BigData and machine learning

An application to the design of car insurances

Gonzalo Ruiz Manzanares

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

Unión Europea
Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"

codeoscopic

Instituto Universitario de Investigación
Biocomputación y Física
de Sistemas Complejos
Universidad Zaragoza

Development group

- Computation Area -> Development group
  - Team
    - Head -> David Íñiguez
    - Members
      - Alfonso Tarancón
      - Alejandro Rivero
      - Alfredo Ferrer
      - Gonzalo Ruiz
      - [Rubén Moreno]
  - Lines of work
    - Data Analysis
    - Advanced Visualization
    - Software development (web, mobile, etc.)

Index

- Introduction
- Internal Data
- External Open Data
- Preliminary data analysis
- Modelling
- Infraestructure
- Results

Introduction

- Nowadays -> insurances for everything
- Insurance sector represents around a 7% of Spain's GDP
- Car insurance is growing more than other insurance types like life, home or accident insurances
- Very important sector
  - Traditional companies
  - New companies (online, phone…)

Introduction

- Codeoscopic -> Car insurance specialized Company
- Their main product is Avant2 -> used by 25% of insurance agents
- Agents introduce data about a specific risk
  - Vehicle, driver, ocasional driver, locality…
- Avant2 evaluate this risk with many insurance companies and get quotes for different modalities

Introduction

- This process is like a black box
- What criteria do insurance companies apply to get this quotes? -> key of each one
- Complex statistic and mathematical models -> actuarial science
- Codeoscopic wants to
  - Understand differences
  - Improve their products
- At the beginning of this Project
  - More than 200 million quotes
- Using BigData and Machine Learning techniques to infer those rules -> project Retos-Colaboración DEEP CODE

Internal data

- 10 of the most important companies in Spain from (2014-2015)
- Around 100 million prices a year
- Around 200.000 policies emitted per year
- Sample of available data:
  - Modality and coverages (third party, comprehensive…)
  - Price (with deductible if applies)
  - Vehicle (Brand, model, age, power, price, displacement, kilometers…)
  - Main and ocasional drivers (gender, age, marital status, driving experience, caused accidents, years insured…)

