common structure of such models assumes that the effect of the covariates influence the mean of the conditional distribution of the response. However, in many real-world applications like the actuarial ones, the effect of the covariates can be different on different parts of the distribution. For instance, the gender of the policyholder may be irrelevant when modeling the average claim severity, while it may be strongly significant when studying the upper quantiles of the claim severity. This idea is supported by the results reported in Heras *et al.* (2018), where the authors show that the significance of a specific categorical variable may vary across quantiles. For this reason, a quantile regression approach in the two-part model may be appropriate.

In this paper, we focus our attention on the effect of covariates on the quantile of the aggregate claim amount S_i . In order to build the two-part quantile model (see Duan *et al.*, 1983; Frees, 2010, and Merlo *et al.*, 2021a), we consider the indicator random variable \mathbb{I}_{N_i} measuring whether the policyholder has zero claims or positive claims (where N_i is the number of claims submitted by the policyholder). If $N_i > 0$ then a positive aggregate claim severity \tilde{S}_i is observed.

Consistently with this approach, given a set covariates \mathbf{x}_i , we model the τ -th conditional quantile of the total claim amount $Q_{S_i}(\tau|\mathbf{x}_i)$ in two stages:

- the first stage allows to estimate the no claim probability $p_i = Pr(\mathbb{I}_{N_i} = 0) = Pr(N_i = 0)$ as function of covariates. To achieve this goal, we use the logistic regression:

$$\log\left(\frac{1-p_i}{p_i}\right) = \mathbf{x}'_i\boldsymbol{\beta}, \quad (1)$$

where the no claim probability is obtained as $p_i = \frac{1}{1-\exp(\mathbf{x}'_i\boldsymbol{\beta})}$;

- The second stage uses p_i to obtain the τ_i^* conditional quantile level of \tilde{S}_i , $Q_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i)$, corresponding to the τ quantile level defined on the total claim amount S_i , $Q_{S_i}(\tau|\mathbf{x}_i)$. Following Heras *et al.* (2018), the τ_i^* level can be calculated as

$$\tau_i^* = \frac{\tau - p_i}{1 - p_i} \quad (2)$$

for which

$$Q_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i) = Q_{S_i}(\tau|\mathbf{x}_i). \quad (3)$$

In the literature, $Q_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i)$ is generally calculated using the well known quantile regression approach of Koenker & Bassett (1978). In this paper, we will generalised this approach by introducing two alternative methods, the QRNN, a particular specification of Regression Neural

Network introduced by Taylor (2000), and the Quantile Combined Actuarial Neural Network (Quantile-CANN) which is a new method to calculate quantiles embedding the CANN approach of Schelldorfer & Wüthrich (2019) in a quantile regression framework. These approaches allow performing quantile estimation without imposing any predetermined structure for the relations between the claim severity and the related covariates.

3. Quantile Claim Severity Models

The standard approach to tackle the problem of quantile claim severity estimation refers to traditional QR models, see for example Kudryavtsev (2009), Heras *et al.* (2018) and Baione & Biancalana (2019). In this paper, we generalise this approach by proposing the use of neural network models to estimate the conditional quantile of the aggregate claim severity. This approach allows performing these calculations without imposing any predetermined structure for the relations between the aggregate claim severity and the related covariates. In this way, we can capture nonlinear and complex patterns in the data and possible interactions between predictors.

The first model we introduce is QRNN, which is basically a feed-forward neural network minimising the same quantile loss function minimised by a QR model. The second one is a new model that generalises the CANN approach introduced by Schelldorfer & Wüthrich (2019) by combining the QR model with a QRNN to improve the model's performance, we call this model Quantile-CANN.

Since all the aforementioned models are somehow based on quantile regression, we give a light insight on QR models in the next subsection.

3.1. Quantile regression

Quantile regression, originally introduced by Koenker & Bassett (1978), is a distribution free method providing a way to model the conditional quantiles of a response variable with respect to a set of covariates in order to have a more robust and complete picture of the entire conditional distribution with respect to the classical mean regression. Quantile regression approach is quite suitable method used in all the situations where specific features, like skewness, fat-tails, outliers, truncation, censoring, and heteroscedasticity arise. In this section, we show how to calculate $Q_{\tilde{S}_i}(\tau_i^* | \mathbf{x}_i)$ using QR standard tools, in particular, we analyse how to calculate the quantile of the $\log(\tilde{S}_i)$. The use of the logarithmic function derives from the need to transform the dependent variable from \mathbb{R}^+ to \mathbb{R} to apply the QR approach. Moreover, the use of the logarithmic transformation is coherent within the insurance pricing context, since it allows to consider multiplicative tariffs. In addition, by considering the equivariance under monotone transformation property of the quantile, it is possible to retrieve the quantity of interest, i.e. $Q_{\log(\tilde{S}_i)}(\tau_i^* | \mathbf{x}_i) = \log(Q_{\tilde{S}_i}(\tau_i^* | \mathbf{x}_i))$. For notational simplicity, hereafter, we will use τ^* instead of τ_i^* in the formulas below.

The quantile regression model can be stated as follows:

$$\log(\tilde{S}_i) = \mathbf{x}'_i \boldsymbol{\beta}(\tau^*) + \varepsilon_i, \quad \text{for all } i = 1, 2, \dots, I, \quad (4)$$

where $\boldsymbol{\beta}(\tau^*) = (\beta_1(\tau^*), \beta_2(\tau^*), \dots, \beta_{q_0}(\tau^*)) \in \mathbb{R}^{q_0}$ is the vector of unknown regression parameters and ε_i having the τ^* -th conditional equal to zero for all $i = 1, 2, \dots, I$. The estimation of the of the regression parameters $\boldsymbol{\beta}(\tau^*)$ can be obtained by solving the following minimisation problem:

$$\hat{\boldsymbol{\beta}}(\tau^*) = \underset{\boldsymbol{\beta}(\tau^*)}{\operatorname{argmin}} \frac{1}{I} \sum_{i=1}^I \rho_{\tau^*}(\log(\tilde{S}_i) - \mathbf{x}'_i \boldsymbol{\beta}(\tau^*)), \quad (5)$$

where ρ_{τ^*} is the quantile loss function defined as

$$\rho_{\tau^*}(u) = u(\tau^* - \mathbb{I}_{(u < 0)}) \quad (6)$$

with $\mathbb{I}_{(u)}$ being the indicator function. The conditional quantile of \tilde{S}_i is then estimated as $\hat{Q}_{\tilde{S}_i}(\tau^* | \mathbf{x}_i) = \exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}}(\tau^*))$.

In what follows we explore the model proposed based on the neural network approach.

3.2. Quantile Regression Neural Networks

QRNN is a modeling technique introduced by Taylor (2000) based on the neural networks that enables to estimate the conditional probability distribution of multiperiod financial return within a quantile regression framework. With this approach, it is possible to estimate potential non-linear quantile relations without imposing any distributional assumption or functional relations between dependent and independent variables. Several applications of the methodology have already been implemented in different fields, see for instance Xu *et al.* (2017) and Cannon (2011). Up to our knowledge, the QRNN methodology has never been considered in an insurance pricing context.

At a deeper insight QRNN model, for a fixed depth $K \in \mathbb{N}$, and fixed τ^* , follows a typical feed-forward structure:

$$Q_{\log(\tilde{S}_i)}(\tau^*) = \exp\left\{\boldsymbol{\theta}^{(K+1)}, \left(\mathbf{z}^{(K)}(\boldsymbol{\theta}^{(K)}) \circ \dots \circ \mathbf{z}^{(s)}(\boldsymbol{\theta}^{(s)}) \circ \dots \circ \mathbf{z}^{(1)}(\boldsymbol{\theta}^{(1)})\right) (\mathbf{x}_i)\right\}, \quad \text{for } i = 1, 2, \dots, I, \quad (7)$$

where the output of the network is given by the exponential activation function applied to the scalar product between the readout parameter vector $\boldsymbol{\theta}^{(K+1)}$ returning one neuron in the output layer and the composition of the different K hidden layers $\mathbf{z}^{(1)}(\boldsymbol{\theta}^{(1)}), \dots, \mathbf{z}^{(K)}(\boldsymbol{\theta}^{(K)})$, where $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(K)}$ are the parameters belonging to each layer.

