Master in Actuarial and Financial Sciences

MASTER THESIS

Title: Social networks Big Data: Personality traits as an explanatory variable in GLM models for insurance claim counts

Author: Andrés Santiago Durán Proaño

Promoter/s: Catalina Bolancé Losilla

Course: Actuarial and Finance Sciences

UNIVERSITAT DE BARCELONA

Facultat d' Economia i Empresa

Faculty of Economics and Business

Universitat de Barcelona

Master thesis

Master in Actuarial and Financial Sciences

Social networks Big Data: Personality traits as an explanatory variable in GLM models for insurance claim counts

Author: Andrés Santiago Durán Proaño

Promoter/s: Catalina Bolancé Losilla

The contents of this document are the sole responsibility of the author, who declares that he has not committed plagiarism and that all references to other authors have been expressed in the text.

To my parents, for the incommensurable gift on life;
to my kids, for being my source of motivation;
to my wife, for her love and support;
and to Ecuadorians how made this scholarship possible.

ABSTRACT

In an industry such as the insurer, highly atomized and competitive, where price comparison engines allow customers to have greater control over information in decision-making; insurance companies are investing great part of their efforts to find new formulas that improve customer loyalty. In that sense, using Big Data generated from social networks such as Facebook, Twitter or YouTube, to know the policyholder's personality, can be used as a strategy that allow companies to compete through personalized service and more competitive premiums. In this study, I analyze the framework to introduce personality as an explanatory variable in Generalized Linear Models for claims count and try to found out any empirical evidence of the relation between personality traits and insurance claims.

KEYWORDS: Big Data, social networks, Big Five personality traits, computational social science, insurance claim counts.

INDEX

FIGURES INDEX

TABLES INDEX

1. INTRODUCTION

Over the past decades, the auto insurance industry has had little changes in the way premiums are calculated. Traditionally, the historical claims data, specific information of the good to insure and policyholder characteristics have been the variables used as the main information to model a set of risks and determine the insurance premium. Several reasons can explain this lethargy of the companies, the main one is that the business model has worked all over that period: the technical methodology is proven and accepted by the regulation institutions.

This perspective is bound to evolve, insofar as other factors (variables) that have not been considered, are incorporated to the analysis. And this seems to be possible due to new and massive information that was not available time ago. Along with technology evolution, digital information has also grown and exponentially: Big Data[1]. Evidently, to the extent that technology has developed, industry has been incorporating it into its processes; but now, it is about implementing these new data.

And, in fact, it is being done. Some insurance companies have begun to take advantage of the Internet of Things (IoT) to monitor drivers' skills and habits. With telematic devices plug in the vehicle, insurers receive valuable information and analytics about policyholders' specific behaviors when driving. So, instead of paying for collision coverage even when the vehicle is parked in the garage, policyholders would only pay for it when the car's being driven.

But telematic devises are not the only source of new information. Social networks' information can be used to accurately predict a range of highly sensitive personal attributes; among them, personality traits (Michal Kosinski et al., 2013). This data could allow insurers to catalogue the risk based on information of the policyholder and offer premiums focus on individuals, not only of on risk groups. Big Five Personality Traits' model, summarize in only five representative traits human behavior and Computational Social Scientists have managed to make it possible to extract these personal patterns from social networks.

The application of these variables is immense in the industry, because all activities subject to risks are performed by a specific person. Let us imagine a bus transportation company, is it the same risk level when *one* or *another* bus driver drives a certain bus, even if it is on the same route? The inevitable answer is no; however, the transport company must pay an equal premium for each bus. Or home insurance segment, would be riskier a policyholder with lack of consciousness, rather than a very conscious one?

It is at least interesting to think about how these variables can customize the insurance premium. But at the same time, doubts arise about the fulfillment of the principle of solidarity in the insurance industry or on the legality and ownership of the information of the social networks. These reflections should be considered when thinking on the possibility of introducing personality traits or other new variables in insurance pricing *a priori* and *a posteriori*.

[1] Big Data is the term that describes a rich and complicated set of characteristics, practices and techniques, ethical issues and outcomes all associated with data (Japec et al., 2015). It is associate with large sets of information (personal, industrial, market, etc.) that are constantly increasing thanks to online connections.

As telematics devises analytics, personality traits information from social networks, are called to be the future of the industry. So, the hypothesis I would like to validate is that the probability of having a claim, also depends on personal constructs[2]. In this sense, the aim of this study are (1) to develop a theoretical framework to incorporate nontraditional variables to the actuarial analysis of claim counts, based on the idea that some personality traits are more prone to have claims; (2) try to find empirical evidence of the relation among personality traits and claim; and (3) get an understanding of the market's feeling when asked to provide access to their social networks account.

My objective is not to dismiss of the analysis traditional variables, systems or procedures; even better, the main idea is to add more information to the models, to create a more accurate risk profile *ergo* an appropriate premium for that level of risk. To develop this idea, I will first introduce the actuarial models in which, this new information coming from the social networks, can be used. Then, present studies related to computational social science and psychological assessment.

2. BACKGROUND

An actuary can be defined as "one who determines the current financial impact of future contingent events[3]". To determine this impact the actuary uses, among other technics, Generalized Linear Models (GLM), through which the mean response can be expressed as a function of linear combination of explanatory variables. In a simple way to express; when an actuary models the probability of having a claim, people most likely to have it, surely must pay a higher premium and vice versa.

In these models that estimate the probability of occurrence of a claim, considering quantitative variables that measure qualitative traits, it would be desirable for them to be mutually exclusive factors, so that they could represent different information (otherwise it would be modeling with variables that describe the same). Somehow, personality traits can be considered mutually exclusive quantitative variables. "Personality is also considered as an important piece of knowledge useful to build effective persuasive systems: people, in fact, may react differently to persuasive stimuli according to their personality" (Lepri et al., 2016).

2.1. GLM Models for claim counts

A non-life insurance contract, among an insurance company and a policyholder, is an agreement where the second one transfers an economic risk to the first, in exchange for a premium. The premium (or appropriately called *pure premium*) is the product of the potential claim frequency and the potential claim severity. In this section I will provide the basics in GLM Models to understand the "claim frequency" from an actuarial point of view; to do so, I basically use the material prepared by Professor Ramon Alemany (Alemany, 2015).

Giving a *price* to a certain risk, from a very basic appreciation, means to understand it and to understand the variables that affect it (Parodi, 2015). In non-life insurance, the aim of *a priori tariff analysis* is to determine how one or more variables, like the number of

[2] In psychology, any hypothetical entity difficult to define within a scientific theory. A construct is something that is known to exist, but whose definition is difficult or controversial. They are constructs intelligence, personality and creativity, for example.
[3] Attributed to Frederick W. Kilbourne.

claims, or the costs of claims, vary with respect to a set of *risk factors*. In other words, how a dependent variable *Y* varies with respect to an explanatory variable *X* in a multiple linear regression.

Nevertheless, in these cases the classic Linear Regression Model (LRM) is not the most appropriate, as for *pricing* purposes. There are two main reason, first LRM assume normality and homoscedasticity in the perturbation term, while other variable like the number of claims follows a discrete distribution, the cost of the claims is a non-negative variable and asymmetric to the right, and the renewal of a policy is a qualitative dichotomic variable, among others. Second, the mean of the LRM is a linear function of the explanatory variables while in other contexts, another functional relationship is needed.

Generalized Linear Model, GLM, are an extension of this LRM in two ways (Ohlsson & Johansson, 2010):

- *Probability distribution*: Instead of assuming the normal distribution of the perturbation term, GLMs work with a general class of distributions (the exponential family) which contains a set of discrete and continuous distributions in particular the Normal, Poisson and Gamma distributions.
- *The model for the mean*. In the LRM the mean is a linear function of the explanatory variables. In GLMs some monotonous transformation of the mean is a linear function of the explanatory variables.

GLMs components

a. Random component: Given y_i , $i = 1, \dots, n,$ independent random variables with density of the exponential family, that is:

$$f(y_i, \theta_i, \emptyset) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\emptyset/\omega_i} + c(y, \emptyset)\right\}.$$

In its canonical form, where:

- $\emptyset$ is a dispersion parameter.
- ω_i is a weighting.
- θ_i is the canonical parameter.
- The functions b(·) and c (·) are known.

b. Systemic component: the linear predictor is:

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik,} \ \ \forall \mathrm{i} = 1, \dots, n,$$

where β is the vector *k x 1* of parameters to estimate and x_{ij} the explanatory variables.

c. Link component: the link function relates the expected value of the dependent variable with the linear predictor (Frees, 2010):

$$\eta_i = h^{-1}(\mu_i) = x_i'\beta, \ \forall \mathrm{i} = 1, \dots, n.$$

A canonical link exists if:

$$\theta_i = \eta_i = x_i'\beta \quad \forall_i = 1, \dots, n.$$

Moments

For these distributions of the exponential family, the *first raw moment* (mean) and the *second central moment* (variance) are:

$$E[y_i] = b'(\theta_i) = \frac{db(\theta_i)}{d\theta} = \mu_i$$

and

$$V[y_i] = \frac{\phi}{\omega_i} b''(\theta_i) = \frac{\phi}{\omega_i}\frac{d^2 b(\theta_i)}{d{\theta_i}^2} = \frac{\phi}{\omega_i} V(\mu).$$

Distributions and links

Table 1 shows a set of probability distributions and their link function.

Distribution	Link	$\boldsymbol{\eta_i = g(\mu_i)}$	$\boldsymbol{\mu_i = g^{-1}(\eta_i)}$	Range for $\boldsymbol{Y_i}$	$\boldsymbol{V(Y_i \vert \eta_i)}$
Normal	Identity	μ_i	η_i	$(-\infty, +\infty)$	$\emptyset$
Binomial	Logit	$\ln \frac{\mu_i}{1-\mu_i}$	$\frac{exp(\eta_i)}{1 + exp(\eta_i)}$	$\frac{0,1,\dots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	$\ln \mu_i$	$exp(\eta_i)$	0, 1, 2, . . .	μ_i
Gamma	Inverse	μ_i^{-1}	η_i^{-1}	$(0, \infty)$	$\emptyset \mu_i^2$

Table 1. Distributions and links of the Exponential Family.
Source: (Alemany, 2015)

Maximum Likelihood Estimation

The logarithm of the likelihood function for an individual observation y_i considering the density of the exponential family, would be:

$$\ln \mathcal{L}(y_i; \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi/\omega_i),$$

and for *n* independent observations, would be:

$$\ln \mathcal{L}(y_1, \dots, y_n; \theta, \phi) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi/\omega i)\right],$$

Besides, if the link function $g(\mu_i) = \eta_i = x_i'\beta$ is canonical, it will happen $g(\mu_i) = \eta_i = x_i'\beta$ and the logarithm of the likelihood function would be:

$$\ln \mathcal{L}(\beta, \phi, |y_1, \dots, y_n) = \sum_{i=1}^{n} \left[\frac{y_i x_i' - b(x_i'\beta)}{\phi/\omega_i} + c(y_i, \frac{\phi}{\omega_i}) \right].$$

The derivate of $ln\, \mathcal{L}(\beta, \phi)$ of with respect to β is:

$$\frac{\partial}{\partial \beta} ln\, \mathcal{L}(\beta, \phi) = \frac{1}{\emptyset} \sum_{i=1}^{n} \big(y_i - b'(x_i'\beta)\big) w_i x_i.$$

Considering that $u_i = b'(\theta_i) = b'(x_i'\beta)$, the Maximum Likelihood Estimators (MLE) of the parameters $\hat{\beta}_{MLE}$ can be obtained through the equations:

$$0 = \sum_{i=1}^{n} w_i (y_i - u_i) x_i.$$

The solution of these equations can be approximated by iterative procedures, such as the Newton-Raphson Algorithm or the Fishing Scoring Method.