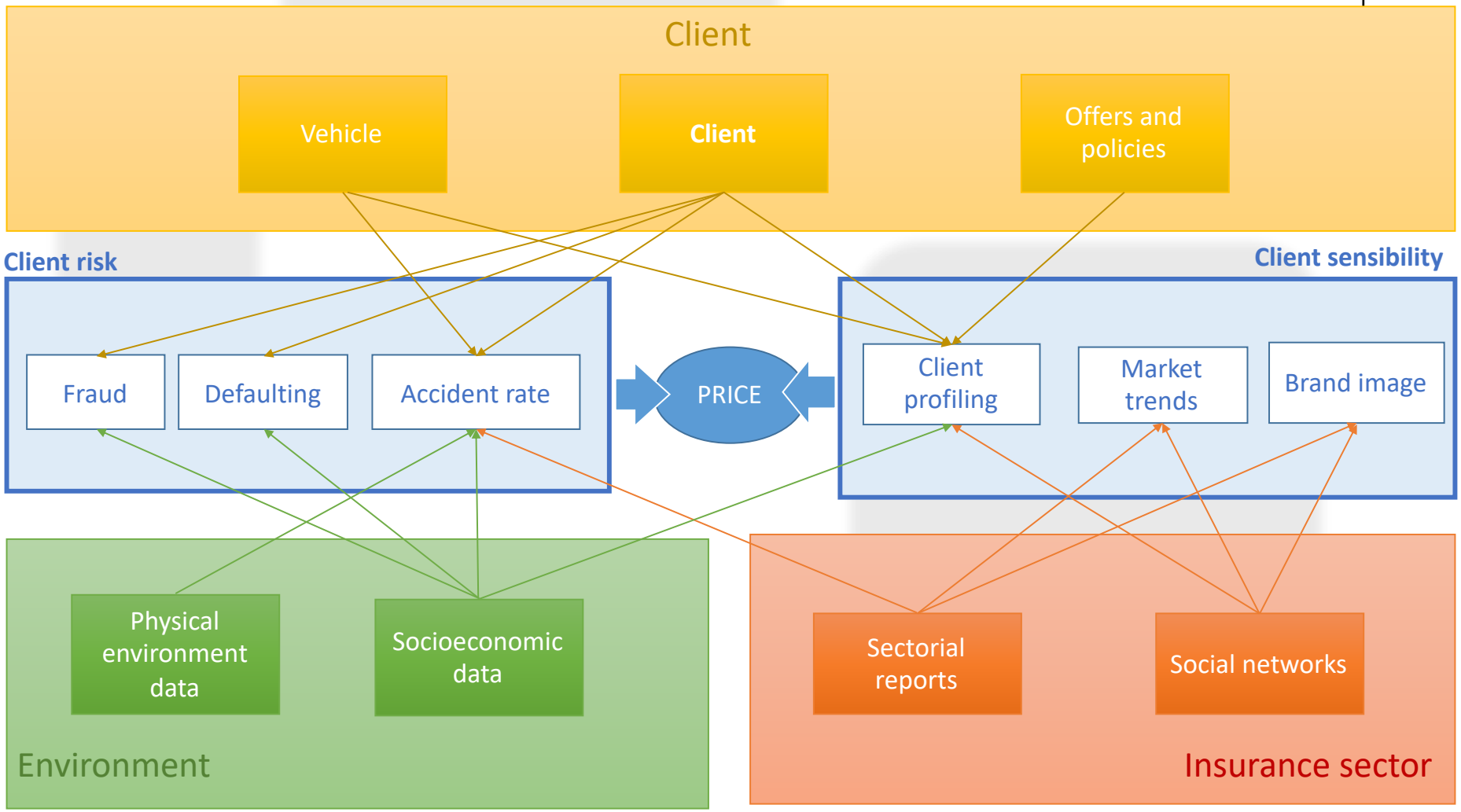
External open data

- There are big differences in prices for very similar risks
- They can't be explained by internal data
- They seem to be connected to the risk location
- Are they related to external data?
  - Do accidents happen more often on rainy regions?
  - Are accidents more frequent in regions with older cars? Or with younger drivers?
  - Do fraud and defaulting affect final prices?

External open data

- Around 10 million registers:
  - DGT: vehicles, accidents, drivers, car transfer, driving licenses, road problems…
  - Ministry of Development: traffic intensity, road speeds
  - Twitter DGT: campaings, alerts, information
  - INE: socio-economic information, census
  - Unemployment rates by age, sex etc.
  - AEMET: historic meteorologic data
  - AEAT: average income