The generic s -th hidden layer $\mathbf{z}^{(s)}(\boldsymbol{\theta}^{(s)})$ of dimension $q_s \in \mathbb{N}$ is defined as

$$\mathbf{z}^{(s)}(\boldsymbol{\theta}^{(s)}) : \mathbb{R}^{q_{s-1}} \rightarrow \mathbb{R}^{q_s}, \quad \mathbf{z}^{(s)}(\boldsymbol{\theta}^{(s)}) = \left(z_1^{(s)}(\boldsymbol{\theta}_1^{(s)}), \dots, z_j^{(s)}(\boldsymbol{\theta}_j^{(s)}), \dots, z_{q_s}^{(s)}(\boldsymbol{\theta}_{q_s}^{(s)})\right)', \quad (8)$$

where the j -th neuron in the s -th hidden layer is given by

$$z_j^{(s)}(\boldsymbol{\theta}_j^{(s)}) = \phi\left(\theta_{j,0}^{(s)} + \sum_{l=1}^{q_{s-1}} \theta_{j,l}^{(s)} \cdot z_l^{(s-1)}(\boldsymbol{\theta}_l^{(s-1)})\right) = \phi\left(\theta_j^{(s)}, \mathbf{z}^{(s-1)}(\boldsymbol{\theta}^{(s-1)})\right), \quad (9)$$

where ϕ is the activation function and $\boldsymbol{\theta}_j^{(s)} = (\theta_{j,0}^{(s)}, \theta_{j,1}^{(s)}, \dots, \theta_{j,q_{s-1}}^{(s)})'$ is the vector of parameters belonging to j -th neuron in the s -th hidden layer. Considering the vector of parameters $\boldsymbol{\theta}_1^{(s)}, \dots, \boldsymbol{\theta}_{q_s}^{(s)}$ for each neuron in (8), we can define the matrix of parameters for the s -th hidden layer as $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\theta}_1^{(s)}, \dots, \boldsymbol{\theta}_{q_s}^{(s)})'$ of dimension $q_s \times (1 + q_{s-1})$ ¹.

Since the network in (7) has K hidden layers, it is possible to denote with $\boldsymbol{\theta}$ the full set of parameters for the network gathering the matrix of parameters of each layer:

$$\boldsymbol{\theta} = \left\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(s)}, \dots, \boldsymbol{\theta}^{(K)}, \boldsymbol{\theta}^{(K+1)}\right\} \quad (10)$$

with dimension r , where $r = \sum_{s=1}^{K+1} q_s(1 + q_{s-1})$.

To obtain the optimal set of parameters $\hat{\boldsymbol{\theta}}$ for (10), we train the network (7) minimising the quantile loss function:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{I} \sum_{i=1}^I \rho_{\tau^*}(\log(\tilde{S}_i) - Q_{\log(\tilde{S}_i)}(\tau^*)) \quad (11)$$

where we fix the starting value of the network parameter $\boldsymbol{\theta}_0$ at the beginning of the training.

From (7) and (11), we estimate $\hat{Q}_{\log(\tilde{S}_i)}(\tau^*)$, then given the equivariance to monotone transformation of the quantile function we retrieve $\hat{Q}_{\tilde{S}_i}(\tau^*) = \exp(\hat{Q}_{\log(\tilde{S}_i)}(\tau^*))$.

¹Note, when $s = 1$ in (9) we have $z^{(0)} = \mathbf{x}_i$ as the input layer in the right hand side of the equation.

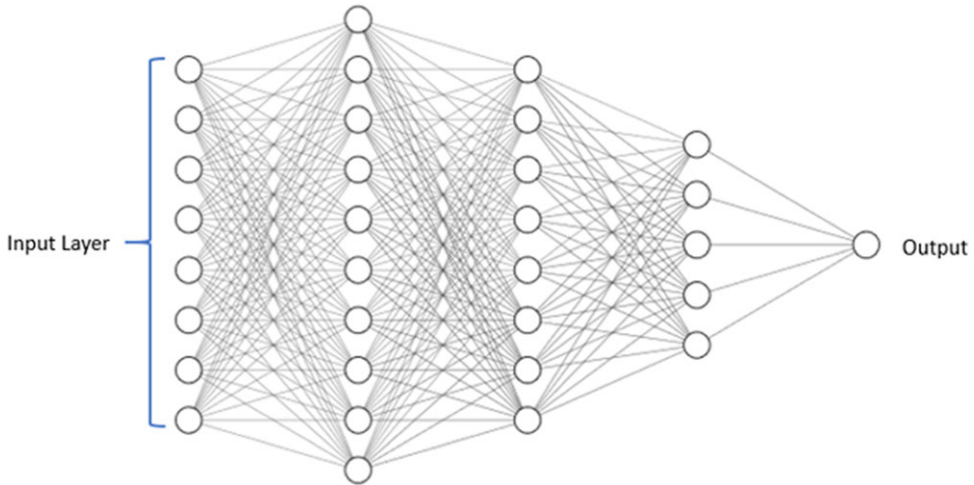


Figure 1. QRNN architecture.

It is worth noting that if we replace the \exp and the ϕ activation functions in (7) and in (9) with the linear activation function and we consider no hidden layers, then the QRNN boils down to the classical QR model in (5). For a visualisation of the QRNN architecture, see Figure 1.

3.3. Quantile-CANN

The innovative model we propose in this section to estimate the quantiles for the distribution of interest is an extension of the combined actuarial neural network (CANN) approach proposed by Schelldorfer & Wüthrich (2019) and launched in the editorial of Wüthrich & Merz (2019). In particular, the CANN framework nests a generic regression model into the neural network architecture to enhance the estimates given by the generic regression model. Following the CANN approach, but in a quantile framework, we boost the QR model with the neural network features proposing the so-called Quantile-CANN model. Our approach allows for exploiting the quantile neural networks to improve the conditional quantile estimates given by a QR model. In such a way, the neural network directly improves the classical QR estimates, preserving the information contained therein.

The main advantage of the Quantile-CANN approach compared to the QRNN is that it combines the flexibility of the networks with the interpretability of a QR approach, so providing higher explainability. According to Wüthrich & Merz (2019), when the reference regression model is already close to optimal, its maximum likelihood estimator can be used as initialisation of the neural network fitting algorithm, we use the QR estimator to initialise the network parameter of the Quantile-CANN, then obtaining lower computational time for the network parameters calibration than the QRNN model.

Formally, we define the Quantile-CANN as

$$Q_{\log(\hat{S}_i)}^{CANN}(\tau^*) = \langle \beta(\tau^*), \mathbf{x}_i \rangle + \left(\theta^{(K+1)}, \left(\mathbf{z}^{(K)}(\theta^{(K)}) \circ \dots \circ \mathbf{z}^{(1)}(\theta^{(1)}) \right) (\mathbf{x}_i) \right), \quad \text{for } i = 1, 2, \dots, I, \quad (12)$$

where the first term of the right hand side of (12) refers to the QR model in (4) with vector of parameters $\beta(\tau^*)$, while the second term is the QRNN model displayed in (7) (except for the missing exponential activation in the output layer). Therefore, the Quantile-CANN model combines the models discussed in the two previous subsections (3.1 and 3.2) by embedding the QR into the network architecture using a skip connection that links the input given by the QR estimates

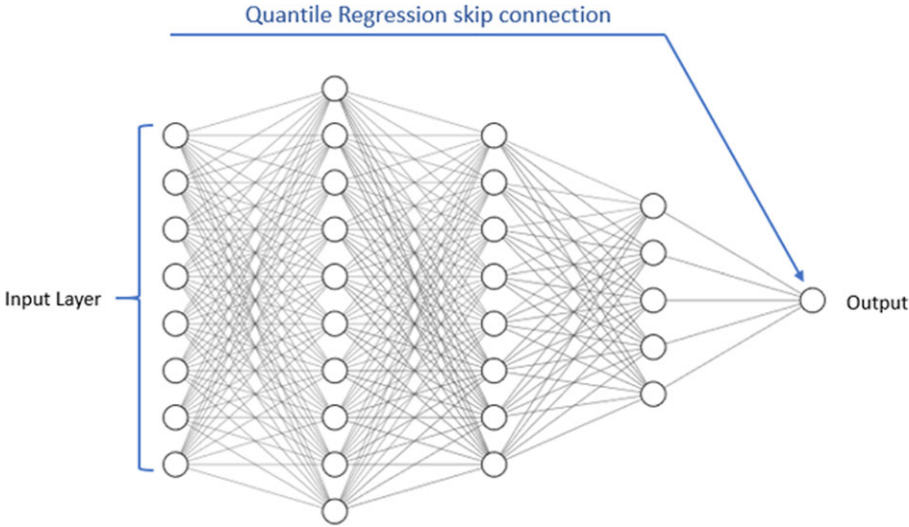


Figure 2. Quantile-CANN architecture.

to the output layer (see Figure 2 for a graphical representation), where the models are merged by summing the two parts as in (12). The network parameter of the Quantile-CANN model is denoted by ϑ and consists of

$$\vartheta = \{\beta(\tau^*), \theta\} = \{\beta(\tau^*), \theta^{(1)}, \dots, \theta^{(K)}, \theta^{(K+1)}\} \tag{13}$$

The optimal set for the network parameter $\hat{\vartheta}$ of (13) is obtained training the Quantile-CANN (12) minimising the quantile loss function:

$$\operatorname{argmin}_{\vartheta} \frac{1}{I} \sum_{i=1}^I \rho_{\tau^*}(\log(\tilde{S}_i) - Q_{\log(\tilde{S}_i)}^{CANN}(\tau^*)). \tag{14}$$

The optimisation process to estimate ϑ in (13) works as follows: we first obtain $\hat{\beta}(\tau^*)$ parameters minimising the quantile loss function for the quantile in (5). Then we use such parameters to initialise the network parameter of the Quantile-CANN by considering $\vartheta_0 = \{\hat{\beta}(\tau^*), \theta_0\}$, where θ_0 is the starting value of the network parameter belonging to the QRNN part of model (12). Therefore, starting from ϑ_0 , we optimise ϑ minimising (14) by means of the gradient descent algorithm². During the optimisation process, both θ_0 and $\hat{\beta}(\tau^*)$ parameters are trained.