Models selection

There are different criteria to measure the goodness of fit considering the number of parameters included in the model. The best known are the Akaike's Information Criterion, AIC, and the Bayesian Information Criterion, BIC, i.e.:

$$AIC = -2 \ln \mathcal{L} + 2p$$

and

$$BIC = -2\ln \mathcal{L} + \mathrm{p} \ln n.$$

When we are estimating models with the same data set, the selected model should be the one with the lowest AIC or BIC. When we have a large number of observations, AIC is preferable to BIC.

Residuals

In GLM models, Pearson Residuals, r_i, are used, instead of the traditional ones. These Pearson residuals are expressed as:

$$r_i = \frac{\hat{\phi}^{1/2} (y_i - \hat{\mu}_i)}{\sqrt{\widehat{Var}(Y_i)}},$$

these residuals allow the construction of the Pearson Statistic, which under the null hypothesis of a good fit is distributed according to a Chi-square. The Pearson statistic is:

$$\sum_{i=1}^{n} r_i^2 \sim X_{n-k}^2.$$

Other option with GLMs is to estimate the Deviance residuals:

$$d_{ri} = signo\ (y_i - \hat{\mu}_i)\sqrt{2\left(\ln \mathcal{L}(\hat{\beta}_R)_i - \ln \mathcal{L}_i(\hat{\beta})_i\right)},$$

where $\ln \mathcal{L}(\hat{\beta}_R)_i$ is the logarithm of the likelihood of the model, estimated with r lineal restrictions over the parameters.

2.1.1. Logit model

The Logistic Regression Model, LOGIT, is applied when the dependent variable is a discrete and dichotomic one, for example: the occurrence of a claim, the renewal of a policy, a fraudulent claim, etc., and the explanatory variables are either continuous and / or discrete. The LRM is not suitable for this type of situation because the basic hypothesis of normality and homoscedasticity of the term of random perturbation are not fulfilled. Also, because the LRM predictions are beyond the range [0,1], which are meaningless in terms of probability; the relation between the explanatory variables and the probability of occurrence of the event is not linear. The adjustment of the model and the interpretation of the parameters is not the same.

Let *Y* be a random variable (r.v.) that takes values 1 if an event occurs and 0 otherwise, *X* is a matrix of variables explaining the occurrence of the event and β is a vector of parameters (Guillén, 2014). Then, with a set of *n* independent individual observations, *Y* follows a Binomial distribution, $\mathcal{B}$(n, π), which is of the exponential family, the canonical link function of the Logit model is:

$$\theta = \ln \frac{\pi}{1-\pi} = \eta = x_i'\beta,$$

then,

$$\pi = P(Y=1) = \frac{e^{\theta}}{1+e^{\theta}} = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \Lambda(x_i'\beta)$$

and

$$1-\pi = P(Y=0) = 1 - \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \frac{1}{1+e^{x_i'\beta}} = 1 - \Lambda(x_i'\beta).$$

The vector of estimated parameters $\hat{\beta}$ is obtained by the criterion of maximum likelihood. Let us suppose observations Y_i, independent and equally distributed, like a Bernoulli. The likelihood function is the joint probability:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | X) = \prod_{y_i=0} [1 - \wedge (x_i'\beta)] \prod_{y_i=1} \wedge (x_i'\beta)$$

or

$$L(\beta) = \prod_{i=1}^{n} [\wedge\, (x_i'\beta)]^{y_i}\, [1 - \wedge\, (x_i'\beta)]^{1-y_i}.$$

Which in logarithms is:

$$\ln\{L(\beta)\} = \sum_{i=1}^{n} [y_i \ln\{\wedge\, (x_i'\beta)\} + (1 - y_i) \ln\{1 - \wedge\, (x_i'\beta)\}].$$

As said before, the maximization of the logarithm of the likelihood function should be performed using numerical optimization methods such as Newton's Method. This the maximum-likelihood estimators will be: unbiased, consistent, asymptotically efficient and distributed asymptotically according to a normal.

2.1.2. Poisson Model

A r.v. *Y* can take on a set of values from the sample space Ω. When Ω is a finite set of real numbers specifically non-negative integers {0,1,2…} associated with the r.v., *f(y)* is a probability function indicating for each variable *y* in Ω, the probability that the *Y* takes value *y* (de Jong & Heller, 2013). In this case we are describing a discrete r.v.

Count data models are those in which the dependent variable is a count of the number of times that an event occurs. These are discrete models, so the variable to explain takes on discrete values 0,1,2… These models are widely used in the actuarial profession because the claims number meets these properties. The Poisson Model is a particular GLM for claims count. However, for the Poisson distribution the mean is equal to the variance, which is an important restriction that will open the door to more general models, such as the Negative Binomial Model that will be analyzed ahead.

The probability function of the Poisson distribution is:

$$P(Y = j) = \frac{u^j}{j!} e^{-u}, \quad j = 0,1,2 \dots,$$

where *u* is the mean number of times that the event occurs in a time interval. Furthermore, the probability generating function is:

$$P(z) = e^{u(z-1)}, \quad u > 0.$$

The mean and variance can be computed from the probability generating function, and their value are:

$$E(Y) = u$$

and

$$V(Y) = u.$$

The idea is that the mean could vary according to the values of some explanatory variables representing various risk factors, like this,

$$E(y_i) = u_i = e^{x_i'\beta}.$$

Even more, the could vary according to the risk exposure, *t_i*, which the proportional part of the year the policyholder has been covered,

$$E(y_i) = u_i = t_i e^{x_i'\beta}.$$

So, let

$$Y_i \sim Poi\left(t_i e^{x_i'\beta}\right), \quad i = 1,2\dots,n \quad i.i.d.,$$

then

$$P(Y_i = y_i) = \frac{(t_i e^{x_i'\beta})^{y_i}}{y_i!} e^{-t_i x_i'\beta} ,$$

which can be expressed like

$$P(Y_i = y_i) = e^{y_i \ln\left(t_i e^{x_i'\beta}\right) - \, t_i e^{x_i'\beta} \, - \, ln\, y_i!} .$$

So, it would be of the exponential family, with:

$$\theta_i = \ln\left(t_i e^{x_i'\beta}\right),$$

$$b(\theta_i) = -t_i e^{x_i'\beta} = e^{\theta_i} ,$$

$$c(y;\phi) = -\ln(y_i) ,$$

$$\phi = 1$$

and link function equal to:

$$g(u_i) = \ln(u_i) = \ln t_i + x_i'\beta$$

The logarithm of the likelihood function is,

$$\ln \mathcal{L}(\beta|y_1,\dots,y_n) = \sum_{i=1}^{n} \left[- t_i e^{x_i'\beta} + y_i \left(\ln t_i + e^{x_i'\beta}\right) - ln\, y_i!\right].$$

That when the first derivatives are equal to zero with respect to the parameters,

$$\frac{\partial}{\partial\beta} \ln \mathcal{L}\big((\beta|y_1,\dots,y_n)\big) = \sum_{i=1}^{n} \left(y_i - t_i e^{x_i'\hat{\beta}}\right) x_i = 0 ,$$

where $\hat{\beta}$ is the maximum likelihood estimator.

2.1.3. Negative Binomial Model

Let Y be a count data r.v., Poisson with mean λ. If λ is not constant, then λ is a r.v., and we assume it is a Gamma distribution with parameters a and b, i.e.:

$$\lambda \sim Gamma\ (a, b),$$

with density function:

$$g(\lambda) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$$

then, the Negative Binomial is obtained as a mixture of the Poisson and the Gamma distribution. The final result is:

$$f(y|a, b) \ = \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^y,$$

with mean

$$E(y) = \frac{a}{b}$$

and variance

$$V(y) = \frac{a}{b}\left(1 + \frac{1}{b}\right) = \bar{\lambda}\left(1 + \frac{\bar{\lambda}}{a}\right).$$

As for the Poisson model, the idea is that the expected number of claims, per time unit, could change according to some risk factors. However, a heterogeneity factor is introduced, because explanatory variables do not capture it among individuals.

$$E(y_i) = \mu_i = e^{x_i'\beta + \epsilon_i}.$$

So, the probability would be calculated as:

$$f(y|a) = \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)} \left(\frac{e^{x_i'\beta}}{a}\right)^y \left(1 + \frac{e^{x_i'\beta}}{a}\right)^{-(y+a)}.$$

The logarithm of the likelihood function is,

$$\ln \mathcal{L}(\beta, a|y_1, \dots, y_n) = \sum_{i=1}^{n} \left\{ \left(\sum_{j=0}^{y-1} \ln(j+a) \right) - \ln y_i! - (y_i + a) \ln\left(1 + \frac{e^{x_i'\beta}}{a}\right) + y_i \ln\left(\frac{e^{x_i'\beta}}{a}\right) \right\}.$$

$\hat{\beta}$ and $\hat{a}$ are estimated solving:

$$\frac{\partial}{\partial \hat{\beta}} \ln \mathcal{L}(\hat{\beta}, \hat{a}) = \sum_{i=1}^{n} \frac{y_i - e^{x_i'\hat{\beta}}}{1 + \frac{e^{x_i'\hat{\beta}}}{\hat{a}}} x_i = 0$$

and

$$\frac{\partial}{\partial \hat{a}} \ln \mathcal{L}(\hat{\beta}\,, \hat{a}) = \sum_{i=1}^{n} \left\{ \hat{a}^2 \left[\ln\left(1 + \frac{e^{x_i'\beta}}{\hat{a}}\right) - \sum_{j=0}^{y-1} \frac{1}{j+\hat{a}} \right] + \frac{\hat{a}(y_i - e^{x_i'\hat{\beta}})}{1 + \frac{e^{x_i'\beta}}{\hat{a}}} \right\} = 0\,,$$

where $\hat{\beta}$ and $\hat{a}$ are the estimators of maximum likelihood.