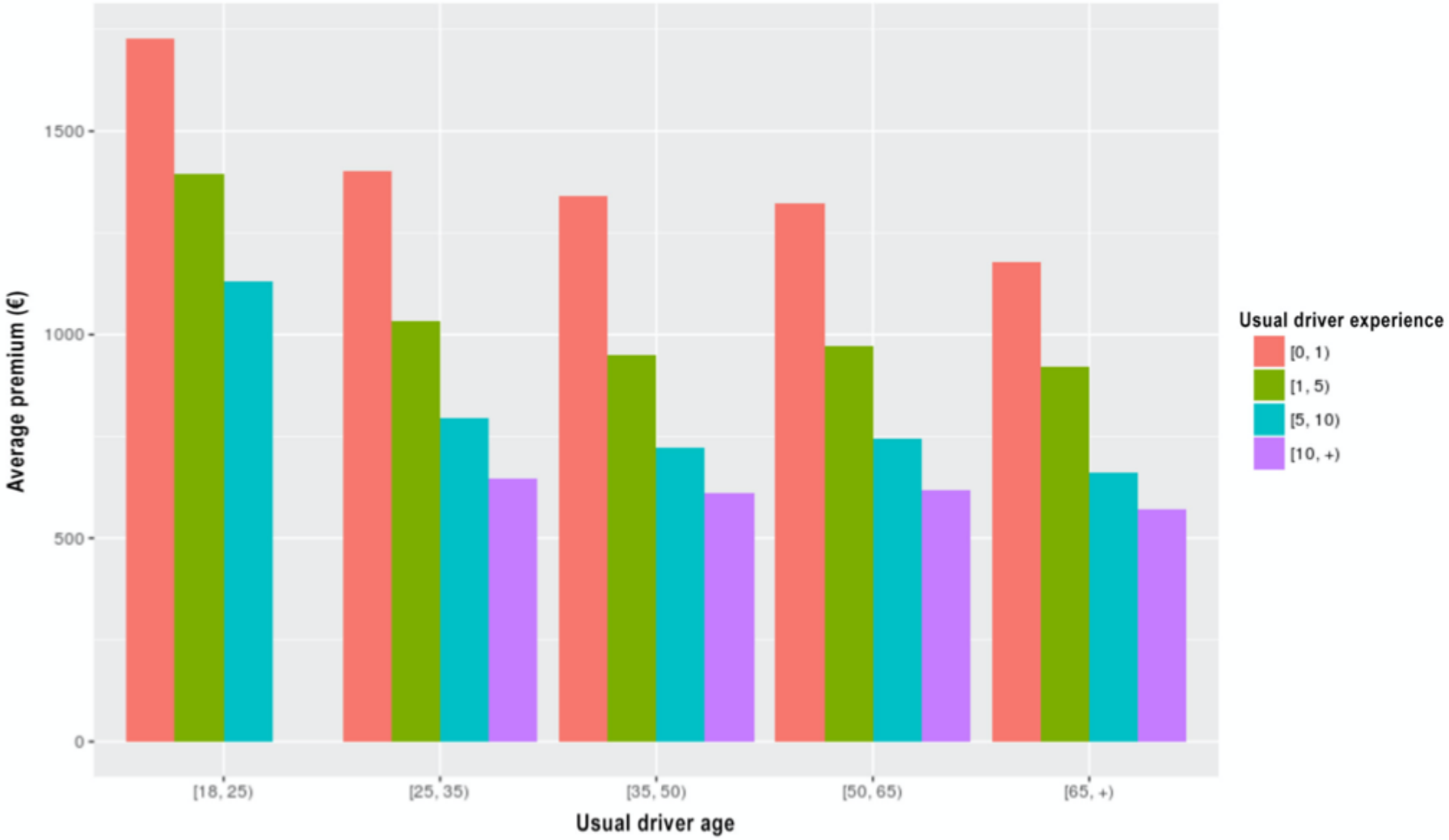
Panoramic view



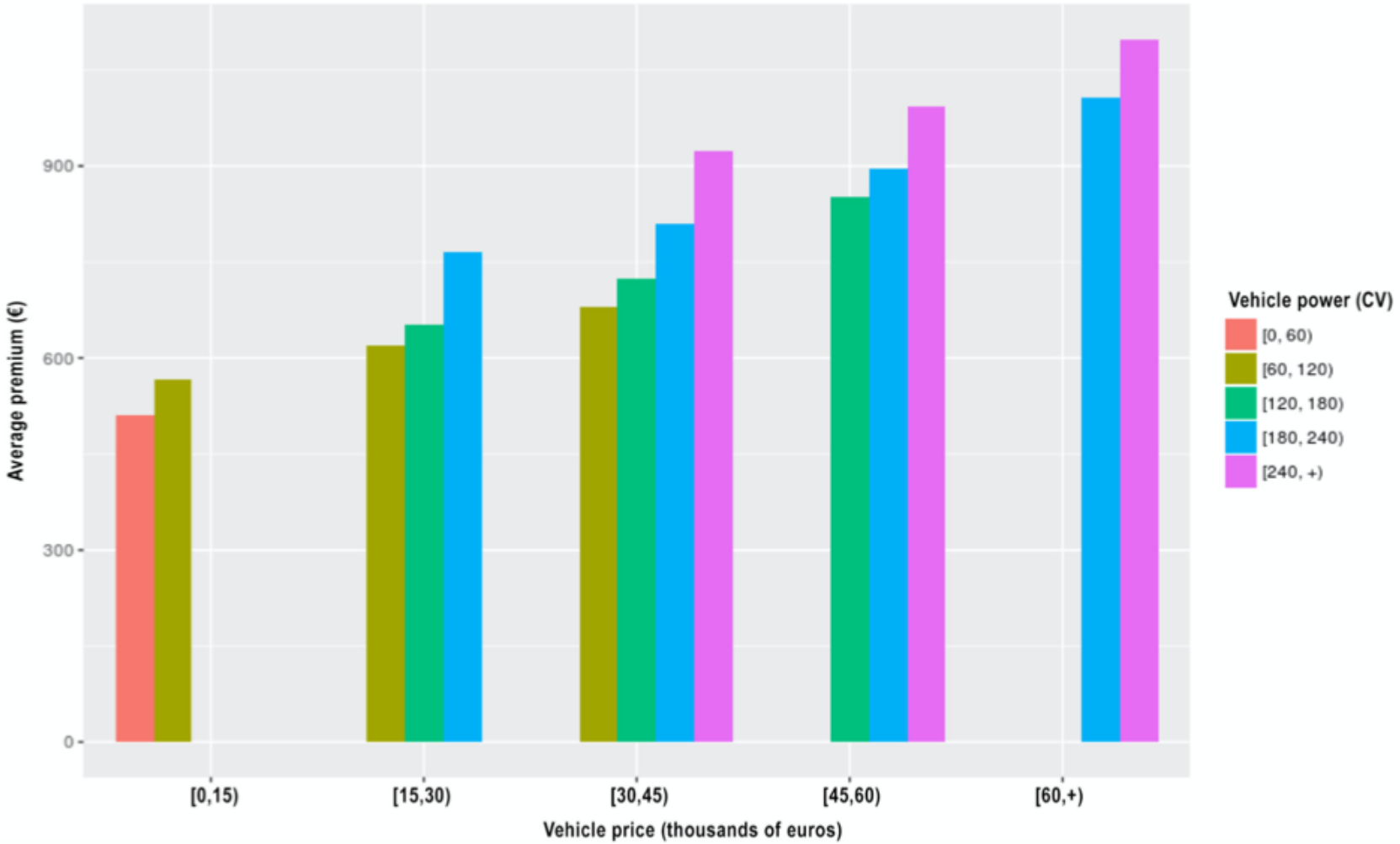
Preliminary data analysis

- BigData infraestructure at BIFI cloud to preprocess all data
  - ElasticSearch: very powerful search engine based on Apache Lucene.
  - Kibana: visualization interface over ElasticSeach
  - 5 OpenStack nodes: m1.large, 1 master and 4 data nodes
  - Search results, filter and plot charts of all registers in seconds