Given the optimal set of parameters $\hat{\vartheta}$, from (12), we estimate

$$\hat{Q}_{\log(\tilde{S}_i)}^{CANN}(\tau^*) = \left\langle \hat{\beta}(\tau^*), x_i \right\rangle + \left\langle \hat{\theta}^{(K+1)}, \left(z^{(K)}(\hat{\theta}^{(K)}) \circ \dots \circ z^{(1)}(\hat{\theta}^{(1)}) \right) (x_i) \right\rangle \tag{15}$$

then, the quantile of the claim severity is obtained as $\hat{Q}_{\tilde{S}_i}^{CANN}(\tau^*) = \exp(\hat{Q}_{\log(\tilde{S}_i)}^{CANN}(\tau^*))$.

Note that if the Quantile-CANN model does not return any improvement with respect to the QR model it means that the latter is already able to capture all the relevant information incorporated in the data.

²Gradient descent is an optimisation algorithm used to find the parameters minimising a loss function. On each iteration, the parameters are updated in the opposite direction of the gradient of the loss function and the local minimum is reached by following downhill the direction of the gradient, see Hastie *et al.* (2009).

Table 1. Summary of the variables available in the dataset.

Variable	Description
Covariates	
AG (age)	Age of the insured (in years)
GE (gender)	Gender of the insured (male/female)
PE (permanence)	Years of permanence in the insurance coverage for the insured
RE (region)	Italian Region of residence for the insured (categorical variables with 21 classes)
DM (dimension)	Dimension of the company the policyholder is working for (number of employees), for insured different than the policyholder the dataset reports the value of the policyholder
Binary dependent variable	
\mathbb{I}_{N_i} (claims binary variable)	Binary variable reporting 1 if the insured filed at least a claim and 0 otherwise
Positive dependent variable	
\bar{S}_i (total claim severity)	Total claim severity submitted by the insured during the year (in euros) If the insured submits no claims the dataset reports a blank

4. Case Study

4.1. Data description

To assess the ability of the models proposed to capture additional information incorporated in the data, we carry out an empirical case study based on an Italian health insurance claim dataset. The data stem from an Italian insurer and report the claims collected in a general health insurance plan during 2018. The plan provides risk coverage for managers and retired managers belonging to companies of a specific industry in Italy. More specifically, the dataset consists of 132,499 policyholders (employees or former employees). Since each policyholder can also enroll his relatives (spouse and children below 25 years of age) in the insurance coverage, we totally have 301,405 insured. For each one of them, the following information is available: a binary variable signaling whether the insured submitted at least a claim during the year, total claim severity per year, age, gender, region, firm dimension, and time of coverage permanence (in years). Table 1 provides a summary for the information available in the dataset.

Among the insured, we count a total of 217,006 claimants, the monetary volume of the submitted claims is about euro 243 m.

In Figure 3 and Table 2, we report the histograms and the frequency tables for the variables in the dataset. The age (AG) variable is strongly concentrated at older ages, we notice a “dip” in the distribution between 25 and 40 years, this is due to the specific subscription policies of the Italian insurer: since the policyholders are managers, it is unfrequent they have less than 40 years³ moreover, they are not allowed to enroll in the insurance coverage their children above 25 years of age. That’s why we observe only a small number of insured between 25 and 40 years. The gender (GE) is rather balanced between the two classes. The permanence (PE) is an integer variable that reports the years of permanence in the insurance plan, minimum is 0 (for new comers) and maximum is 41 (for early adopters). Observing the plot in Figure 3, we notice a decreasing trend; however, there is strong peak around 38 years, probably due to the subscription of the policy by a consistent number of companies in the industry. For the region (RE), we observe a strong concentration in 2 of 21 Italian regions: “Lombardia” and “Lazio”, where most firms have their head office. Dimension (DM) is the number of managers of the company to which the policyholder belongs, the value of this variable is the same across all the insured belonging to the same family and ranges from 1 to 1,500, with a strong concentration below 100, which represents the small to medium sized firms (that are specific to the Italian economy).

³The Italian labor market is *seniority driven*, hence it is particularly difficult to become a manager at younger ages.

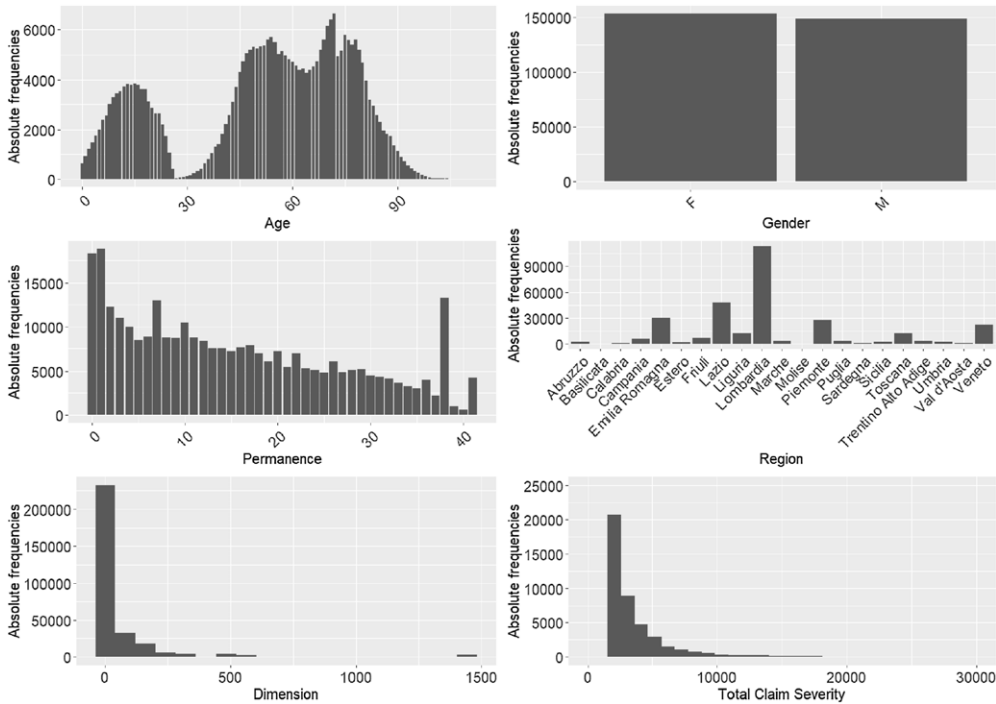


Figure 3. Histograms for the covariates and the total claim severity.

In the right bottom panel in Figure 3, we also plot the histogram of the aggregate claim severity for the 217,006 claimants. We recall that the aggregate claim severity is defined as the sum of the cost of all claims submitted by a given claimant. As it is common in this field, the distribution is right-skewed, from the plot we observe the existence of some large claims; however, the tail doesn't seem to be particularly fat.

4.2. Evaluate model performance

To assess the general performance of the QR, QRNN, and Quantile-CANN models, we use the aforementioned dataset at different τ^* levels. Specifically, the performance of each model is measured in terms of quantile loss function (see Equation (6)), where the lower the loss the better the model.

The results reported in Table 3 are obtained by means of five-fold cross validation, where the dataset is divided in five folds and at each iteration, three of the five folds are used as training set, one as validation and one as testing set. The network models are trained over 2,000 epochs using early stopping on the validation set to avoid overfitting. For the early stopping, we employ a patience parameter of 200 epochs; hence, the training stops when the validation loss does not decrease for at least 200 epochs. When training stops, the network weights obtained at the 200th last epoch are restored and saved as the optimal parameters for the model. Both models adopt a three hidden layer structure of dimension (20, 15, 10), we consider the hyperbolic tangent⁴ activation function for QRNN, while we use the ReLU⁵ activation function for Quantile-CANN. As for the variables presented in Table 1: AG, PE, DM are Min-Max scaled, GE is dummy encoded

⁴The hyperbolic tangent activation function is defined as: $\phi(x) = \tanh(x)$.

⁵The rectified linear unit (ReLU) activation function is defined as: $\phi(x) = x \cdot \mathbb{I}_{x \geq 0}$.

Table 2. Frequency tables for the covariates and the total claim severity.