2.1.4. Zero Inflated Models

Frequently, when analyzing claims insurance databases, there is an excess of zeros in the number of claims (which is desirable). A model can be zero inflated, if it can be written as a mixture with π_i probability, of a Dirac distribution in zero and a count model like Poisson or Negative Binomial (Charpentier, 2015), i.e.:

$$P_r(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)g_i(0) & j = 0 \\ (1-\pi_i)g_i(j) & j = 1,2,\ldots \end{cases}.$$

The ZIP model is explained as follows:

$$g(y_i|\lambda_i) = \frac{e^{-\lambda i}\lambda_i^{yi}}{y_i!}\ ,$$

$$\lambda_i = e^{x_i'\beta},$$

$$\pi_i = \frac{e^{z_i'\gamma}}{1 + e^{z_i'\gamma}},$$

$$P_r(y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda i} & y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\lambda i}\ \lambda_i^{yi}}{y_i!} & y_i = 1,2,\ldots \end{cases},$$

with mean

$$E(y_i) = \mu_i = (1-\pi_i)\lambda_i$$

and variance

$$Var(y_i) = \mu_i + \frac{\pi_i}{1-\pi_i}\mu_i^2 = (1-\pi_i)\lambda_i + \pi_i(1-\pi_i)\lambda_i^2.$$

On the other hand, the ZINB model is specified as:

$$g(y_i|\lambda_i) = \frac{\Gamma(a+y)}{\Gamma(y+1)\Gamma(a)}\left(\frac{a}{a+\lambda_i}\right)^a \left(\frac{\lambda_i}{a+\lambda_i}\right)^y,$$

$$\lambda_i = e^{x_i'\beta},$$

$$\pi_i = \frac{e^{z_i'\lambda}}{1 + e^{z_i'\lambda}},$$

$$P_r(y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{a}{a+\lambda_i}\right)^a & y_i = 0 \\ (1-\pi_i)\ \frac{\Gamma(a+y)}{\Gamma(y+1)\Gamma(a)}\left(\frac{a}{a+\lambda_i}\right)^a\left(\frac{\lambda_i}{a+\lambda_i}\right)^y & y_i = 1,2,\ldots \end{cases},$$

with mean

$$E(y_i) = \mu_i = (1-\pi_i)\lambda_i$$

and variance

$$Var(y_i) = \mu_i + \left(\frac{\pi_i}{1-\pi_i} + \frac{1/a}{1-\pi_i}\right)\mu_i^2 = (1-\pi_i)\lambda_i + (\pi_i + 1/a)(1-\pi_i)\lambda_i^2.$$

2.2. Trait theory

Trait theory is the branch of psychology that studies human personality. It is also called dispositional theory. A trait can be defined as a habitual pattern of behavior, thought, and emotion (Matthews et al., 2009). The first step of psychologist of personality consist of developing a science of traits by measurement and classification of traits. Traits are relatively stable over time and differ across individuals, so a basic technic to measure personality is to ask a person to rate how well traits adjectives, like quiet, sincere, mean, liar, etc., apply to himself or herself.

2.2.1. The lexical hypothesis

Each person has its own characteristics; and this diversity difficult researchers to understand personality (Ashton & Lee, 2005). However, one way to obtain a set of variables that is representative of an individual, is by considering the lexical hypothesis in personality analysis. This concept has been around since the 1800s and, basically, this hypothesis states that the most important personality characteristics of a person will eventually become a part of his or her language and more likely, encoded as a single word (John et al., 1988). It means that everything we write or talk leaves a trace of our personality; with language as a resource and a sample, a full spectrum and taxonomy of personality traits can be assembled.

This hypothesis lead to the development of the Big Five Personality Traits (Goldberg, 1993), a prominent widely accepted multi factorial model that embraces in five traits the human behavior. Is a model based on common language descriptors of personality summarizes characteristic behaviors that distinguish a person throughout different contexts of their daily life.

2.2.2. The Big Five personality traits

The Big Five personality traits subject of this study are:

- Openness
- Conscientiousness
- Extroversion
- Agreeableness
- Neuroticism

These traits can be scored in several ways, depending on the methodology; but are typically expressed in a range that goes from zero to one (0,1). So, a high score on Openness, for example, would be a value of 0,8 to 1; an average score would be 0,4 to 0,6 and a low score would be 0 to 0,2.

Johnson (2016), based on an extensive reading of the scientific literature on personality measurement, summarized this five personality traits, as it is shown in Table 2:

Openness (OPE)	Is defined as "openness to experience". It distinguishes imaginative and creative people, from down to earth or conventional people. People who are Open to Experience are intellectually curious, appreciate art and beauty. Are closed people, because they are more aware of their feelings. Characteristics of these people include individualistic thinking or acting and nonconforming ways. Intellect is probably best regarded as one aspect of openness to experience.	
	Facets	Imagination. Artistic Interests. Emotionality. Adventurousness. Intellect. Liberalism.
	(-) Low score	Down to earth, practical, conservative, think in simple terms.
	Average score	These people enjoy tradition but do not close their mind to try new experiences, even though their thinking is neither simple nor complex. Usually well-educated people but not intellectuals.
	(+) High score	Intellectuals typically score high on Openness to Experience. They enjoy variety and change. Usually are curious, imaginative and creative people.
Conscientiousness (CON)	It concerns to the way in which people control, regulate and direct their impulses. To be impulsive is not necessarily bad, sometimes restrictions require a snap decision and acting with impulse can be effective. Impulsive individuals can be seen by others as colorful or fun to be with. But also, acting on impulse can lead to troubles. Some impulses are antisocial and uncontrolled antisocial acts harm other members of society and also can result in retribution toward the perpetrator of such impulsive acts. Another problem with impulsive acts is that they often produce immediate rewards but undesirable, long term consequences. Impulsive behavior reduces a person's efficiency, because it disallows contemplating alternative courses of action, some of which would have been wiser than the impulsive choice. On the other hand, a cautiousness person describes the disposition to think through possibilities before acting.	
	Facets	Self-Efficacy. Orderliness. Dutifulness. Achievement-Striving. Self-Discipline. Cautiousness. Responsibility.
	(-) Low score	This people like to live for the moment and do what feels good now. Normally, their work tends to be careless and disorganized.

	Average score	This means he/she is a reasonably reliable, organized and self-controlled person.
	(+) High score	Is a person seen by others as reliable and hard working.
Extroversion (EXT)	Its main characteristic is the engagement with the external world. Extraverts appreciate company, talking to people, tend to be full of energy and enthusiastic. Usually experience positive emotions and are more likely to say "Yes!" to opportunities. On the other hand, introverts lack this energy and activity levels; they are quiet, thoughtful and disengaged from the social world. It does not mean they are depressed and should not be necessarily associated with shyness.	
	Facets	Friendliness. Gregariousness. Assertiveness. Activity Level. Excitement-Seeking. Cheerfulness.
	(-) Low score	Indicates an introverted, reserved and quiet person. Someone who enjoys solitude and solitary activities. He/She has a few close friends.
	Average score	Indicates neither a subdued loner nor a jovial chatterbox person, that enjoys time with others but also time alone.
	(+) High score	Indicates a sociable, outgoing and lively individual, that prefers to be around people much of the time.
Agreeableness (AGR)	Agreeable people value getting along with others; they are respectful, tolerant, considerate, generous and helpful. They also have an optimistic view of human nature, so believe people are basically honest or decent or trustworthy. Disagreeable people place self-interest above getting along with others. They do not care others' well-being and are skeptic of others, so they tend to be unfriendly and uncooperative.	
	Facets	Trust. Morality. Altruism. Cooperation. Modesty. Sympathy.
	(-) Low score	Indicates less concern with others' needs. This people are tough, critical, and uncompromising.
	Average score	Indicates some concern with others' needs; but not to sacrificing yourself for others.
	(+) High score	Indicates that this person is interested in other ones' needs and wellbeing. This people are sympathetic and cooperative.

Neuroticism (NEU)	To explain neuroticism, I will refer to people with high scores of neuroticism. Measures a reaction to complicated situations. It refers to the tendency to experience negative feelings like anxiety, anger or depression, these people are emotionally reactive to events that would not affect most people. Their reactions tend to be more intense than normal because they interpret current situations as threatening. This people are often in a bad mood and cannot think clearly, make decisions or deal with stress.	
	Facets	Anxiety. Anger. Depression. Self-Consciousness. Immoderation. Vulnerability.
	(-) Low score	It refers to a calm and composed persona that do not react with intense emotions.
	Average score	Indicates that the level of emotional reactivity is typical of the general population. Stressful and frustrating situations upset this people, but can get over these feelings.
	(+) High score	These persons are easily upset with the normal demands of living; sensitive and emotional.

Table 2. Big Five personality traits model.
Source: Own elaboration based on (Johnson, 2016).

3. RELATED WORK

The objective of traditional/empirical social science has been make inferences about a population from available sources of data (Foster et al., 2017). Today, Computational Social Science can achieve this goal but with more precise sources, tools and technics such as scraping, machine learning, relational database management systems, etc.

Michal Kosinski, David Stillwell, and Thore Graepel carried out one of the firsts and most important studies about predicting personal traits and attributes from digital records (Michal Kosinski et al., 2013). They show that a wide variety of people's personal attributes can be inferred using their Facebook Likes. Users and their Likes were represented as a sparse user–Like matrix and its dimensionality was reduced using singular value decomposition. Numeric variables (age, intelligence) were predicted using a linear regression model; and dichotomous variables (gender, sexual orientation) were predicted using a logistic regression.

Through a machine learning process, they could determine men's sexual orientation with an accuracy of 88%, ethnicity 95%, religion 82% and political views 85%. They also predicted numeric attributes and individualities like personality traits, intelligence, happiness, use of addictive substances, etc. For that, these researchers used a significant sample of volunteers from the *myPersonality* project, which was a Facebook application that allowed users to take real psychometric tests, and record their psychological and Facebook profiles. Similar studies have been done with other social networks like Twitter (Preot et al., 2017) or Youtube (Biel & Gatica-perez, 2013).

According to (Michal Kosinski, 2017b) we are leaving thousands of *Digital Footprints*[4] behind while using Internet and other digital services and products, like a smartphone, which is constantly recording our location, conversations, pictures, emails and, now, even our heart beat. We are generating an incredible amount of information through digital devices, sometimes, even without knowing it. Our credit card number and name, for instance, are associated with our consuming preferences, so the credit card companies and the groceries stores are exchanging this information with the objective of understanding us, the clients.

Psychologists have determine our behavior is not random (Kennedy, 2012). We do not buy a pair of jeans, or watch certain shows, or listen to music randomly. When we buy something, normally, we are revealing we belong to a determinate group an advertiser fixed. Or when we watch a show is maybe because a friend of us recommended to do so; and again, that is not random, because we make friends with people with whom we share interests, education, political views, etc. Now, if everything we do is not random, there is a connection among every aspect of our behavior. This connection is to slight for human beings to perceive that an algorithm is needed. The algorithm can look at millions of people and find little correlations, to combine these thousands of pieces of information to give an accurate profile (Michal Kosinski, 2017a).

This revolution of knowing the individual in an unprecedent way, has lead in recent years to the development of a diversity of companies selling algorithms information, among witch personality traits are pursued. Some of the webpages that offer this service include: *Personality Insights* or *Tone Analyzer* powered by IBM Watson. And also academically, there has been a boom of publications describing techniques to build predictive models that can infer the scores users would receive on the Big Five Personality Traits, using their social networks data; for example (Michal Kosinski, 2013), (Farnadi et al., 2016) or (Whitty et al., 2014). The basic idea is to use the dimensions or clusters extracted to build predictive models in a cross-validated way, using R code and a data set from *myPersonality* project. This database currently has more than six million test results, together with more than four million individual Facebook profiles; and a wide variety of data.