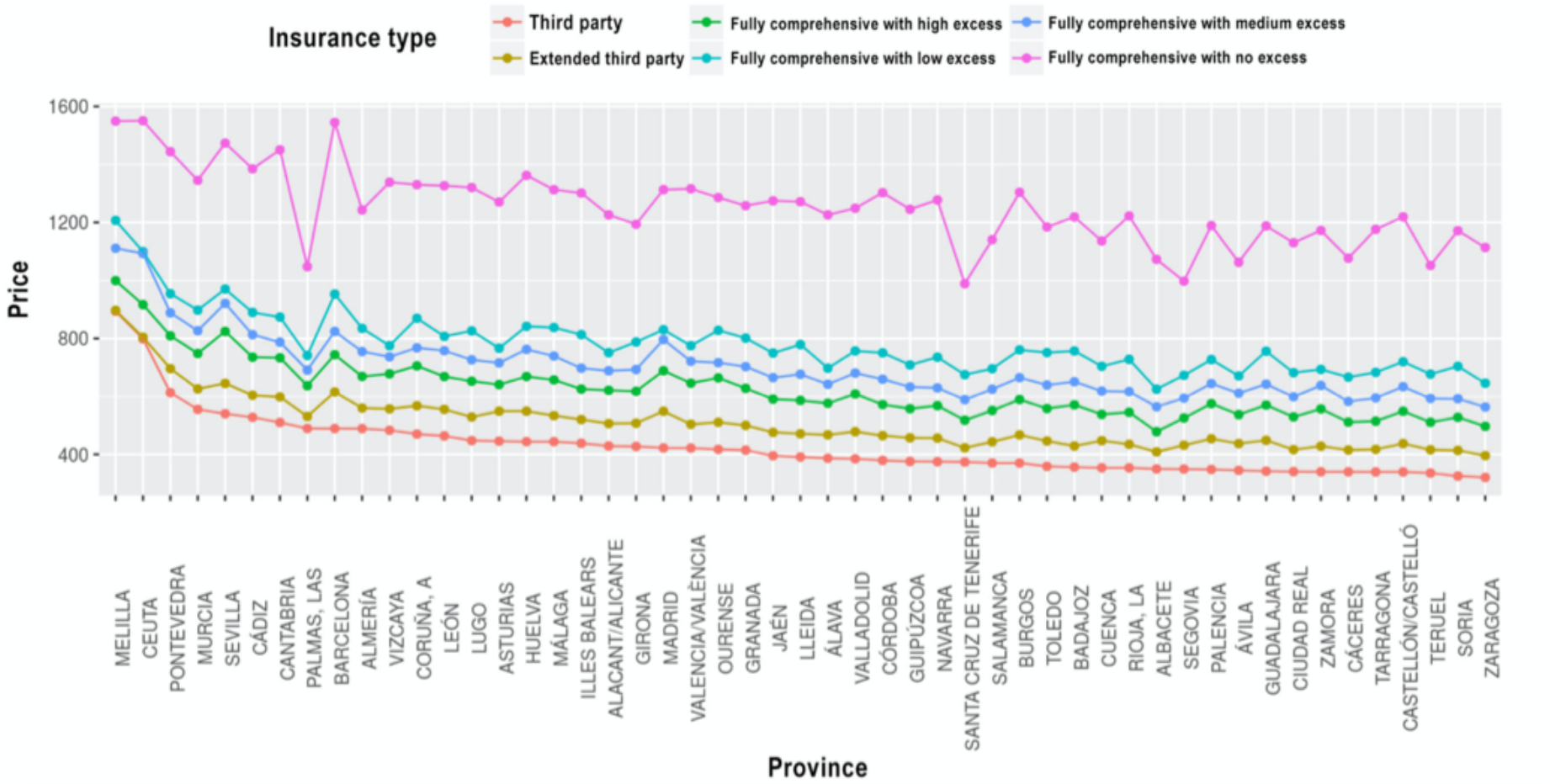
Preliminary data analysis



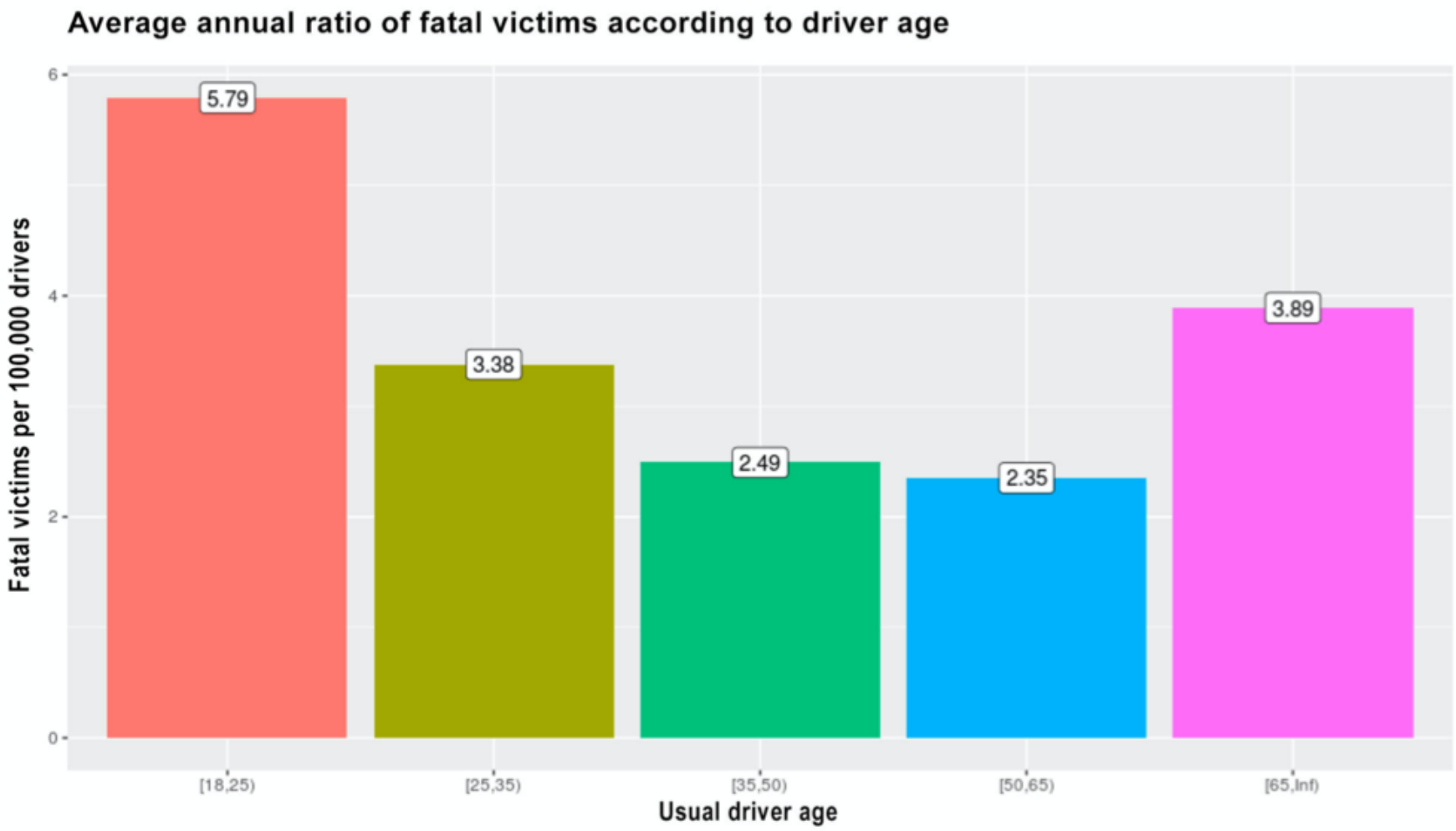
Preliminary data analysis



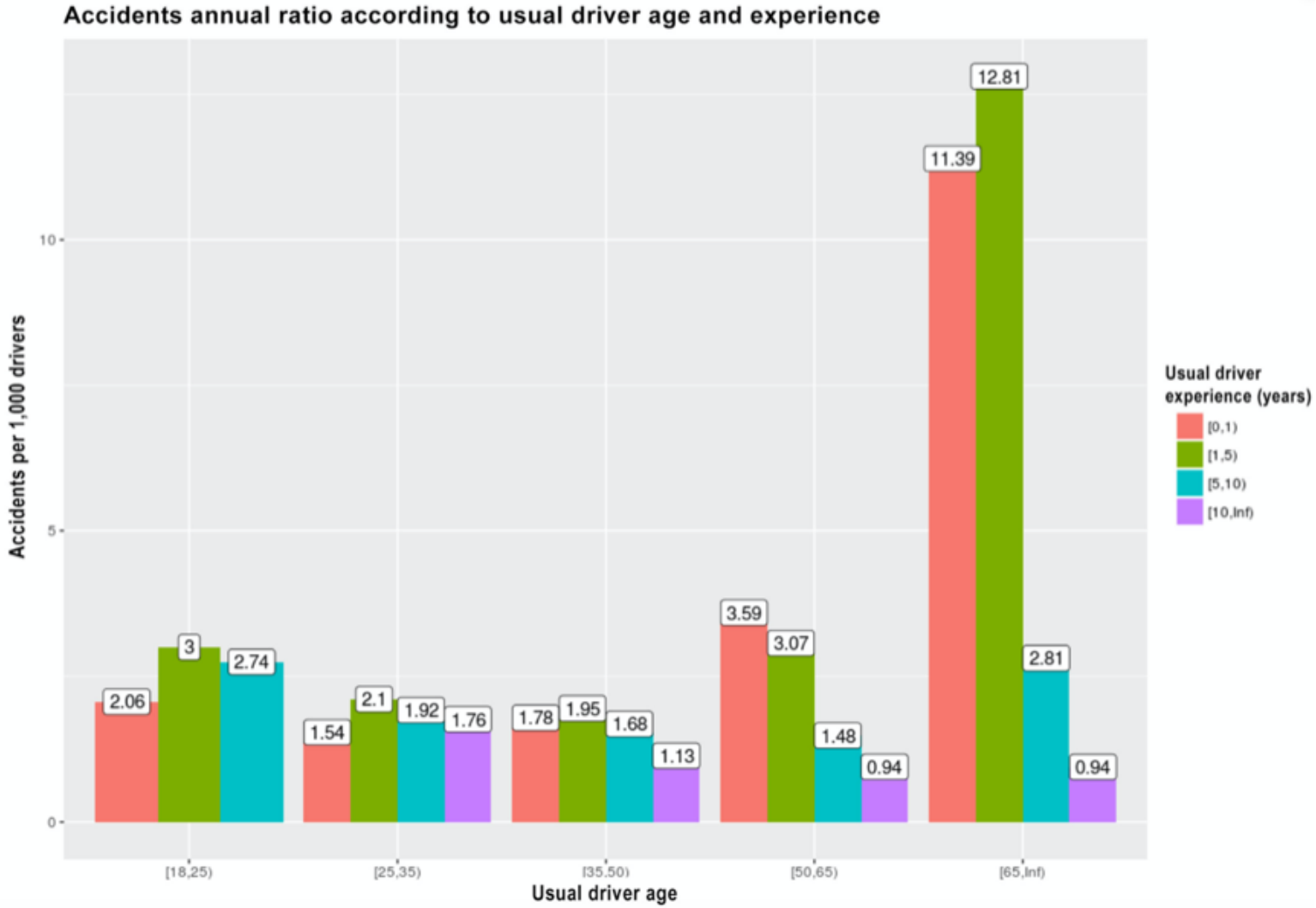
Preliminary data analysis



Preliminary data analysis



Preliminary data analysis



Modelling

- Preselection of the most significant variables:
  - Main driver age
  - Driving experience
  - Last five years accidents
  - Vehicle type
  - Vehicle power
  - Vehicle price
  - Vehicle age