Variable	Absolute frequency	Relative frequency	Variable	Absolute frequency	Relative frequency
AG (age)			RE (region)		
[0,10)	19,140	0.063579	Lombardia	113,562	0.377227
[10,20)	36,260	0.120448	Lazio	48,066	0.159664
[20,30)	13,639	0.045306	Emilia Romagna	30,063	0.099862
[30,40)	6,333	0.021037	Piemonte	27,436	0.091136
[40,50)	37,997	0.126217	Veneto	22,243	0.073886
[50,60)	52,742	0.175197	Toscana	12,821	0.042588
[60,70)	46,869	0.155688	Liguria	12,227	0.040615
[70,80)	56,896	0.188996	Others	34,626	0.11502
[80,90)	26,454	0.087874	DM (dimension)		
[90,∞)	4,714	0.015659	[0,9)	97,256	0.323062
GE (gender)			[10,19)	10,6676	0.354354
F	152,835	0.507683	[20,49)	35,131	0.116697
M	148,209	0.492317	[50,99)	21,389	0.071049
PE (permanence)			[100,∞)	40,592	0.134837
[0,5)	70,481	0.234122	\tilde{S}_i (total claim severity)		
[5,10)	47,728	0.158542	[1,250)	59,886	0.275965
[10,15)	42,634	0.14162	[250,500)	43,237	0.199243
[15,20)	35,653	0.118431	[500,1,000)	46,224	0.213008
[20,25)	30,173	0.100228	[1,000,2,000)	36,067	0.166203
[25,30)	26,178	0.086957	[2,000,5,000)	24,145	0.111264
[30,35)	19,873	0.066014	[5,000,10,000)	5,845	0.026935
[35,∞)	28,324	0.094086	[10,000,∞)	1,602	0.007382

Table 3. In-sample and out-of-sample quantile loss function at the different τ^* levels.

τ^*	In-sample			Out-of-sample		
	QR	QRNN	Q-CANN	QR	QRNN	Q-CANN
0.7	487.91	485.00	484.32	488.09	485.84	482.58
0.75	486.34	482.25	481.61	486.55	483.06	480.38
0.8	472.49	467.35	466.75	472.74	468.51	465.32
0.85	441.02	436.11	435.26	441.33	437.04	434.30
0.9	382.15	378.30	377.75	382.52	379.40	376.82

and RE is treated using a $d = 1$ embedding layer. As for the QR model, we consider a splines function to model the AG and the PE effects.

For each model, we estimate the conditional quantile of the total claim amount at levels $\tau^* = (0.7, 0.75, 0.80, 0.85, 0.9)$ and compute the respective in sample and out of sample quantile loss function to evaluate their performance.

For our dataset, QRNN and Quantile-CANN exhibit an overall better performance in terms of the quantile loss function compared to the classical QR for each quantile level τ^* (see Table 3). It is interesting to note that Quantile-CANN always yields a lower quantile loss function compared

to the QRNN, suggesting that building the network around the QR has not only improved the performance given by the QR model but also beats the QRNN. The results above illustrated seem to suggest that using a neural network approach can be appealing. However, such models often face one major drawback: the lack of explainability. Indeed, neural networks have a huge number of parameters and a complex inner structure made up of several hidden layers, that make very difficult for the modeler to understand the results. To overcome such limitations, in recent years, a wide literature covering the topic of model agnostic tools has flourished, see Friedman & Popescu (2008a) and for an actuarial case study Lorentzen & Mayer (2020) or Henckaerts *et al.* (2021), aiming at providing interpretative tools for machine learning models.

The corresponding literature has a plethora of well established model agnostic techniques. In this work, we exploit the permutation variable importance of Breiman *et al.* (1984) to gauge the relevance of each covariate in the model; the ICE curves and the PD plots to showcase the marginal effect of a covariate over the prediction produced by the model; the H-squared statistic Friedman & Popescu (2008a) in order to spot potential interaction effects between the covariates.

4.3 Variable importance

Permutation variable importance of Breiman *et al.* (1984) measures the increase in the deviance of a model after permuting the values of a given covariate. The basic idea is quite simple: the importance of a covariate is measured by calculating the increase in the model's deviance after permuting the covariate. In other terms, the variable importance is the increase in model deviance when the information provided by an explanatory variable is destroyed. Such variable is deemed important if randomly shuffling its values increases the model deviance, because in this case the model relied on the feature for the prediction. While, a covariate is not important if shuffling its values produces little to no increase in the model's deviance

In what follows the general framework for the algorithm.

Let f be the generic prediction function given by a model, where \mathbf{x} is the feature matrix, y is the vector of observations and $D(y, f)$ is the model's deviance. For instance in QR we have $f = \exp(\mathbf{x}'\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}^*))$ and $D(y, f) = \rho_{\boldsymbol{\tau}^*}(y - f)$.

1. We estimate the original model Deviance $D_0 = D(y, f(\mathbf{x}))$;
2. For each covariate $j = 1, \dots, p$:
 - take the covariate matrix \mathbf{x}_j , that can also be represented as $\mathbf{x}_j = \{\mathbf{x}_{.,j}, \mathbf{x}_{.,-j}\}$, where $\mathbf{x}_{.,j}$ is the column vector belonging to covariate j and $\mathbf{x}_{.,-j}$ is the matrix for the other covariates. Get the set of covariates \mathbf{x}^{perm} by permuting the values in $\mathbf{x}_{.,j}$;
 - estimate model deviance $D_{perm} = D(y, f(\mathbf{x}^{perm}))$;
 - compute the Permutation Variable Importance for covariate j as $I_j = D_{perm} - D_0$
3. Sort covariates by descending order I_j for $j = 1, \dots, p$.

Permutation variable importance can be either performed on the training set or the test set. Performing the analysis on the test set informs the modeler on how much the model counts on each covariate for the predictions, while using the test set would give a hint on the relevance of the covariate for the performance of the model on new and unseen data.

In Figure 4, we report the variable importance metric to find the most relevant variables in our dataset for the quantile models. The variables are ranked from top to bottom, starting with the most important one as measured by the variable importance. For all models, the most important variable is the Age (AG) followed by the spatial variable (RE), the other variables are far less relevant; however, they seem to have a somewhat higher importance in network models with respect to the QR.

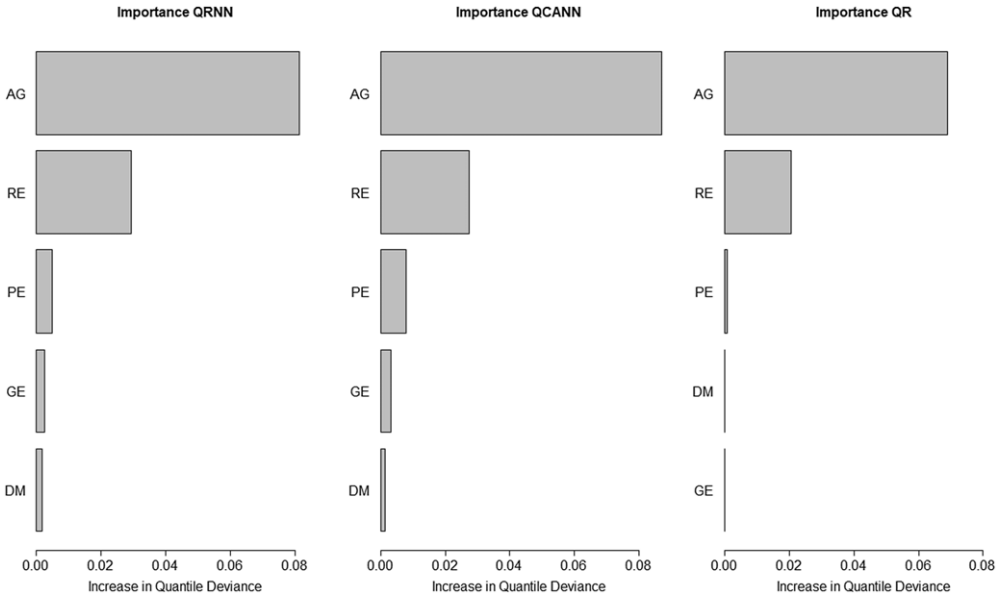


Figure 4. Variable importance for QRNN (left), Quantile-CANN (middle) and QR (right), trained at level $\tau^* = 0.8$. The results report are obtained on the first fold of the 5-fold cross validation.

4.4 Main effects

ICE profiles are a useful tool to study the marginal effect of a covariate over the response provided by the model. Such a profile, for a given covariate j , shows how the prediction provided by the model for an observation $obs_i = \mathbf{x}_i$ reacts when the covariate $\mathbf{x}_{i,j}$ slides over its range of possible values.

In particular, producing ICE profiles for a set of observations gives a hint on how the response evolves with respect to the different values of the variable. An ICE plot represents the relationship between the prediction and a specific covariate for each single observation separately, producing one profile (or line) per observation. The values for a line of the data matrix is obtained by fixing the values of all other covariates and creating variants of this observation by replacing the covariate’s value with values coming from a grid and producing predictions with the model for these new observations. This procedure, for a specific observation, results in a set of points given by the feature value from the grid and the respective predictions. From an algorithmic standpoint, we have:

1. take an observation $obs_i = \mathbf{x}_i$ and the corresponding prediction $f(\mathbf{x}_i)$ given by the model.
2. \mathbf{x}_i is a p dimensional vector that can also be represented as $\mathbf{x}_i = \{\mathbf{x}_{i,j}, \mathbf{x}_{i,-j}\}$, where j is the covariate we want to study.
3. Consider the grid of V possible values (v_1, v_2, \dots, v_V) for the selected covariate j .