Big Data from Social Networks offers the opportunity to study everyday behavior at a scale never before possible (Kulkarni et al., 2017). This technology is being used in several fields; Human Resources Departments, for example, have used psychological test for decades to hire the correct person, but now they are able to identify employees' talent to develop and engage their potential to increase organizational effectiveness (Chamorro-Premuzic et al., n.d.). Or in Health industry, also, where Facebook information can be used to construct a reasonable index of population-scale schizophrenia consciousness (Saha et al., 2017), or monitoring Twitter to find at-risk for depression users (Jamil, 2017).

Having this knowledge of the people, could be also useful for further applications in the insurance industry like developing specific products, services or marketing strategies; for example, insurance advertisements might emphasize security when facing a neurotic user; but if he/she is emotionally stable, then potential threats should be emphasized (Lambiotte & Kosinski, 2014).

[4] For one individual, it is a unique and traceable set of digital activities, actions and communications on the Internet or on a computer or other digital device. An example of Digital Footprint can be our online browsing habits.

4. DATA ANALYSIS

4.1. Insurance and Facebook databases

As social networks information is private (in a certain way); first it is very important to ask the policyholders consent to use their information and data. Then a database with the scores of the Big Five Personality Traits from the policyholders, needs to be created; for example, using the techniques described in (Michal Kosinski, 2013).

The development of an algorithm to extract/predict personality traits would not be part of this study, because there is a vast offer of this services, as seem before. To be able to proof the hypothesis, it would be needed a traditional claims database, which should include the number of claims, from an insurance company; along with a database containing social networks information, from the same policyholders included in the first database. As it is not available, I will use two databases and combine them to create a database like the one is needed to obtain empirical evidence. These databases are described, respectively, in Table 3 and 4.

First database (A) (claims): policyholder claims information

Variable	Detail	Type
client_id	policyholder identity number	Numeric
nclaims_md	number of claims for material damage	Numeric
nclaims_bi	number of claims for body injuries	Numeric
nclaims_auto	total number of claims	Numeric
exposi_auto	time of the policyholder in the company	Numeric
client_sex	policyholder gender as categorical variable	Text
client_gender	policyholder gender as numerical variable: 0 for male, 1 for female	Numeric
client_age	policyholder age	Numeric
client_age2	policyholder age squared	Numeric
client_genage	product: gender * age	Numeric
zone1	Barcelona and Madrid	Numeric
zone2	north zone of Spain	Numeric

Table 3. Claims dataset of 162.019 policyholders.
Source: Private insurance company.

Second database (B) (users): psychodemographic profiles of 110.728 Facebook users and their Facebook Likes.

Variable	Detail	Type
userid	anonimized user ID	Alphanumeric
gender	user gender as numerical variable: 0 for male, 1 for female	Numeric
age2	user age squared	Numeric
genage	product: gender * age	Numeric
age	user age	Numeric

ope	scores measuring openness	Numeric
con	scores measuring conscientiousness	Numeric
ext	scores measuring extroversion	Numeric
agr	scores measuring agreeableness	Numeric
neu	scores measuring neuroticism	Numeric

Table 4. Psychodemographic profiles of 110.728 Facebook users and their Facebook Likes.
Source: myPersonality project (M. Kosinski et al., 2015).

In both databases the variables *age squared* and *gender*age* were generated. The reason to incorporate the age squared is to generate a quadratic curve. Thereby, a (+) effect of variable *age* and a (–) effect of the variable *age squared*, means that as the individual gets older, the effect of age is modest. On the other hand, (+) effect of the variable *age* and a (+) effect of the variable *age squared* means that as the individual gets older the effect of age is stronger. And, when introducing the variable *gender*age*, the pursued objective was to capture the effect of being female, rather than male.

4.2. Modelling claims frequency with *nontraditional* variables

To model the claims frequency it was used the R software and its *Integrated Development Environment*, R-Studio; following the procedures and techniques explained by Charpentier (2015). As a previous step for modelling, preparing the datasets cleaning missing values, eliminating wrong ages or correcting misspellings, was first done. Besides, as far as I am concern, this is the first academic trial to model claims frequency with personality traits as explanatory variable; thus, there is no a unique dataset containing traditional variables and these other nontraditional variables, as explained in the last section.

Therefore, a two steps idea has been developed to make it possible to have an insurance company database along with personality traits for each policyholder, to be able to proof any relation among claims frequency and personality traits. The fists step consists in explaining each empirical personality trait (e.g.openness), in the database B, as a function of the traditional variables *age*, *gender*, *age squared* and *gender*age*, to obtain these variables coefficients. The second step implies the estimation of all the personality traits (e.g. $\overline{\text{openness}}$) with the same variables, *age*, *gender*, *age squared* and *gender*age*, but from the database A. Thereby it would be possible to use personality traits, even though they are not empirical, but estimated. Ahead the details.

To explain the personality traits of the database B as a function of the variables *age*, *gender*, *age squared* and *gender*age*, a LRM was used, as a mean to associate these variables. The formulas in R are:

```
users$ope ~ users$gender + users$age + users$age2 + users$genage
users$con ~ users$gender + users$age + users$age2 + users$genage
users$ext ~ users$gender + users$age + users$age2 + users$genage
users$agr ~ users$gender + users$age + users$age2 + users$genage
users$neu ~ users$gender + users$age + users$age2 + users$genage
```

This procedure allowed to obtain the coefficients of the traditional variables, explaining each personality trait. Most of the coefficients were significant at levels of 1%, 5% and 10%, although the adjusted R-squared was low in all the cases. Regression's results are summarized in Table 5.

	coef_ope	coef_con	coef_ext	coef_agr	coef_neu
intercept	-0,897631	-1,688531	0,006890	-0,086594	-0,014723
users$gender	0,374897	-0,010762	-0,064126	-0,187603	0,693029
users$age	0,047236	0,081632	-0,002615	-0,003354	-0,010451
users$age2	-0,000505	-0,000818	0,000033	0,000098	0,000101
users$genage	-0,012398	0,002208	0,003781	0,011784	-0,008900

Table 5. Traditional variables coefficients for personality traits.
Source: own elaboration.

Once the coefficients are obtained, the following step consists in estimating personality traits for the insurance company database (claims), using its analog traditional variables *age*, *gender*, *age squared* and *gender*age*. Personality traits for every policyholder were estimated as a matrix product of the variables and the previously calculated coefficients. For example, to Neuroticism it is obtained:

```
x_neu   <- cbind(claims$client_gender, claims$client_age,
               claims$client_age2, claims$client_genage)
x_neu   <- as.matrix(x_neu)
n       <- nrow(x_neu)
ones    <- matrix(1,n,1)
x_neu   <- cbind(ones,x_neu)
hat_neu <- matrix(0,n,1)
for(i in 1:n){
  hat_neu[i]<-(x_neu[i,]%*%coef_neu)
}
```

To make easier to compare scores of the estimated personality traits, it was decided to work with standardized values.

```
hat_opes <- (hat_ope - mean(hat_ope)) / sd(hat_ope)
hat_cons <- (hat_con - mean(hat_con)) / sd(hat_con)
hat_exts <- (hat_ext - mean(hat_ext)) / sd(hat_ext)
hat_agrs <- (hat_agr - mean(hat_agr)) / sd(hat_agr)
hat_neus <- (hat_neu - mean(hat_neu)) / sd(hat_neu)
```

In Figure 1 it is described the distribution of these variables.

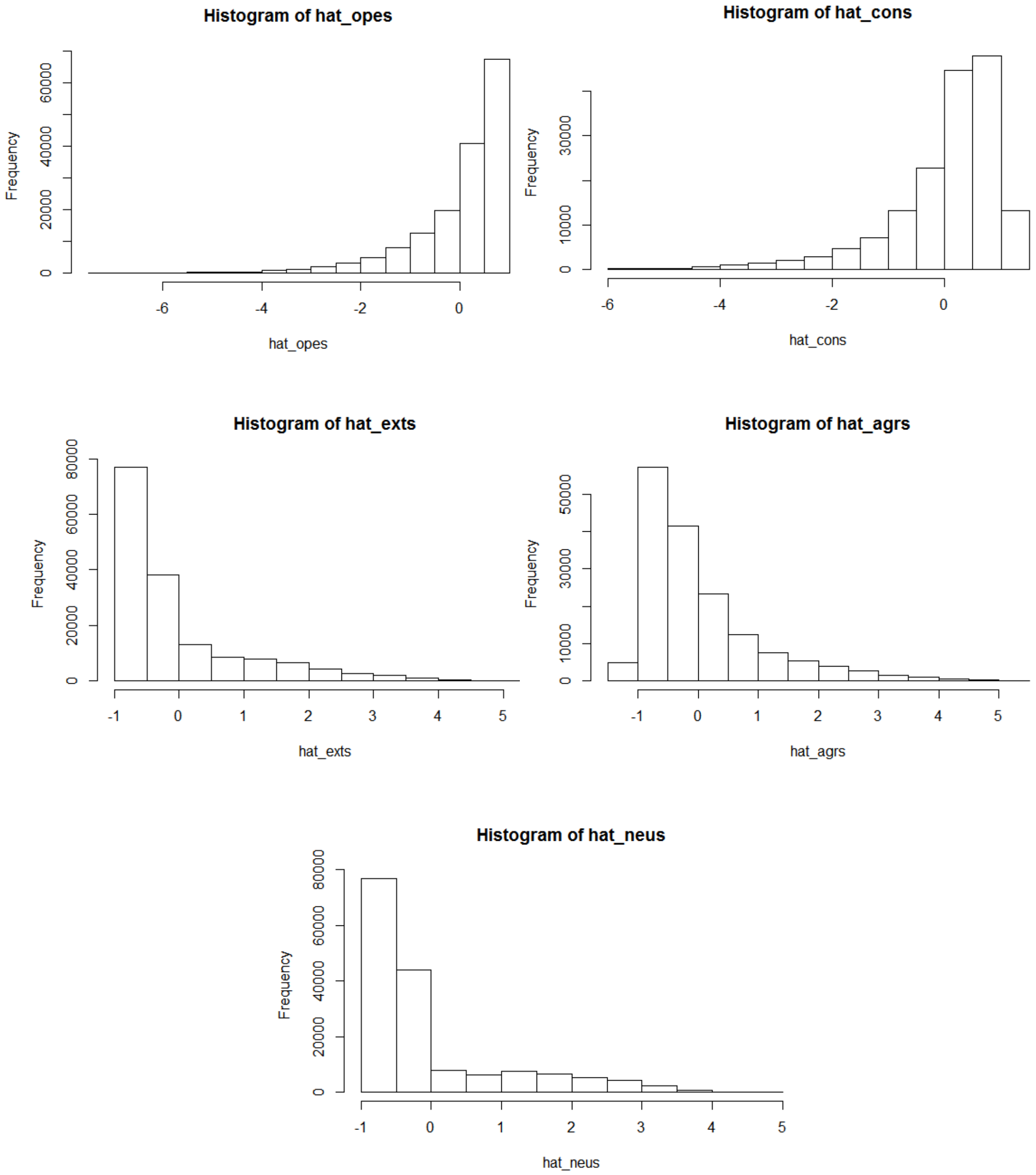


Figure 1. Standardized estimated personality traits.
Source: Own elaboration.

Now, the insurance company database B (claims), includes the claims frequency and the estimated personality traits for each policyholder; making it possible to study their relationship. To understand this relation of the number of claims as a dependent variable of the personality traits, it had been used GLM models for claims count: Poisson, Negative Binomial, Zero Inflated Poisson and Zero Inflated Negative Binomial.