- Binning -> group in ranges
  - Avoid errors

Modelling

- Adjust models adding some external variables (per province) to improve results reducing remainders
  - Number of drivers
  - Ratio of drivers between 18 and 25
  - Ratio of drivers older than 65
  - Total number of vehicles
  - Vehicles between 15 and 20 years old
  - Income per capita
  - Percent of homes with payment delays
  - Unemployment rate
  - Days per year with rain, ice or fog

Modelling

- Regression modelling
  - Using regression techniques in R suite
    - Y: price
    - X: variable
    - B: adjustment paramenters

$$\hat{y}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_n x_{n,i}$$

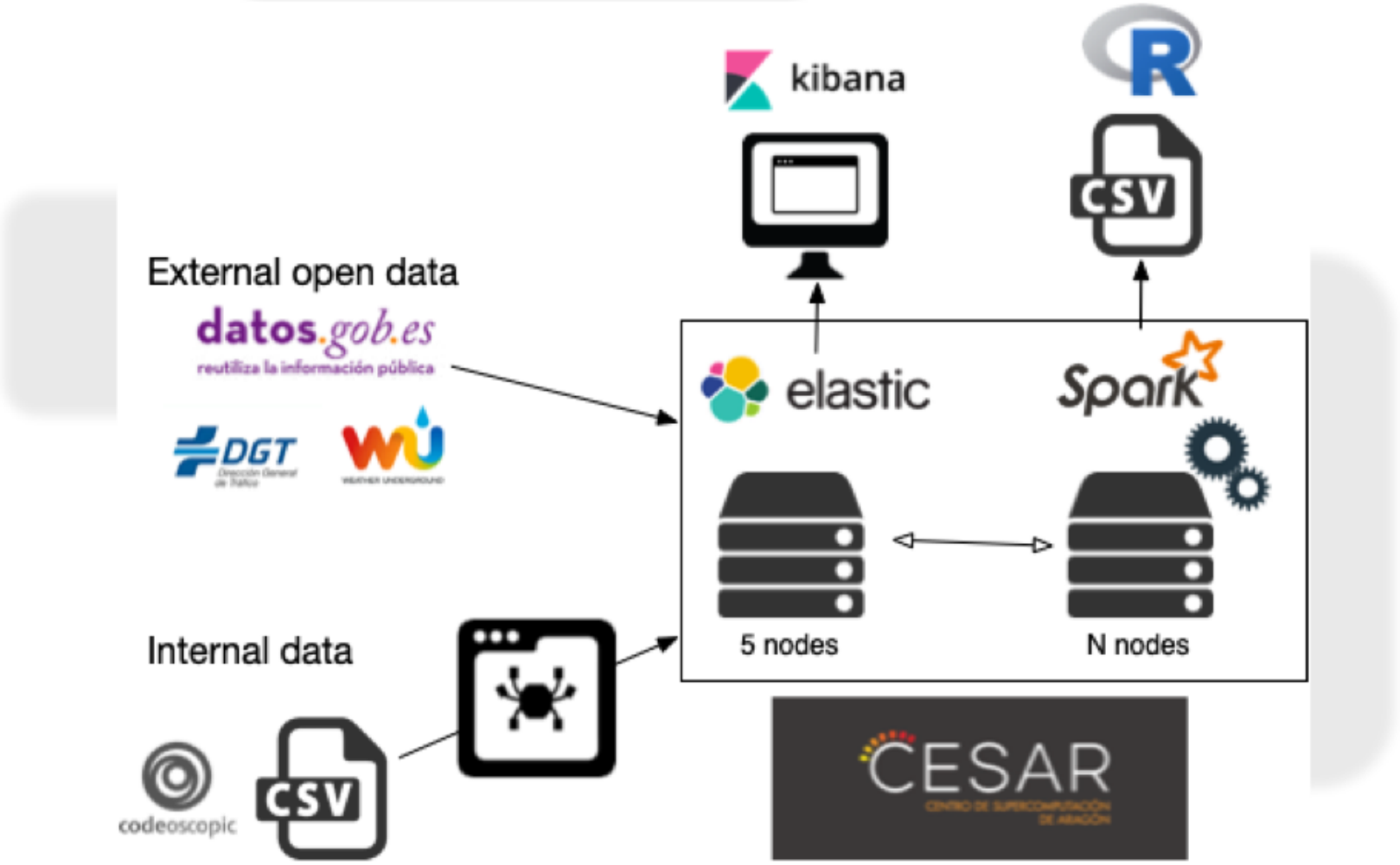
- Neural network modelling
  - Tensorflow -> open source software library for dataflow programming developed by Google Brain team
- Similar results but Tensorflow is more comfortable, very easy to integrate, part of Amazon Services