Then for each v_k with $k = 1, \dots, V$ we repeat:

- $\mathbf{x}_{i,j} = v_k$.
- $obs_i = \mathbf{x}_i = \{v_k, \mathbf{x}_{i,-j}\}$.
- $ICE_{i,v_k}^j = f(obs_i)$.

Once the process ends, we obtain a curve $\{ICE_{i,v_k}^j\}_{k=1}^V$ corresponding to $\{v_k, \mathbf{x}_{i,-j}\}_{k=1}^V$. The process is repeated potentially for each observation in the dataset and each variable.

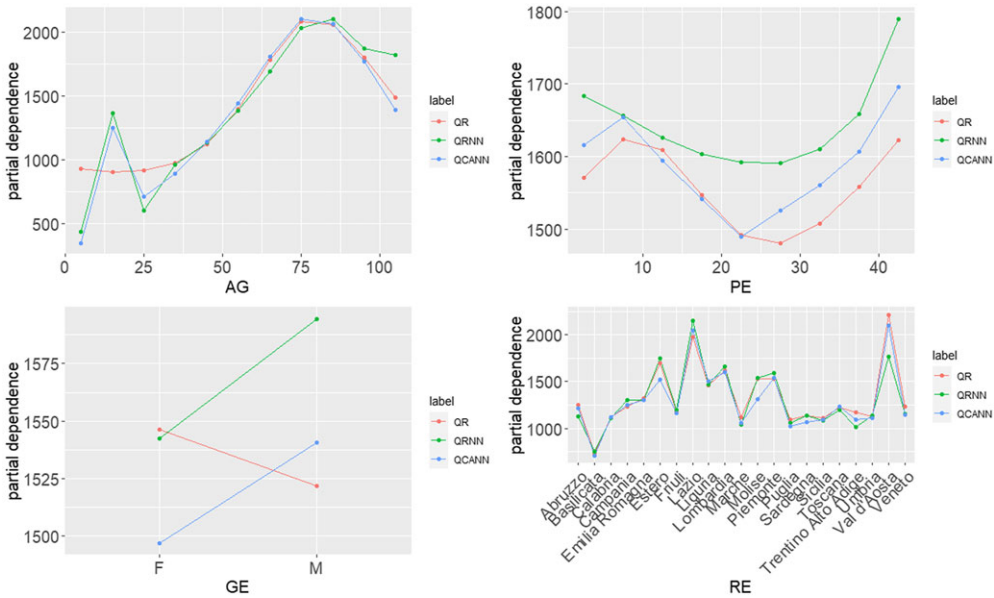


Figure 5. Partial dependence plot for the age of the insured (AG), years of permanence (PE), gender (GE), and region of the insured (RE). The models are trained on the first fold of the 5-fold cross validation and fitted at level $\tau^* = 0.8$. The curves are obtained averaging ICE profiles plotted using 200 randomly selected observations in the training set.

The ICE profiles are useful to highlight the presence of interactions. Infact, the stronger the interaction effects associated with the variable j , the greater the differences in shape observed across ICE profiles. However, this model agnostic tool does not reveal with which other variable the interaction arises. Note that, by construction, ICE profiles of a given covariate j are parallel as long as the underlying model does not incorporate interactions (like in a GLM or in a QR on the log scale transformation).

The partial dependence (PD) profiles of Friedman & Popescu (2008b) are obtained averaging different ICE profiles of a given variable j . PD profiles can be viewed as the main effect of covariate j merged over all observations. In other terms, they represent the average effect of variable j and are able to display whether the relationship between the response and a covariate is linear, monotonic, or more complex.

PD marginalises the model’s prediction over the distribution of the covariates in $\mathbf{x}_{\cdot,-j}$, so that the profile displays the relationship between variable j and the output produced by the model.

If we consider the ICE profiles obtained for variable j over a set of n observations, ICE_{i,v_k}^j where $i = 1, \dots, n$ and $k = 1, \dots, V$, the PD profile is obtained as

$$PD^j = \left(PD_{v_1}^j, \dots, PD_{v_V}^j \right), \quad \text{with} \quad \left\{ PD_{v_k}^j = \frac{1}{n} \sum_{i=1}^n ICE_{i,v_k}^j \right\}_{k=1}^V, \quad (16)$$

where, again, V is the grid of possible values for the selected covariate. The $PD_{v_k}^j$ function, for a given value v_k of variable j , reveals the average marginal effect on the prediction returned by the model. Note that PD profiles are not restricted to a single variable, it is also possible to consider multiple variables at the same time in order to study their joint effect on the response.

In Figures 5 and 6, we consider PD plots and individual conditional expectations (ICEs) to gain an understanding of the main effects of the variables over the conditional quantile of the total claim severity, for the different models. The different ICE profiles in Figure 6 were coloured

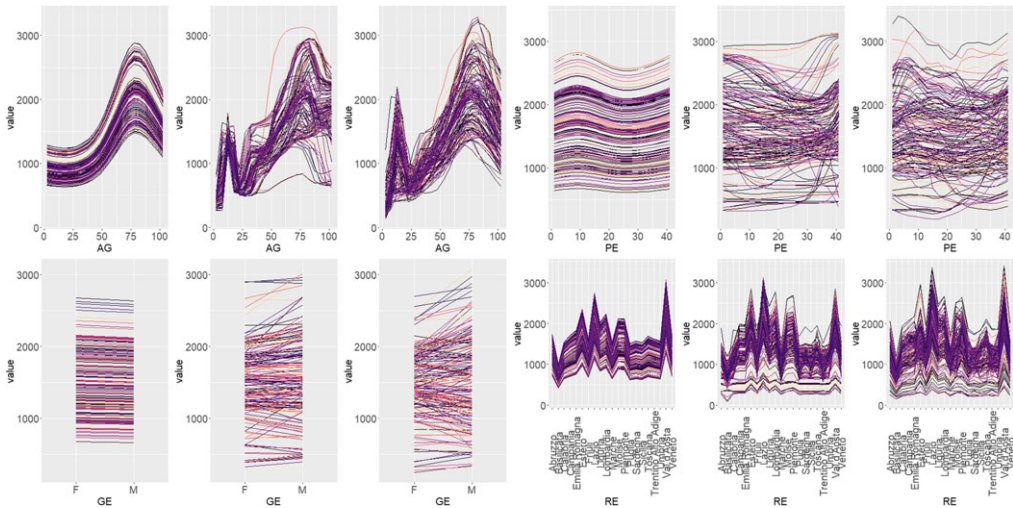


Figure 6. ICE profiles for the age of the insured (AG), years of permanence (PE), gender (GE), and region of the insured (RE). The models are trained on the first fold of the 5-fold cross-validation and fitted at level $\tau^* = 0.8$. The curves are plotted using 200 randomly selected observations in the training set.

to report the amount of the observed claim (the darker the colour the higher the amount) to detect possible peculiar profiles. The top-left plot in Figure 5 compares the PD plots for the AG variable produced by the different models. At first glance, the curves look quite similar, however taking a closer look we notice an important difference in the leftmost part of the plot, for the values below 25 years of age. More specifically, for both QRNN and Quantile-CANN, we observe an upward trend in the riskiness of the insured, starting from 500 euros and peaking at 1,500 euros at around 15 years of age, then the curve falls off down to 700 euros at 25 years. While the QR shows a flat trend kicking off at around 1,000 euro and slowly increasing after 25 years of age. The behaviour displayed by network models seems to be more reasonable, since it captures the cost for dental treatment in younger ages, as it is usually low for children below 10 years, while it raises significantly for teenagers often associated with the use of dental braces. The ICE curves associated with the PD plot displayed in the top-left plot of Figure 6 do not seem to reveal specific interactions within the neural network models, since all the profiles look almost parallel in the plot, which is the case when the variable has little to no interactions with other variables. Of course, the ICE profiles for the QR are always parallel (on a log scale), since the variables are always modelled additively.

The permanence (PE) PD plots display a similar evolution but at different levels (top-right pane in Figure 5). Furthermore, for this variable, we observe some wild ICE profiles (top-right pane in Figure 6), that may signal the presence of interactions with other variables. The Gender (GE) variable does not seem to have a particular behaviour, in fact, the range for the y -axis is pretty narrow. As for the regional variable (RE), the PD plots show almost identical profiles, with a particularly riskiness associated with insured living in Lazio and Valle d'Aosta.