But when doing the modeling, it was evident that existed multicollinearity among the estimated personality traits, it is a strong correlation between the explanatory variables of the model. So, it was decided to carry out a Principal Components Analysis (PCA), to work with synthetic variables that represent the estimated personality traits and reduce the dimensions of these dataset.

The PCA shows an evident concentration of the variance in the three first principal components, as it can be seen in Figure 2, together they represented the 99,71% of the information.

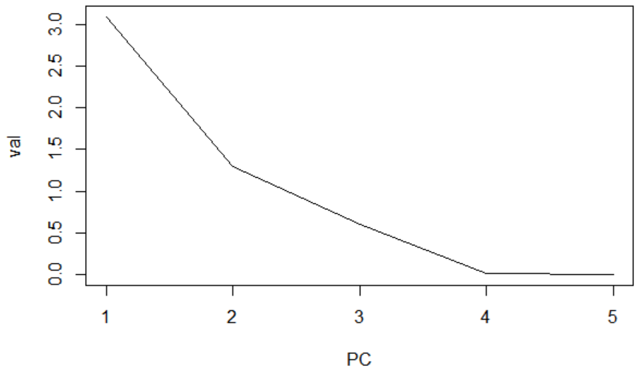
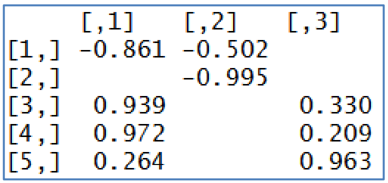


Figure 2. Principal Components values.
Source: Own elaboration.

To assig which principal component goes with each estimated personality trait, a *Varimax Orthogonal Rotation* of the Principal Components is applied, to simplify the structure of the columns of the eigenvectors matrix; it means that it increases the higher values and diminishes the lower ones. The loadings[5] obtained to select the principal component that represents the estimated personality traits, is presented in Figure 3:



```
      [,1]   [,2]   [,3]
[1,] -0.861 -0.502
[2,]        -0.995
[3,]  0.939         0.330
[4,]  0.972         0.209
[5,]  0.264         0.963
```

```
Columns represent the principal components and rows represent the estimated
personality traits in the following order: 1. Openness, 2. Conscientiousness,
3. Extroversion, 4. Agreeableness and 5. Neuroticism.
```

Figure 3. Loadings of the PCA and the Varimax Orthogonal Rotation.
Source: Own elaboration.

The relationship between the main components and the estimated personality traits is described in Table 6.

Principal Component	Describes
Component 1 (CPr 1)	- negatively openness (-0.861) - positively extraversion (0.939) - positively agreeableness (0.972)
Component 2 (CPr 2)	- negatively consciousness (-0.995)
Component 3 (CPr 3)	- positively neuroticism (0.963)

Table 6. Relationship between Principal Components and Estimates Personality Traits.
Source: Own elaboration.

[5] Loadings are the covariances/correlations between the original variables and the unit-scaled components in PCA or factor analysis.

Resuming the modeling, the explanatory variables will be the three Principal Components (representing the estimated personality traits), zone 1 and zone 2 and the response variable will be the number of claims.

Thereby, the models used to calculate the claims frequency with estimated personality traits as explanatory variables are Poisson, Negative Binomial, Zero Inflated Poisson and Zero Inflated Negative Binomial. Their respectively R formula is detailed above and the estimated coefficients (outputs) of the first three models can be seen in the appendix *A1*, *A2* and *A3*; Zero Inflated Negative Binomial is explained in the next section. The complete R code formulation can be found in appendix *A5*.

Poisson

```
glm(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                          + claims$zone1 + claims$zone2
                          + offset(log(claims$exposi_auto)),
                            family=poisson)
```

Negative Binomial

```
glm.nb(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                             + claims$zone1 + claims$zone2
                             + offset(log(claims$exposi_auto)))
```

Zero Poisson Inflated

```
zeroinfl(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                                + claims$zone1 + claims$zone2
                                + offset(log(claims$exposi_auto))
                                 |CPr[,1] +  CPr[,2] +  CPr[,3]
                                + claims$zone1 + claims$zone2,
                                  dist="poisson")
```

Zero Negative Binomial Inflated

```
zeroinfl(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                                + claims$zone1 + claims$zone2
                                + offset(log(claims$exposi_auto))
                                 |CPr[,1] +  CPr[,2] +  CPr[,3]
                                + claims$zone1 + claims$zone2,
                                  dist="negbin")
```

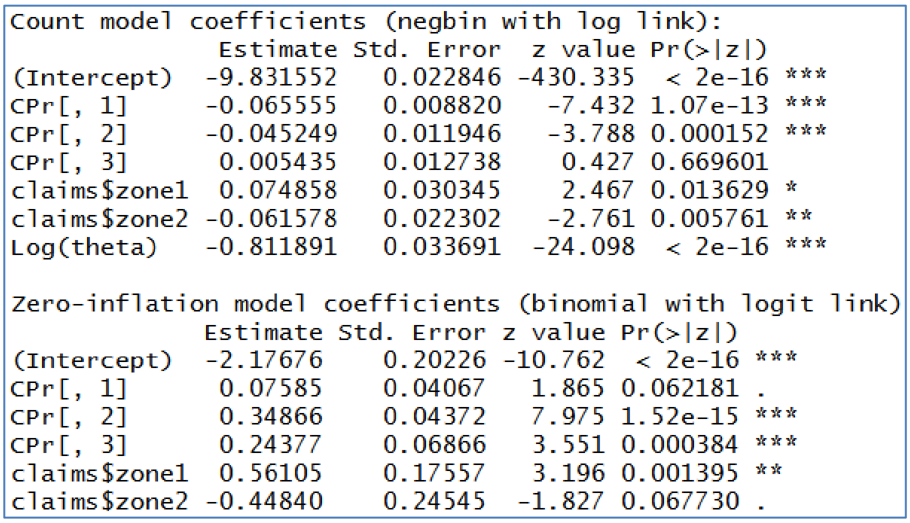
5. RESULTS

To choose one of the previous models, the Akaike Information Criteria and the Bayesian Information Criteria were considered, and they are shown in Table 7. As the Zero Inflated Negative Binomial has the lowest AIC and BIC, it is chosen as the best model to explain claims frequency with estimated personality traits as explanatory variables.

	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
AIC	208.069	183.071	189.687	183.019
BIC	208.129	183.141	189.806	183.138

Table 7. AIC and BIC comparative models' results.
Source: Own elaboration.

Figure 4 shows the estimated coefficients for the Principal Components with the Zero Inflated Negative Binomial model. First, we analyze the count model coefficients. In it we can see that the parameter associated with variable CPr1 is negative (-0,06555) and significantly different from zero (small *p*-value). This results means that the bigger CPr1 is, smaller OPE would be and, considering that parameter is negative, the lesser expected number of claims. The opposite for EXT and AGR. The bigger CPr1 is, higher EXT and AGR are, and the expected number of claims decreases.



```
Count model coefficients (negbin with log link):
              Estimate Std. Error  z value Pr(>|z|)    
(Intercept)  -9.831552   0.022846 -430.335  < 2e-16 ***
CPr[, 1]     -0.065555   0.008820   -7.432 1.07e-13 ***
CPr[, 2]     -0.045249   0.011946   -3.788 0.000152 ***
CPr[, 3]      0.005435   0.012738    0.427 0.669601    
claims$zone1  0.074858   0.030345    2.467 0.013629 *  
claims$zone2 -0.061578   0.022302   -2.761 0.005761 ** 
Log(theta)   -0.811891   0.033691  -24.098  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link)
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -2.17676    0.20226 -10.762  < 2e-16 ***
CPr[, 1]      0.07585    0.04067   1.865 0.062181 .  
CPr[, 2]      0.34866    0.04372   7.975 1.52e-15 ***
CPr[, 3]      0.24377    0.06866   3.551 0.000384 ***
claims$zone1  0.56105    0.17557   3.196 0.001395 ** 
claims$zone2 -0.44840    0.24545  -1.827 0.067730 .  
```

Figure 4. Zero Inflated Negative Binomial coefficients and p-value.
Source: Own elaboration.

When analyzing CPr2, we observe it is also negative and significant (-0,045249). This means that when CRr2 increases, the variable CON is smaller and the expected number of claims is also lesser. On the other hand, CPr3, is negative but not significantly different from zero. The associate estimated personality trait to this variable is NEU. So, according to this model, NEU has no influence over the claims frequency.

The variable zone1 has an associated parameter significantly different from zero and positive sign (0,074858); and, variable zone2 also has a significant coefficient, but with negative sign (-0,061578). This means that the expected number of claims increases for people driving in zone1; and, decreases for drivers in zone2.

Then, analyzing the Logit model, we observe that all the coefficients of the variables are significant. For every increase of CPr1, which is positive, grows the probability of declaring zero claims; but the effect on the variable is different: OPE is smaller while EXT and AGR are bigger. CPr2 is also positive, when it increases CON turns smaller and grows the probability of not declaring any claim. A positive CPr3 parameter, increases NEU when it is bigger, and with it, also increases the zero claims probability.

As for the variables zone1 and zone2, they have an opposite effect. Those driving in zone1 are not likely to report zero claims, while those who drive in zone2 are less likely to report claims zero claims.

In general terms, it could be inferred from the Zero Inflated Negative Binomial model, that:

OPE – people open to new experiences and curious have a higher expected number of claims and are more likely to declare claims.

EXT and **AGR** – introverted, reserved, quiet persons (EXT) and people less concern with others' needs, tough, critical, and uncompromising (AGR) have a bigger expected number of claims and are more likely to declare zero claims.

CON – hard workers, responsible and self-discipline people have a greater expected number of claims and are more likely to declare claims. One of the characteristics of people with high scores in CON are cautiousness, so it may not be complete coherent with the results.

NEU – emotional stability does not have influence over the claims frequency, because its parameter is not significant. But according to the logit model, people who easily get upset have more probabilities of declaring zero claims; which does not seem to be logical.

Zone1 and **Zone2** – the expected number of claims for policyholders in Madrid and Barcelona is bigger than for those who drive in the north zone of Spain. However, people in zone1 are more likely not to file a claim, contrary to what a person in the zone2 would do.

5.1. Markets’ view

Aside from the results obtained, it is quite logical to think that if the methodology of using personality traits to estimate claims probabilities is not accepted by the insured, it will have no place in the market. Based on this reasoning, I decided to make a survey, that can found in appendix *A4*, to know the interest that could be in the market to use this technic.

Using a Google Forms, I reached a small sample of 170 volunteers, basically from Ecuador, Spain and United States; to whom I asked 5 questions:

1. *Gender.*
2. *Age.*
3. *What insurance contract(s) do you have?*
   - *None*
   - *Car insurance*
   - *Motorbike insurance*
   - *Home insurance*
   - *Tech/gadgets insurance*
   - *Other (non-life) insurance*
4. *How many accidents have you declared in the last year (in total according to the previous insurance list)?*
5. *Would you be willing to provide an insurer with a text of 300 to 500 words written by you in a natural way (an email, for example); or, alternatively, to allow the insurer to access to the information from your social networks, in a legal framework, in addition to the information traditionally provided, to estimate more accurately the price of its insurance premium?*

The answers are very consistent despite of the gender, age or the number of claims. First to mention is that almost 60% of the answers belong to females and 40% to males. Age range goes from 17 to 69 years old with a clear concentration on people between 17 to 35; basically millennials. Almost 80% of the sample has not report a claim in the past year and 15% has declared only 1 claims; Figure 5 summarizes this claims frequency.