Infraestructure

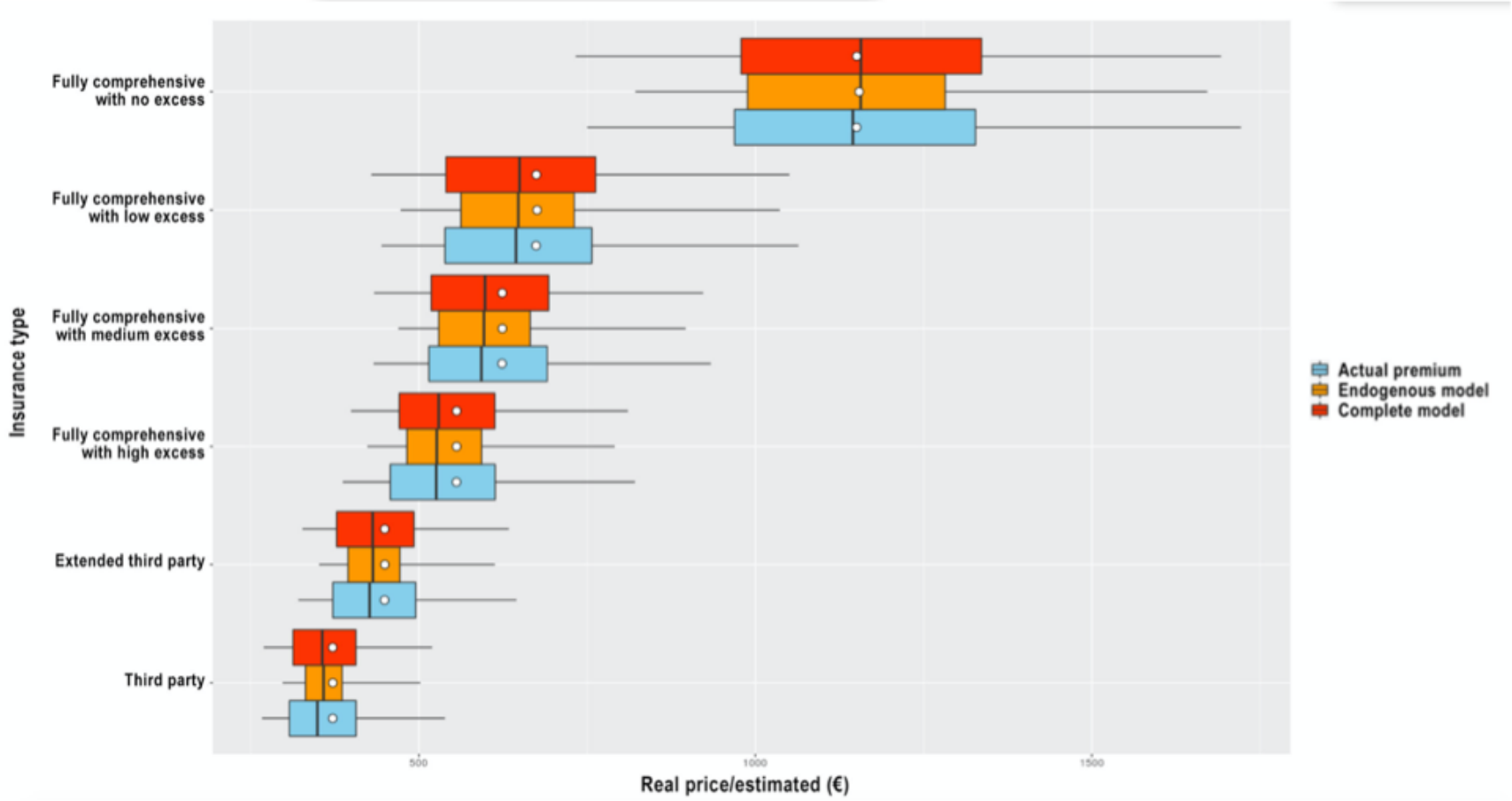
- Process data to generate models
- Regression
  - Spark: BigData platform for exploiting massive data
    - 4 nodes: 1 master and 3 slaves m1.xlarge
    - Prepare data for regression
    - Filter extreme cases and erroneous values -> they do not reflect reality
  - R: programming language and environment for statistic analysis and modelling
- Tensorflow
  - Amazon Web Services -> Sagemaker