4.5 Interaction effects

After having a look at main effects and at some clues on possible interactions in Section 4.4, here we closely study possible interaction effects between covariates captured by the network models. When a given model incorporates an interaction effect, its predictions cannot be expressed as the sum of its variables' main effects, because the effect of one variable depends on the value of another variable. To assess the presence of interaction effects, we adopt the H-statistic introduced

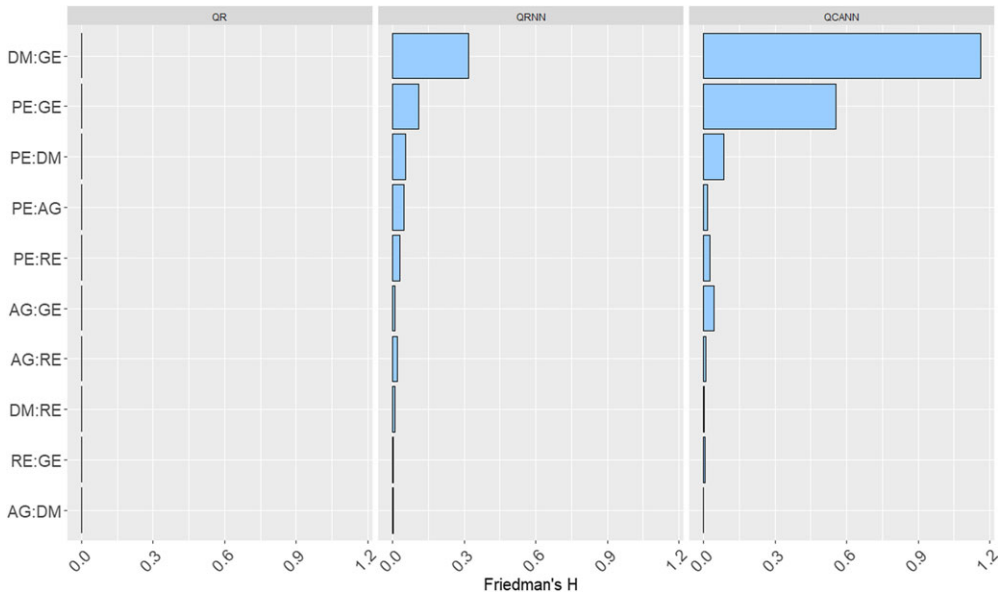


Figure 7. H-statistic of the possible two-way interactions for the different quantile models, fitted at level $\tau^* = 0.8$ and trained on the first fold of the 5-fold cross-validation.

by Friedman & Popescu (2008b), which estimates the interaction strength between two covariates by measuring how much of the prediction variance originates from their interaction. To measure pairwise interaction strength between covariates j and k , H-statistic is defined as follows:

$$H_{jk}^2 = \sum_{i=1}^n \left[\bar{PD}^{jk} - \bar{PD}^j - \bar{PD}^k \right]^2 / \sum_{i=1}^n (\bar{PD}^{jk})^2 \quad (17)$$

where the sums run over a subset of n randomly selected observations, \bar{PD}^k is the centered version⁶ of the PD profile for variable k , and \bar{PD}^{jk} is the centered two-way PD for variable j and k . In other words, H_{jk} measures the proportion of variability in the joint effect of $\mathbf{x}_{.,j}$ and $\mathbf{x}_{.,k}$ unexplained by their main effects. A value close to zero indicates almost no pairwise interaction, while a value close to one means that most effects come from the pairwise interaction. The H-statistic can also be larger than 1, this can happen when the variance of the joint interaction is larger than the variance of the 2-dimensional PD plot. Hence, the interaction strength is measured as the share of variance explained by the interaction.

In Figure 7, we plot the values of the H-statistic for each model and each possible pairwise interaction. As discussed above, the QR model does not display any interaction between covariates, since it is not designed to do so. While, both QRNN and Quantile-CANN display some specific interactions. The strongest common interaction between the two models is the one between the firm's dimension (DM) and the gender of the insured (GE), followed by the one between the Permanence (PE) and, again, the Gender (GE). To gain insight on the behaviour of the interaction effects, in Figures 8 and 9, we report PD plots for Dimension (DM) and Permanence (PE), grouped with respect to Gender (GE). The interaction appears to be relevant if the curves display a different behaviour when conditioned to the different values of the gender variable.

⁶The centered version for the partial dependence profile is obtained via Equation (16) where instead of using the ICE profiles discussed in Section 4.4 we consider a set of ICE profiles that are centered around zero.

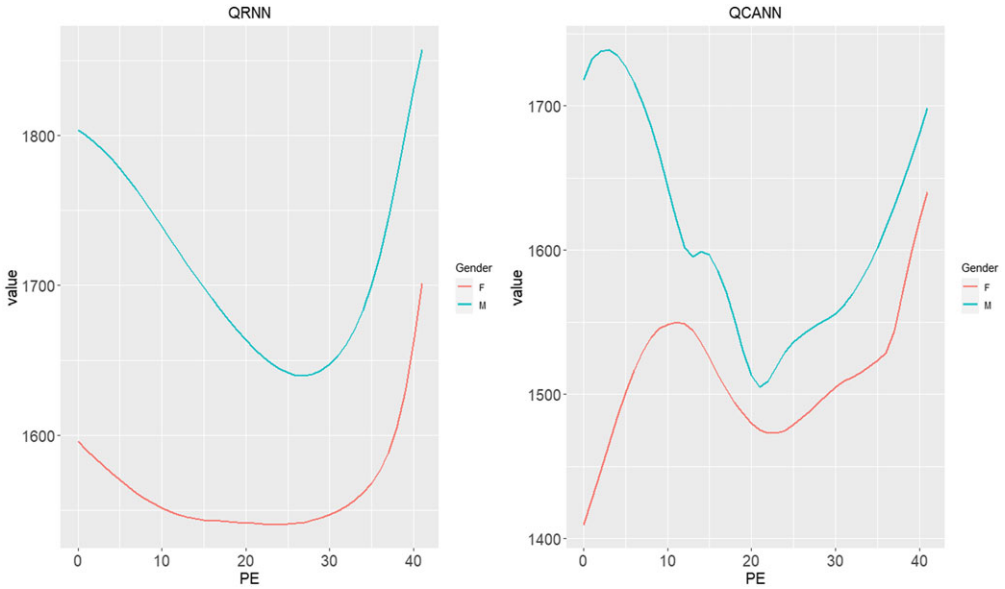


Figure 8. Grouped partial dependence plots for the permanence variable with respect to gender.

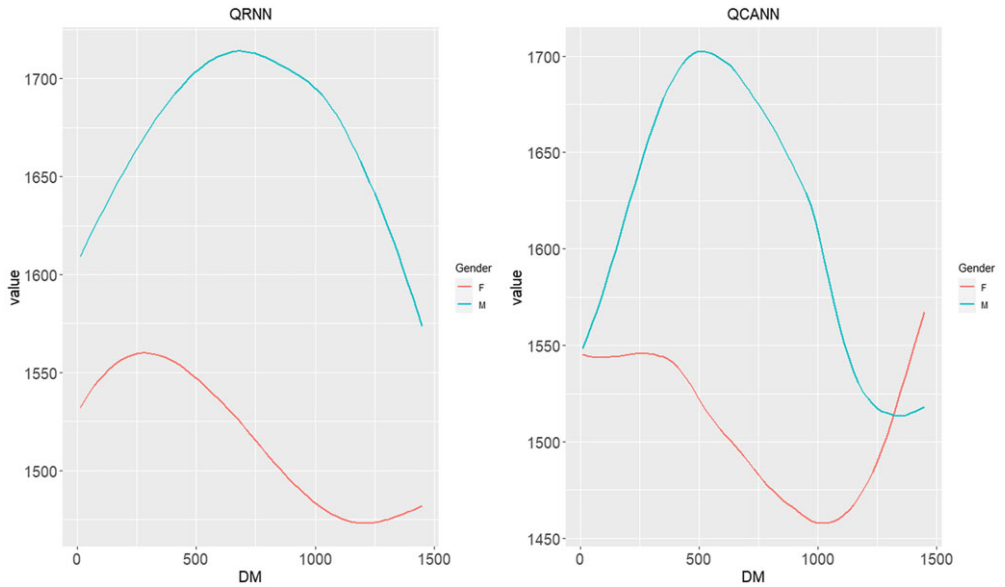


Figure 9. Grouped partial dependence plots for the dimension variable with respect to gender.

For the first interaction, in Figure 8, we notice that both network models recognise higher riskiness to male insured, indeed in the two panels, the blue line (in Figure 8) lies always above the red curve. However, there are some relevant differences between QRNN and Quantile-CANN that are worthy of discussion. In particular, in the QRNN plot, the two curves display a downward parabolic trend which is rather steep for male insured and far flatter for females. Both curves reach their minimum at around 25 years of permanence and then they start increasing again,

reaching 1, 900 euro for males and 1, 700 euros for females, the difference between the two curves is rather stable across the plot (though not constant). For Quantile-CANN, we observe a somewhat different behaviour, in fact, the two curves display a large distance for the insured with less than 10 years of permanence, while in the right part of the plot the distance in terms of potential riskiness between males and females is almost negligible.