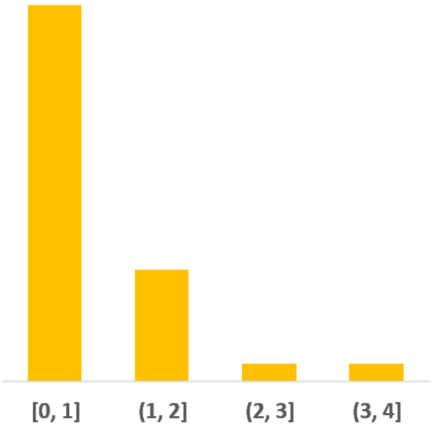


Figure 5. Histogram of declared claims in the survey.
Source: Own elaboration.

When answering the main question of the survey "*Would you be willing to provide an insurer with a text of 300 to 500 words written by you in a natural way (an email, for example); or, alternatively, to allow the insurer to access to the information from your social networks, in a legal framework, in addition to the information traditionally provided, to estimate more accurately the price of its insurance premium?";* there was a consistent response around 65% who said NO and 35% who said YES.

By categories, I received the following answers:

- General answer. Figure 6 shows that in general terms, 34% of the sample would be willing to provide access to an insurance company to their social networks.

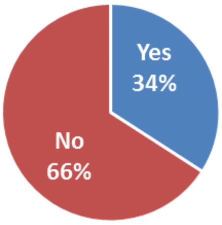


Figure 6. General response from the market survey.
Source: Own elaboration.

- Millenials. We observe in Figure 7 that 63% of people from the sample, born between 1981 and 1995, approximately; who are described as technology and telecommunications lovers, won't accept to provide access to their social networks to obtain a more accurate premium.

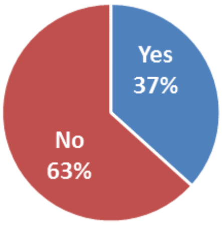


Figure 7. Millenials response from the market survey.
Source: Own elaboration.

- People older than 60 years. We can see on Figure 8, that 40% of the old people from the sample will accept providing a personal text to an insurer, in order to pay a more precise premium.

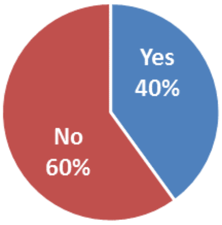


Figure 8. People older than 60 years response from the market survey.
Source: Own elaboration.

- No claims. Figure 9 summarizes the answers of people who have not had any claim; and 67% of them does not agree on sharing personal information to an insurance company.

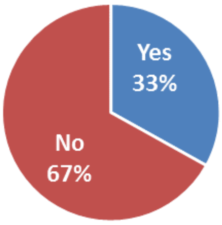


Figure 9. No claims response from the market survey.
Source: Own elaboration.

- One claim. On the other hand, Figure 10 shows people with one declared claim in the past year; and 35% of this collective believes in providing access to personal information to receive a benefit/punishment on the premium they pay.

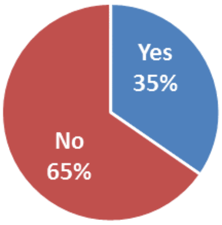


Figure 10. One claim response from the market survey.
Source: Own elaboration.

- Two or more claims. Figure 11 shows, maybe, the most interesting fact: 50% of the people who have two or more claims, will accept to provide access to their social networks to receive an accurate premium.

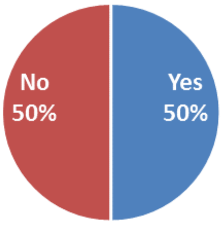


Figure 11. Two or more claims response from the market survey.
Source: Own elaboration.

As it is a small sample, I could not disaggregate more the answers, because more specific cases are represented only by a few individuals. Even though, some information can be inferred. Two thirds of the surveyed would not be willing to provide their Social Networks information to have a more accurate premium. I cannot tell the exact reasons, but I guess is basically for privacy issues.

6. CONCLUSIONS

The idea proposed in this thesis is based on recent researches that have proven that it is possible to categorize people’s personality into five general factors, from data collected on social networks. The intention is to show one of the applications that Big Data of social networks could have in the insurance industry, specifically in relation to the modeling of the number of claims.

It is possible to offer to the policyholder a premium adjusted to a more real risk profile; based on their presence on the social networks, in addition to the information traditionally considered. The practical result of this hypothesis could lead to a fair trait for the customers and to an increase of the revenue of the companies. Incorporating the variable *personality*, could benefit both, policyholders and the insurance company. Clients could pay the premium that matches their real risk, instead of paying an extra premium for their age, type of car or even color of vehicle.

Reviewing the best model estimates, Zero Inflated Negative Binomial, we can conclude, for the count model, that people scoring high in OPE, it is curious, creative and open to new experiences, have a higher expected number of claims is more, than for those who are practical and conservative. Individual with positives scores in AGR and EXT, that includes empathetic subjects and people seeking adventure, will have a greater expected number of claims. We can also conclude that high scores for CON are prone to a higher expected number of claims; this includes people discipline and hardworking people.

Evidently, a deeper study is needed, because we cannot anticipate a reaction on the number of claims when the explanatory variable is a personality trait. These nontraditional variables need to be highly understood in their context and relation among them. I must mention that the resources, specially information, were limited, so it is not possible to accept or deny any hypothesis. An idea has been launched and further studies are needed in other to identify if there is a real correlation among certain personality traits and claims frequency.

Finally, I center my attention on the portion of the sample who would provide their information. Today people has control over the market prices and insurance companies struggle with fidelity problems (retention ratio), because policyholders go to the cheapest offer (considering that the product is almost the same). A personalized premium can be used as a marketing strategy for new business. I believe that when people realize that they are paying for the risk they represent, they will stop floating from one company to another, and maybe, they will stay with that one company who let them know their real risk profile. 35% of the surveyed are a lot of potential policyholders expecting to pay, more or less, but for a more accurate premium.

7. REFERENCES

Alemany, R. (2015). Modelos Lineales Generalizados. Barcelona (España): Departament d'Econometria, Estadística i Ec. Espanyol.

Ashton, M., & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality*, *19*(April 2003), 5–24. http://doi.org/10.1002/per.541

Biel, J., & Gatica-perez, D. (2013). The YouTube Lens : Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs, *15*(1), 41–55.

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (n.d.). The Datafication of Talent: How Technology is Advancing the Science of Human Potential at Work. *Current Opinion in the Behavioral Sciences*.

Charpentier, A. (2015). *Computational Actuarial Science with R*. Boca Raton (U.S.): CRC Press.

de Jong, P., & Heller, G. Z. (2013). *Generalized Linear Models for Insurance Data*. New York (U.S.): Cambridge University Press.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., … De Cock, M. (2016). Computational personality recognition in social media. *User Modelling and User-Adapted Interaction*, *26*(2–3), 109–142. http://doi.org/10.1007/s11257-016-9171-0

Foster, I., Ghani, R., Jarmin, R., Kreuter, F., & Lane, J. (2017). *Bid Data and Social Science*. Boca Raton (U.S.): CRC Press.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. New York (U.S.): Cambridge University Press.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *The American Psychologist*, *48*(1), 26–34. http://doi.org/10.1037/0003-066X.48.12.1302

Guillén, M. (2014). Regression with categorical dependent variables. In *Predictive modelling applications in Actuarial Science, Volume I: Predictive Modelling Techniques* (pp. 65–78). New York (U.S.): Cambridge University Press.

Jamil, Z. (2017). *Monitoring tweets for depression to detect at-risk users*. University of Ottawa.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., … Usher, A. (2015). *Big data in survey research: Aapor task force report*. *Public Opinion Quarterly* (Vol. 79). http://doi.org/10.1093/poq/nfv039

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, *2*(3), 171–203. http://doi.org/10.1002/per.2410020302

Johnson, J. A. (2016). Descriptions Used in the IPIP-NEO Narrative Report. Retrieved May 27, 2017, from http://www.personal.psu.edu/~j5j/IPIP/IPIPNEOdescriptions.html

Kennedy, W. G. (2012). Modelling Human Behaviour in Agent-Based Models (pp. 167–179). London (UK): Springer. http://doi.org/10.1007/978-90-481-8927-4

Kosinski, M. (2013). Mining Big Data to Extract Patterns and Predict Real-Life Outcomes. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. http://doi.org/10.1017/CBO9781107415324.004

Kosinski, M. (2017a). Interview with Fredrik Skavlanle. Sweden. Retrieved from https://wn.com/interview_with_fredrik_skavlan_(2017)

Kosinski, M. (2017b). The End of Privacy. Hannover: CeBIT Global Conferences. Retrieved from http://www.cebit.de/event/the-end-of-privacy/VOR/74931

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015). Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. American Psychologist. Retrieved from http://mypersonality.org/wiki/doku.php?id=download_databases

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5. http://doi.org/10.1073/pnas.1218772110

Kulkarni, V., Kern, M. L., Stillwell, D., Kosinski, M., Matz, S., Ungar, L., … Schwartz, H. A. (2017). Latent Human Traits in the Language of Social Media: An Open-Vocabulary Approach, 1–21. Retrieved from http://arxiv.org/abs/1705.08038

Lambiotte, R., & Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, *102*(12), 1934–1939. http://doi.org/10.1109/JPROC.2014.2359054

Lepri, B., Staiano, J., Shmueli, E., Pianesi, F., & Pentland, A. (2016). The role of personality in shaping social networks and mediating behavioral change. *User Modelling and User-Adapted Interaction*, *26*(2–3), 143–175. http://doi.org/10.1007/s11257-016-9173-y

Matthews, G., Deary, I., & Whiteman, M. (2009). *Personality Traits* (Cambridge). Cambridge.

Ohlsson, E., & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. London (UK): Springer. http://doi.org/10.1007/978-3-642-10791-7

Parodi, P. (2015). *Pricing in General Insurance*. Boca Raton (U.S.): CRC Press.

Preot, D., Hopkins, D. J., & Ungar, L. (2017). Beyond Binary Labels : Political Ideology Prediction of Twitter Users.