Also for the interaction between dimension (DM) and gender (GE), the grouped PD plots (Figure 9) report an higher riskiness associated with male insured. Both panels show an upward parabolic trend with a maximum at around 500 for both plots for the PD plots associated with males insured. In the QRNN panel, we notice a quasi-sinusoidal trend for the females grouped PD plot, with a very low variability for the values in the y -axis. In other terms, for the QRNN, the dimension variable only has some sort of significant effect for males, while for females the effect of this variable seems negligible. As for the Quantile-CANN female plot, we have first had a substantially stable trend, then followed by a downward parabolic trend, with a good variability for the insured potential riskiness.

Thus, even though the gender variable did not appear to be significant in Figures 4 and 7, it has some relevance when interacting with other covariates. As shown above, network models have proven to be good at detecting such interactions.

4.6 Ratemaking

In Section 4.2, we have investigated models' behaviours evaluated at different quantile levels. We are now interested in evaluating models' performances when the focus is estimating the quantile premium principle introduced by Heras *et al.* (2018), where the premium paid by the insured is loaded according to its potential riskiness. In particular, we consider the convex combination of the quantile claim severity $Q_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i)$ and the conditional expected value of the total claim amount S_i :

$$P_i = \gamma \cdot \hat{Q}_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i) + (1 - \gamma) \cdot E(S_i|\mathbf{x}_i), \quad (18)$$

where $0 < \gamma < 1$ is the loading factor and $E(S_i|\mathbf{x}_i)$ the conditional expected value of the total claim amount that can be decomposed as $E(S_i|\mathbf{x}_i) = Pr(N_i > 0|\mathbf{x}_i) \cdot E(\tilde{S}_i|\mathbf{x}_i)$. Note that $Pr(N_i > 0|\mathbf{x}_i) = 1 - p_i$ can be obtained from Equation (1), while $E(\tilde{S}_i|\mathbf{x}_i)$ is estimated using a Gamma regression model as in Frees (2010). In order to compute (18), we need to obtain $\hat{Q}_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i)$ using the two-part model discussed in Section 2.

The two-part model is estimated in taking the first fold in the 5-fold cross-validation, where we consider a 60-20-20 split between learning, validation, and testing set. The first step of the two-part model consists in estimating the no claim probability p_i for each insured using logistic regression. Then as discussed in Section 2, the estimated p_i is employed to compute the τ_i^* level as in Equation (2) for each insured setting $\tau = 0.95$. As a result of this first step, we obtain 47,120 unique values for τ_i^* ranging between 0.799 and 0.896. Following the two-part approach designed by Heras *et al.* (2018) would involve fitting a regression model for each different quantile level τ_i^* . Doing this would be rather time-consuming. Thus, to avoid that, we approximate the τ_i^* values up to the second digit, where the second digit is rounded to the closest even digit⁷ we perform the second step of the two-part model by computing the conditional quantile of the claim severity $\hat{Q}_{\tilde{S}_i}(\tau_i^*|\mathbf{x}_i)$ using QR, QRNN, and Quantile-CANN.

In order to test the ability of the different models to accurately estimate the desired quantile level of \tilde{S}_i , we use the backtest criteria approach. More specifically, the models are backtested using the unconditional coverage (UC) test proposed by Kupiec (1995), which is a widespread testing technique generally employed to validate VaR models in the financial literature. This technique

⁷For instance: $\tau_i^* = 0.815$ is rounded up to 0.82, while $\tau_i^* = 0.805$ is rounded down to 0.80. This approximation results in six different quantile levels τ_i^* for the testing set.

Table 4. For the different models we report the values for the LR_{uc} statistic and its corresponding p-values. The critical values of the LR_{uc} statistic is 3.84, denoting that the null hypothesis is rejected at the 5% significance level. The asterisk indicates that the model passes the test.

τ^*	LR_{uc}			P-values		
	QR	QRNN	Q-CANN	QR	QRNN	Q-CANN
0.8	0.02	9.19	5.51	0.86*	0.00	0.01
0.82	3.94	0.02	1.19	0.04	0.86*	0.27*
0.84	0.53	0.00	0.09	0.46*	0.99*	0.75*
0.86	18.68	0.35	0.08	0.00	0.55*	0.77*
0.88	1.51	1.30	0.73	0.21*	0.25*	0.39*
0.9	0.09	0.31	0.37	0.76*	0.57*	0.54*

consists in a binomial test checking if the proportion of insured with a claim severity above the conditional quantile is consistent with the predefined quantile level τ^* . The UC test performs a likelihood ratio test, where the null hypothesis of the test states that the probability of a violation⁸ is equal to $\tau_c = (1 - \tau^*)$.

More formally, given a generic insured i , we can define the violation function as

$$I_i(\tau^*) = \begin{cases} 1 & \text{if } \tilde{S}_i > \hat{Q}_{\tilde{S}_i}(\tau_i^* | \mathbf{x}_i) \\ 0 & \text{if } \tilde{S}_i \leq \hat{Q}_{\tilde{S}_i}(\tau_i^* | \mathbf{x}_i) \end{cases} \tag{19}$$

Hence, given a portfolio composed of I insured, we define the total amount of the violations occurred in the portfolio as $I(\tau^*) = \sum_{i=1}^I I_i(\tau^*)$, and the ratio of occurred violations as

$$\hat{\tau}_c = \frac{I(\tau^*)}{I} \tag{20}$$

It is now possible to define the UC test statistic as

$$LR_{uc} = -2 \log \left[\left(\frac{1 - \tau_c}{1 - \hat{\tau}_c} \right)^{I - I(\tau)} \left(\frac{\tau_c}{\hat{\tau}_c} \right)^{I(\tau)} \right] \tag{21}$$

The null and the alternative hypothesis for the UC test are defined as

$$H_0 : \hat{\tau}_c = \tau_c \quad H_1 : \hat{\tau}_c \neq \tau_c \tag{22}$$

In other terms, considering a portfolio of I insured, if the number of occurred violations $\hat{\tau}_c \cdot I$ is close enough to $\tau_c \cdot I$, the test statistic LR_{uc} is low, and the null hypothesis is not rejected. There is no evidence of any inadequacy for the tested quantile estimate. While, if the number of violations strongly differs from $\tau_c \cdot I$, the test statistic increases indicating growing evidence that the proposed quantile either systematically understates or overstates the portfolio’s potential riskiness, and thus the null hypothesis is rejected. In Table 4, we report the backtesting results for QR, QRNN, and Quantile-CANN. In the left part of the table, we report the unconditional coverage test statistic LR_{uc} , while on the right side of the table, we display the corresponding p-values, keeping in mind that the null hypothesis is not rejected when the p-value is larger than 0.05. From Table 4, we observe that QRNN and Quantile-CANN pass the backtest at almost all quantile levels since the null hypothesis of unconditional coverage is not rejected. More specifically, the QRNN and Quantile-CANN pass the test at all quantile levels, with the sole exception of the

⁸We have a violation when we observe an insured submitting a larger total claim severity w.r.t. the estimated conditional quantile at level τ_c .

Table 5. Two-way comparison of Gini Indices for the models.

Base model	Competing model		
	QR	QRNN	Q-CANN
QR	–	4.47	5.83
QRNN	0.31	–	2.83
Q-CANN	–1.26	0.22	–

$\tau^* = 0.8$ level, while QR fails the backtest at two quantile levels: 0.86 and 0.82. Therefore, QRNN and Quantile-CANN seem more able to accurately estimate the quantile of the total claim severity $\hat{Q}_{\hat{S}_i}(\tau_i^* | \mathbf{x}_i)$.

Once the models are estimated, we are able to calculate tariffs P_i as in Equation (18) and evaluate them using the ordered Lorenz curve introduced by Frees *et al.* (2014). This tool is a twist on the classical Lorenz curve usually employed in welfare economics to represent social inequality via the Gini index, see Farris (2010). In insurance literature, the ordered Lorenz curve is employed to compare different tariff structures issued by a set of competing models. Given a base tariff structure P_i^{base} and a competing tariff P_i^{comp} the Lorenz curve proposed by Frees *et al.* (2014) is ordered with respect to the relativity r_i :

$$r_i = \frac{P_i^{comp}}{P_i^{base}} \quad (23)$$

A relativity r_i consistently below 1 reveals a largely profitable policy for the company, that is likely to be lost to a competing insurance company proposing a cheaper premium. Instead a relativity r_i greater than 1 signals an underpriced policy. Of course these statement hold true only if we assume P_i^{comp} to give a sharper representation of the real risk compared to P_i^{base} .

Given r_i , the ordered Lorenz curve can be defined as follows:

$$\left(\frac{\sum_{i=1}^n S_i \mathbb{I}\{F_n(r_i) \leq s\}}{\sum_{i=1}^n S_i}, \frac{\sum_{i=1}^n P_i^{base} \mathbb{I}\{F_n(r_i) \leq s\}}{\sum_{i=1}^n P_i^{base}} \right). \quad (24)$$

for $s \in [0, 1]$ where $F_n(r_i)$ is the empirical cumulative distribution function of the relativities r_i . The idea behind the ordered Lorenz curve is that a model producing tariffs with a greater Gini index produces a stronger separation among premiums paid by the insured, signaling that such model is more capable to distinguish good risks from bad risks. Hence, a tariff structure P_i^{comp} that yields a larger Gini index is likely to result in a more profitable portfolio because of a better risk differentiation.

Table 5 displays the two-way comparison of Gini indices for the network models and the QR. The rows report the model generating the base tariff structure P_i^{base} whereas the column stores the model from which the competing tariff structure P_i^{comp} is generated. The approach we use for selecting a tariff based on the Gini index is the “mini-max” strategy designed by Frees *et al.* (2014), which consists of selecting the model that provides the minimum Gini index among the maximal Gini indices taken over the competing models. The strategy is rather intuitive: if we have to choose a base premium we chose the one that has the minimal maximum improvement when compared to other models, meaning that the selected base premium is the least vulnerable to alternatives. In practice, we look for the base model with the lowest value in bold in Table 5. The Quantile-CANN appears as a clear winner since the other models are not able to achieve an high Gini index when considered as an alternative (see the last row), signaling that this model leads to a tariff structure that is the least likely to incur in adverse selection. The QRNN tariff structure achieves second place, followed by QR.

5. Conclusions

In the domain of insurance ratemaking, neural networks may be considered a tool to perform high-dimensional nonlinear regressions. Following this line, in this paper, we extend such techniques in order to estimate the conditional quantile of the total claim amount in the context of a two-part model devoted to the definition of a loaded premium. To fulfill the quantile estimation task, we propose two models. First, we use QRNN of Taylor (2000), which is a particular neural network specification that has never been applied before in actuarial sciences. Next, we generalise the CANN approach of Schelldorfer & Wüthrich (2019) introducing the Quantile-CANN, a flexible tool that merges the classical QR with a QRNN allowing to improve the results given by the QR model and the QRNN.

In the first part of the empirical application, Section 4.2, we test the performance of the proposed network models against the traditional QR model over a health insurance claim dataset. The results show that our models outperform QR in terms of quantile loss function. Furthermore, for QRNN and Quantile-CANN, we apply the set of model agnostic tools discussed in Lorentzen & Mayer (2020) to gain additional insight in the dataset.

In the second part of the empirical application, Section 4.6, we compute the Quantile Premium Principle introduced by Heras *et al.* (2018) using the different models discussed in the paper (QRNN, Quantile-CANN, and QR). To compare the tariff structures issued by the models, we adopt the ordered Lorenz designed by Frees *et al.* (2014). According to this technique, the proposed network models exhibit a better tariff structure w.r.t QR since they provide a better risk differentiation for the portfolio. This feature is paramount for the insurer since a better differentiation between good and bad risks is likely to increase profits.

This work focuses on the comparison between the neural network-based models and the quantile regression, further research could explore different network configurations, i.e., increasing the number of dimensions for the embedding layer used for the regional variable or even different machine learning models, such as tree-based models or gradient boosting machines. Another possible development of this work could consider a multivariate QRNN or Quantile-CANN approach in order to jointly model the conditional quantile of the total claim severity for different and possibly correlated claim types.

References

- Adrian, T. & Brunnermeier, M.K. (2016). Covar. *American Economic Review*, **106**(7), 1705–1741, predicting and measuring a financial institution's contribution to systemic risk that internalizes externalities and avoids procyclicality. URL [StaffReports](#)
- Baione, F. & Biancalana, D. (2019). An individual risk model for premium calculation based on quantile: a comparison between generalized linear models and quantile regression. *North American Actuarial Journal*, **23**(4), 573–590.
- Breiman, L., Friedman, J., Stone, C. & Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis. Available online at the address <https://books.google.it/books?id=JwQx-WOmSyQC>
- Cannon, A.J. (2011). Quantile regression neural networks: implementation in r and application to precipitation down-scaling. *Computers and Geosciences*, **37**(9), 1277–1284. Available online at the address <http://www.sciencedirect.com/science/article/pii/S009830041000292X>
- Duan, N., Manning, W.G., Morris, C.N. & Newhouse, J.P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, **1**(2), 115–126. Available online at the address <http://www.jstor.org/stable/1391852>
- Farris, F.A. (2010). The gini index and measures of inequality. *The American Mathematical Monthly*, **117**(10), 851–864. Available online at the address <http://www.jstor.org/stable/10.4169/000298910x523344>
- Frees, E.W. (2010). *Regression Modeling with Actuarial and Financial Applications*. International Series on Actuarial Science. Cambridge University Press, New York, USA.
- Frees, E.W.J., Meyers, G. & Cummings, A.D. (2014). Insurance ratemaking and a gini index. *The Journal of Risk and Insurance*, **81**(2), 335–366. Available online at the address <http://www.jstor.org/stable/24546807>
- Friedman, J.H. & Popescu, B.E. (2008a). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, **2**(3), 916–954. Available online at the address <http://www.jstor.org/stable/30245114>

- Friedman, J.H. & Popescu, B.E. (2008b). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 916–954. Available online at the address <https://doi.org/10.1214/07-AOAS148>
- Gabrielli, A., Richman, R. & Wüthrich, M.V. (2020). Neural network embedding of the over-dispersed poisson reserving model. *Scandinavian Actuarial Journal*, 2020(1), 1–29. Available online at the address <https://doi.org/10.1080/03461238.2019.1633394>
- Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3), 3659–3667.
- Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition. Springer. Available online at the address <http://www-stat.stanford.edu/tibs/ElemStatLearn/>
- Henckaerts, R., Côté, M.-P., Antonio, K. & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2), 255–285.
- Heras, A., Moreno, I. & Vilar-Zanón, J. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 2018, 1–17.
- Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1), 33–50.
- Kudryavtsev, A.A. (2009). Using quantile regression for rate-making. *Insurance: Mathematics and Economics*, 45(2), 296–304. Available online at the address <https://EconPapers.repec.org/RePEc:eee:insuma:v:45:y:2009:i:2:p:296-304>
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2), 73–84.
- Laporta, A.G., Merlo, L. & Petrella, L. (2018). Selection of value at risk models for energy commodities. *Energy Economics*, 74, 628–643. Available online at the address <http://www.sciencedirect.com/science/article/pii/S0140988318302548>
- Lorentzen, C. & Mayer, M. (2020). Peeking into the black box: an actuarial case study for interpretable machine learning. Available at SSRN. <https://ssrn.com/abstract=3595944>
- Merlo, L., Maruotti, A. & Petrella, L. (2021a). Two-part quantile regression models for semi-continuous longitudinal data: a finite mixture approach. *Statistical Modelling*, 1471082X21993603.
- Merlo, L., Petrella, L. & Raponi, V. (2021b). Forecasting var and es using a joint quantile regression and its implications in portfolio allocation. *Journal of Banking & Finance*, 133, 106248. Available online at the address <https://www.sciencedirect.com/science/article/pii/S0378426621002077>
- Petrella, L. & Raponi, V. (2019). Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. *Journal of Multivariate Analysis*, 173(C), 70–84. Available online at the address <https://ideas.repec.org/a/eee/jmvana/v173y2019icp70-84.html>
- Schelldorfer, J. & Wüthrich, M.V. (2019). Nesting classical actuarial models into neural networks. Available at SSRN. <https://ssrn.com/abstract=3320525> or <http://dx.doi.org/10.2139/ssrn.3320525>
- Spedicato, G.A., Dutang, C. & Petrini, L. (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMs. *Variance*, 12(1), 69–89. Available online at the address <https://hal.archives-ouvertes.fr/hal-01942038>
- Taylor, J.W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19, 299–311.
- Taylor, J.W. (2007). Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of Financial Econometrics*, 6(3), 382–406. Available online at the address <https://doi.org/10.1093/jjfnec/nbn007>
- Taylor, J.W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2), 428–441. Available online at the address <https://www.sciencedirect.com/science/article/pii/S0169207019301918>
- White, H., Kim, T.-H. & Manganelli, S. (2015). Var for var: measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics*, 187(1), 169–188. Available online at the address <http://www.sciencedirect.com/science/article/pii/S0304407615000287>
- Wüthrich, M.V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6), 465–480. Available online at the address <https://doi.org/10.1080/03461238.2018.1428681>
- Wüthrich, M.V. (2020). Bias regularization in neural network models for general insurance pricing. *European Actuarial Journal*, 10(1), 179–202.
- Wüthrich, M.V. & Merz, M. (2019). Editorial: Yes, we can! *ASTIN Bulletin*, 49(1), 1–3.
- Xu, Q., Deng, K., Jiang, C., Sun, F. & Huang, X. (2017). Composite quantile regression neural network with applications. *Expert Systems with Applications*, 76, 129–139. Available online at the address <http://www.sciencedirect.com/science/article/pii/S0957417417300726>

Cite this article: Laporta AG, Levantesi S and Petrella L (2023). Neural networks for quantile claim amount estimation: a quantile regression approach, *Annals of Actuarial Science*, 1–21. <https://doi.org/10.1017/S1748499523000106>