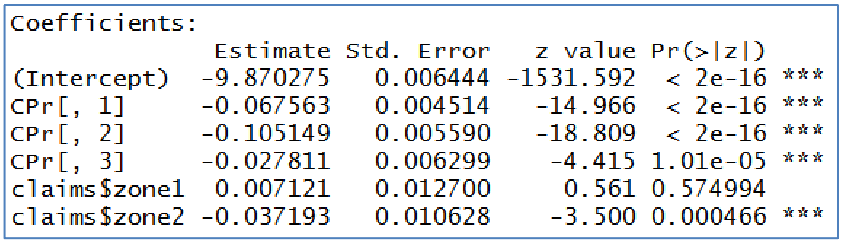
Saha, K., Weber, I., Birnbaum, M. L., & De Choudhury, M. (2017). Characterizing

Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates. *Journal of Medical Internet Research*, *19*(5), e156. http://doi.org/10.2196/jmir.6815

Whitty, M., Doodson, J., Creese, S., & Hodges, D. (2014). Social Computing and Social Media. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8531*, 528–537. http://doi.org/10.1007/978-3-319-07632-4

8. APPENDIX

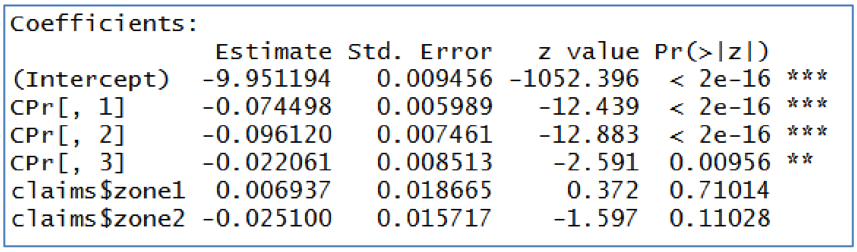
A1. Poisson Model R output.



```
Coefficients:
               Estimate Std. Error   z value Pr(>|z|)    
(Intercept)   -9.870275   0.006444 -1531.592  < 2e-16 ***
CPr[, 1]      -0.067563   0.004514   -14.966  < 2e-16 ***
CPr[, 2]      -0.105149   0.005590   -18.809  < 2e-16 ***
CPr[, 3]      -0.027811   0.006299    -4.415 1.01e-05 ***
claims$zone1   0.007121   0.012700     0.561 0.574994    
claims$zone2  -0.037193   0.010628    -3.500 0.000466 ***
```

Figure 12. Poisson coefficients and p-value.
Source: Own elaboration.

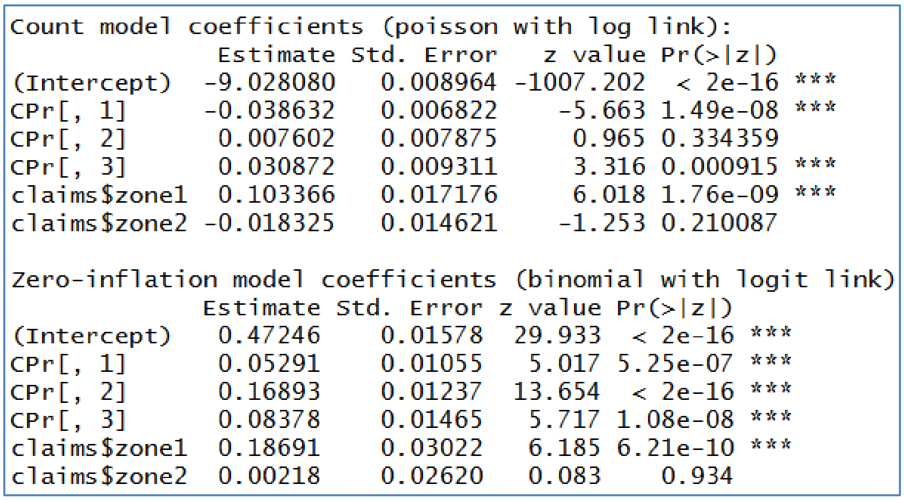
A2. Negative Binomial Model R output.



```
Coefficients:
               Estimate Std. Error   z value Pr(>|z|)    
(Intercept)   -9.951194   0.009456 -1052.396  < 2e-16 ***
CPr[, 1]      -0.074498   0.005989   -12.439  < 2e-16 ***
CPr[, 2]      -0.096120   0.007461   -12.883  < 2e-16 ***
CPr[, 3]      -0.022061   0.008513    -2.591  0.00956 ** 
claims$zone1   0.006937   0.018665     0.372  0.71014    
claims$zone2  -0.025100   0.015717    -1.597  0.11028    
```

Figure 13. Negative Binomial coefficients and p-value.
Source: Own elaboration.

A3. Zero Inflated Poisson Model R output.



```
Count model coefficients (poisson with log link):
              Estimate Std. Error   z value Pr(>|z|)    
(Intercept)  -9.028080   0.008964 -1007.202  < 2e-16 ***
CPr[, 1]     -0.038632   0.006822    -5.663 1.49e-08 ***
CPr[, 2]      0.007602   0.007875     0.965 0.334359    
CPr[, 3]      0.030872   0.009311     3.316 0.000915 ***
claims$zone1  0.103366   0.017176     6.018 1.76e-09 ***
claims$zone2 -0.018325   0.014621    -1.253 0.210087    

Zero-inflation model coefficients (binomial with logit link)
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   0.47246    0.01578  29.933  < 2e-16 ***
CPr[, 1]      0.05291    0.01055   5.017 5.25e-07 ***
CPr[, 2]      0.16893    0.01237  13.654  < 2e-16 ***
CPr[, 3]      0.08378    0.01465   5.717 1.08e-08 ***
claims$zone1  0.18691    0.03022   6.185 6.21e-10 ***
claims$zone2  0.00218    0.02620   0.083    0.934    
```

Figure 14. Zero Inflated Poisson coefficients and p-value.
Source: Own elaboration.

A4. Survey

N	1. Gender	2. How old are you?	3. What insurance contract(s) do you have?	4. How many accidents have you declared in the last year (in total according to the previous insurance list)?	5. Would you be willing to provide an insurer with a text of 300 to 500 words written by you in a natural way (an email, for example); or, alternatively, to allow the insurer to access to the information from your social networks, in a legal framework, in addition to the information traditionally provided, to estimate more accurately the price of its insurance premium?
1	Male	30	None	0	Yes
2	Female	26	None	0	No
3	Male	31	Car insurance	0	No
4	Female	56	Car insurance	0	Yes
5	Female	23	None	0	No
6	Male	26	Car insurance	0	No
7	Male	28	None	0	Yes
8	Female	26	Car insurance	0	No
9	Female	57	None	0	No
10	Male	28	Car insurance, Home insurance, Tech gadgets (cell phone, laptop, etc.) insurance	0	No
11	Male	56	Car insurance, Home insurance	0	No
12	Male	27	Car insurance	0	No
13	Female	54	Car insurance	0	No
14	Male	28	Motorbike insurance	0	Yes
15	Male	62	None	0	No
16	Male	19	Car insurance, Home insurance	0	No
17	Female	27	None	0	Yes
18	Male	57	Car insurance	0	No
19	Female	34	Car insurance	0	No
20	Male	34	Car insurance, Home insurance	0	Yes
21	Male	63	None	0	No
22	Male	42	Car insurance	0	Yes
23	Male	54	None	0	No
24	Female	32	None	0	No

25	Female	39	Car insurance, Home insurance	0	Yes
26	Male	61	None	0	No
27	Female	69	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	0	No
28	Male	43	None	0	Yes
29	Female	56	Car insurance	0	No
30	Male	50	Car insurance, Home insurance	0	Yes
31	Female	58	Car insurance	0	Yes
32	Female	46	None	0	Yes
33	Male	46	None	0	Yes
34	Female	38	Others (only insurance contracts for things not for people)	0	No
35	Male	45	Car insurance	0	Yes
36	Female	44	None	0	No
37	Male	53	Car insurance, Home insurance , Tech gadgets (cell phone, laptop, etc.) insurance, Others	0	No
38	Female	57	None	0	Yes
39	Female	57	None	0	Yes
40	Male	45	Home insurance	0	No
41	Female	27	Others (only insurance contracts for things not for people)	0	No
42	Male	54	Car insurance	0	No
43	Female	27	Car insurance	0	No
44	Male	32	Car insurance	0	No
45	Male	50	Car insurance, Home insurance	0	No
46	Male	64	Car insurance, Home insurance, Others	0	Yes
47	Female	49	Car insurance, Home insurance, Tech gadgets (cell phone, laptop, etc.) insurance	0	No
48	Female	27	None	0	No
49	Female	49	Car insurance	0	Yes
50	Female	21	Tech gadgets (cell phone, laptop, etc.) insurance	0	No
51	Female	25	Car insurance	0	No
52	Female	28	None	0	No
53	Female	26	Car insurance	0	No
54	Female	34	None	0	No
55	Female	34	None	0	No

56	Male	47	Car insurance, Home insurance	0	No
57	Female	59	Car insurance	0	No
58	Female	57	None	0	No
59	Female	27	None	0	No
60	Female	59	Car insurance	0	No
61	Male	23	None	0	No
62	Female	27	None	0	Yes
63	Female	26	Home insurance	0	Yes
64	Female	22	None	0	No
65	Female	23	None	0	No
66	Female	22	None	0	No
67	Male	27	None	0	No
68	Female	22	None	0	No
69	Female	23	None	0	No
70	Male	21	None	0	No
71	Female	24	None	0	Yes
72	Female	22	Home insurance	0	No
73	Female	54	None	0	Yes
74	Female	22	None	0	No
75	Male	25	None	0	No
76	Female	22	None	0	Yes
77	Female	59	Car insurance	0	No
78	Female	24	None	0	Yes
79	Female	59	Car insurance	0	No
80	Female	59	Car insurance	0	No
81	Female	20	None	0	Yes
82	Male	22	None	0	Yes
83	Female	59	Car insurance	0	No
84	Female	26	Car insurance, Home insurance	0	No
85	Female	59	Car insurance	0	No
86	Female	29	Car insurance, Home insurance	0	No
87	Male	40	Home insurance, Tech gadgets (cell phone, laptop, etc.) insurance	0	No
88	Female	19	Tech gadgets (cell phone, laptop, etc.) insurance	0	No

89	Female	55	Car insurance, Home insurance	0	No
90	Female	20	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	0	No
91	Female	20	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	0	No
92	Male	20	None	0	No
93	Male	49	Car insurance, Home insurance	0	No
94	Female	59	None	0	No
95	Male	24	None	0	No
96	Female	55	Car insurance, Motorbike insurance, Home insurance, Tech gadgets (cell phone, laptop, etc.) insurance, Others (only insurance contracts for things not for people)	0	No
97	Female	41	None	0	No
98	Female	35	None	0	No
99	Female	54	None	0	No
100	Female	30	None	0	No
101	Male	57	Car insurance	0	No
102	Female	21	None	0	Yes
103	Male	22	None	0	Yes
104	Female	23	None	0	Yes
105	Male	38	Car insurance, Home insurance	0	No
106	Female	28	None	0	Yes
107	Male	42	None	0	No
108	Male	22	Car insurance	0	Yes
109	Female	26	None	0	No
110	Female	57	None	0	No
111	Female	28	None	0	No
112	Female	34	None	0	No
113	Male	28	Others (only insurance contracts for things not for people)	0	No
114	Female	49	None	0	No
115	Male	50	Car insurance, Home insurance	0	Yes
116	Male	29	None	0	No
117	Female	51	Car insurance, Home insurance	0	No
118	Male	18	None	0	No
119	Female	63	None	0	Yes

120	Female	21	None	0	Yes
121	Female	62	None	0	No
122	Male	22	None	0	Yes
123	Male	31	Car insurance	0	No
124	Male	27	Car insurance, Home insurance	0	No
125	Female	22	Others (only insurance contracts for things not for people)	0	No
126	Male	30	None	0	No
127	Male	19	Car insurance	0	Yes
128	Female	55	Car insurance	0	Yes
129	Female	32	None	0	Yes
130	Female	17	None	0	Yes
131	Female	29	None	0	Yes
132	Female	47	None	0	Yes
133	Male	44	Car insurance, Home insurance	0	Yes
134	Male	27	None	0	Yes
135	Female	27	Tech gadgets (cell phone, laptop, etc.) insurance	0	Yes
136	Male	23	Car insurance	0	Yes
137	Male	67	Car insurance, Others (only insurance contracts for things not for people)	1	Yes
138	Male	65	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	1	Yes
139	Male	53	Car insurance, Home insurance , Tech gadgets (cell phone, laptop, etc.) insurance, Others	1	Yes
140	Male	53	Car insurance, Home insurance , Tech gadgets (cell phone, laptop, etc.) insurance, Others	1	Yes
141	Female	30	Car insurance	1	Yes
142	Female	22	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	1	Yes
143	Male	41	Car insurance, Home insurance, Tech gadgets (cell phone, laptop, etc.) insurance	1	Yes
144	Female	31	Car insurance	1	No
145	Female	55	Car insurance	1	No
146	Female	33	None	1	Yes
147	Male	23	Car insurance, Home insurance	1	No
148	Female	42	Tech gadgets (cell phone, laptop, etc.) insurance	1	No

149	Male	29	Car insurance	1	No
150	Male	45	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	1	No
151	Female	32	Home insurance	1	No
152	Female	39	Car insurance, Home insurance	1	No
153	Male	52	Car insurance, Home insurance, Others	1	No
154	Female	55	None	1	No
155	Female	26	Motorbike insurance	1	No
156	Female	23	Motorbike insurance	1	No
157	Female	23	Car insurance	1	No
158	Female	43	Car insurance	1	No
159	Male	35	Car insurance, Home insurance	1	No
160	Male	61	Car insurance, Home insurance	1	No
161	Male	37	Car insurance	1	No
162	Male	23	Motorbike insurance, Tech gadgets (cell phone, laptop, etc.) insurance	1	Yes
163	Female	23	Others (only insurance contracts for things not for people)	2	Yes
164	Female	21	Tech gadgets (cell phone, laptop, etc.) insurance	2	Yes
165	Female	21	Tech gadgets (cell phone, laptop, etc.) insurance	2	Yes
166	Female	22	Tech gadgets (cell phone, laptop, etc.) insurance	2	No
167	Male	23	None	2	No
168	Male	25	Car insurance	2	No
169	Male	23	Car insurance, Tech gadgets (cell phone, laptop, etc.) insurance	3	No
170	Male	29	None	4	Yes

A5. R Code

```
############################
#      MASTER THESIS       #
#   Andrés Durán Proaño    #
#       June / 2017        #
#   Universy of Barcelona  #
############################

setwd("C:/Users/Andres/Desktop/UB/IV. Feb.Jul2017/TFM/Soporte investigativo")

######  Read-in files:
# I. #  - mypersonality database
######  - claims database

# users DB
users <- read.csv("users.csv")
users[,10] <- users$age^2
colnames(users)[10] <- "age2"
users[,11] <- users$gender*users$age
colnames(users)[11] <- "genage"

summary(users$ope)
summary(users$con)
summary(users$ext)
summary(users$agr)
summary(users$neu)
summary(users$age)
summary(users$age)
hist(users$ope)
hist(users$con)
hist(users$ext)
hist(users$agr)
hist(users$neu)
hist(users$age)
hist(users$age2,100)

# claims DB
claims <- read.table("claims_dat.csv", header = TRUE, sep = ";", dec = ",",
                      row.names =   NULL)
claims <- claims[(claims$client_age > 17 & claims$client_age <= 90),]
nrow(claims)
claims <- na.omit(claims)
nrow(claims)

claims[,22] <- claims$client_gender*claims$client_age
colnames(claims)[22] <- "client_genage"

#######   1. Estimate for each personality trait its coefficients in function of the age
# II. #      and sex, through a Linear Regression Model.
#######   2. Infer each personality trait for clients in claims_data database.

# 1. Openness - Multiple
reg_ope <-lm(users$ope~users$gender+users$age+users$age2+users$genage)
coef_ope <-as.matrix(coef(reg_ope))
summary(reg_ope)

x_ope <- cbind(claims$client_gender,claims$client_age,claims$client_age2,
               claims$client_genage)
x_ope <-as.matrix(x_ope)
n <-nrow(x_ope)
ones <-matrix(1,n,1)
x_ope <-cbind(ones,x_ope)

hat_ope <- matrix(0,n,1)
for(i in 1:n){
  hat_ope[i] <- (x_ope[i, ]%*%coef_ope)
}
```

```
summary(hat_ope)
hist(hat_ope)

hat_opes <- (hat_ope - mean(hat_ope))/sd(hat_ope)
summary(hat_opes)
hist(hat_opes)

# 2. Conscientiousness - Multiple
reg_con<-lm(users$con~users$gender+users$age+users$age2+users$genage)
coef_con<-as.matrix(coef(reg_con))
summary(reg_con)

x_con <- cbind(claims$client_gender,claims$client_age,claims$client_age2,
               claims$client_genage)
x_con<-as.matrix(x_con)
n<-nrow(x_con)
ones<-matrix(1,n,1)
x_con<-cbind(ones,x_con)

hat_con<-matrix(0,n,1)
for(i in 1:n){
  hat_con[i]<-(x_con[i,]%*%coef_con)
}
summary(hat_con)
hist(hat_con,100)

hat_cons<- (hat_con-mean(hat_con))/sd(hat_con)
summary(hat_cons)
hist(hat_cons)

# 3. Extroversion - Multiple
reg_ext<-lm(users$ext~users$gender+users$age+users$age2+users$genage)
coef_ext<-as.matrix(coef(reg_ext))
summary(reg_ext)

x_ext <- cbind(claims$client_gender,claims$client_age,claims$client_age2,
               claims$client_genage)
x_ext<-as.matrix(x_ext)
n<-nrow(x_ext)
ones<-matrix(1,n,1)
x_ext<-cbind(ones,x_ext)

hat_ext<-matrix(0,n,1)
for(i in 1:n){
  hat_ext[i]<-(x_ext[i,]%*%coef_ext)
}
summary(hat_ext)
hist(hat_ext,100)

hat_exts<- (hat_ext-mean(hat_ext))/sd(hat_ext)
summary(hat_exts)
hist(hat_exts)

# 4. Agreeableness - Multiple
reg_agr <-lm(users$agr~users$gender+users$age+users$age2+users$genage)
coef_agr <-as.matrix(coef(reg_agr))
summary(reg_agr)

x_agr <- cbind(claims$client_gender,claims$client_age,claims$client_age2,
               claims$client_genage)
x_agr <-as.matrix(x_agr)
n<-nrow(x_agr)
ones <-matrix(1,n,1)
x_agr <-cbind(ones,x_agr)

hat_agr<-matrix(0,n,1)
for(i in 1:n){
  hat_agr[i] <-(x_agr[i,]%*%coef_agr)
}
```

```
summary(hat_agr)
hist(hat_agr,100)

hat_agrs <- (hat_agr - mean(hat_agr))/sd(hat_agr)
summary(hat_agrs)
hist(hat_agrs)

# 5. Neurosis - Multiple
reg_neu <-lm(users$neu~users$gender+users$age+users$age2+users$genage)
coef_neu <-as.matrix(coef(reg_neu))
summary(reg_neu)

x_neu <- cbind(claims$client_gender,claims$client_age,claims$client_age2,
               claims$client_genage)
x_neu <-as.matrix(x_neu)
n <-nrow(x_neu)
ones <-matrix(1,n,1)
x_neu<-cbind(ones,x_neu)

hat_neu<-matrix(0,n,1)
for(i in 1:n){
  hat_neu[i]<-(x_neu[i,]%*%coef_neu)
}
summary(hat_neu)
hist(hat_neu,100)

hat_neus<- (hat_neu-mean(hat_neu))/sd(hat_neu)
summary(hat_neus)
hist(hat_neus)

# Coefficients summary
betas_m <- cbind(coef_ope, coef_con, coef_ext, coef_agr, coef_neu)
betas_m <- as.matrix(betas_m)
colnames(betas_m)<- c("b_ope","b_con","b_ext","b_agr","b_neu")
write(t(betas_m), file = "betas_m", ncolumns = 5, sep = "\t" )

########
# III. #    Hypothesis validation: GLM Models for claims count
########

# Principal components analysis

per<-as.matrix(cbind(hat_opes, hat_cons, hat_exts, hat_agrs, hat_neus))
colnames(per)<- c("hat_opes","hat_cons","hat_exts","hat_agrs","hat_neus")
summary(per)

SX<-cov(per)
det(SX)
valvec<-eigen(SX)
valvec
val<-valvec$values
val
PC<-c(1,2,3,4,5)
plot(PC,val,"l")
varexp<-(val/sum(val))*100
varexp
vec<-valvec$vectors
B<-vec[,1:3]
B
omega<-diag(val[1:3])
omega
BB<-B%*%sqrt(omega)
BB
varimax(BB, normalize = TRUE, eps = 1e-5)

rot<-c(0.9053860989, 0.216325653,  0.365348086,
       -0.0173513138, 0.878606492, -0.477231144,
       -0.4242345386, 0.425739174,  0.799231638)
rot<-matrix(rot, 3,3)
```

```
rot<-t(rot)

BB

BBr<-BB%*%rot
BBr

CP<-per%*%valvec$vectors[,1:3]
CPr<-CP%*%rot

claims_pts<-as.matrix(cbind(claims, CPr))

# CLAIMS MODELS

# Poisson
poisson_auto_m <- glm(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                      +  claims$zone1  + claims$zone2  + offset(log(claims$exposi_auto)),
family=poisson)
summary(poisson_auto_m)
BIC(poisson_auto_m)

# Negative Binomial
library(MASS)

nb_auto_m <- glm.nb(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                    + claims$zone1 + claims$zone2 + offset(log(claims$exposi_auto)))
summary(nb_auto_m)
BIC(nb_auto_m)

# Zero Inflated Poisson
library(pscl)

zip_auto_m <- zeroinfl(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                       + claims$zone1 + claims$zone2 + offset(log(claims$exposi_auto))
                       |CPr[,1] +  CPr[,2] +  CPr[,3]
                       + claims$zone1 + claims$zone2, dist="poisson")
summary(zip_auto_m)
AIC(zip_auto_m)
BIC_zip <- -2*zip_auto_m$loglik + 12*log(zip_auto_m$n)
BIC_zip

zinb_auto_m <- zeroinfl(claims$nclaims_auto ~ CPr[,1] +  CPr[,2] +  CPr[,3]
                        + claims$zone1 + claims$zone2 + offset(log(claims$exposi_auto))
                        |CPr[,1] +  CPr[,2] +  CPr[,3]
                        + claims$zone1 + claims$zone2, dist="negbin")
summary(zinb_auto_m)
AIC(zinb_auto_m)
BIC_zinb <- -2*zinb_auto_m$loglik + 12*log(zinb_auto_m$n)
BIC_zinb
```