Infraestructure

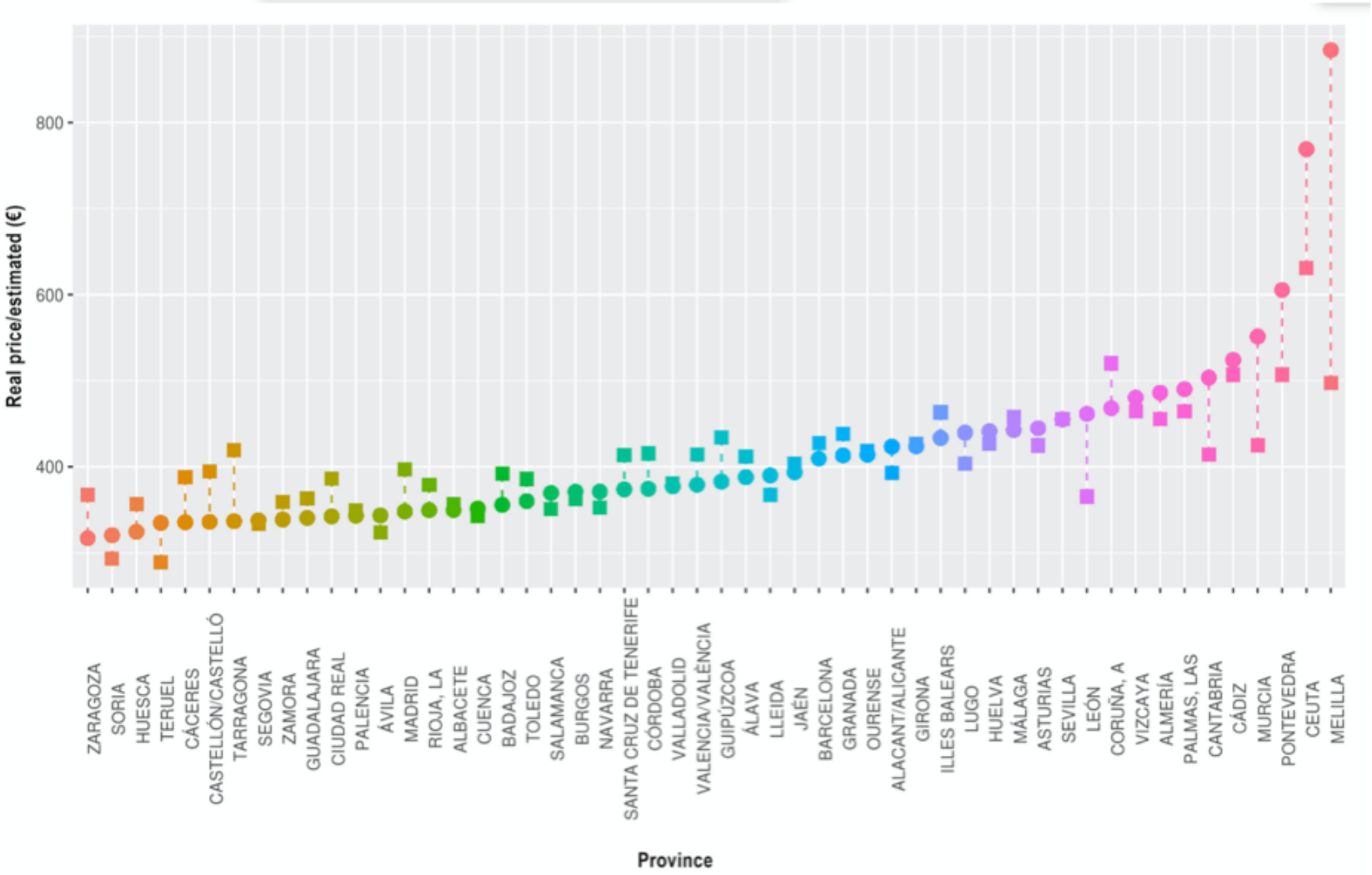
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Unión Europea
Fondo Europeo de Desarrollo Regional "Una manera de hacer Europa"
Instituto Universitario de Investigación Biocomputación y Física de Sistemas Complejos Universidad Zaragoza
codeoscopic



Results


Fully comprehensive with no excess
Fully comprehensive with low excess
Fully comprehensive with medium excess
Fully comprehensive with high excess
Extended third party
Third party
Insurance type
500
1000
1500
Real price/estimated (€)
Actual premium
Endogenous model
Complete model

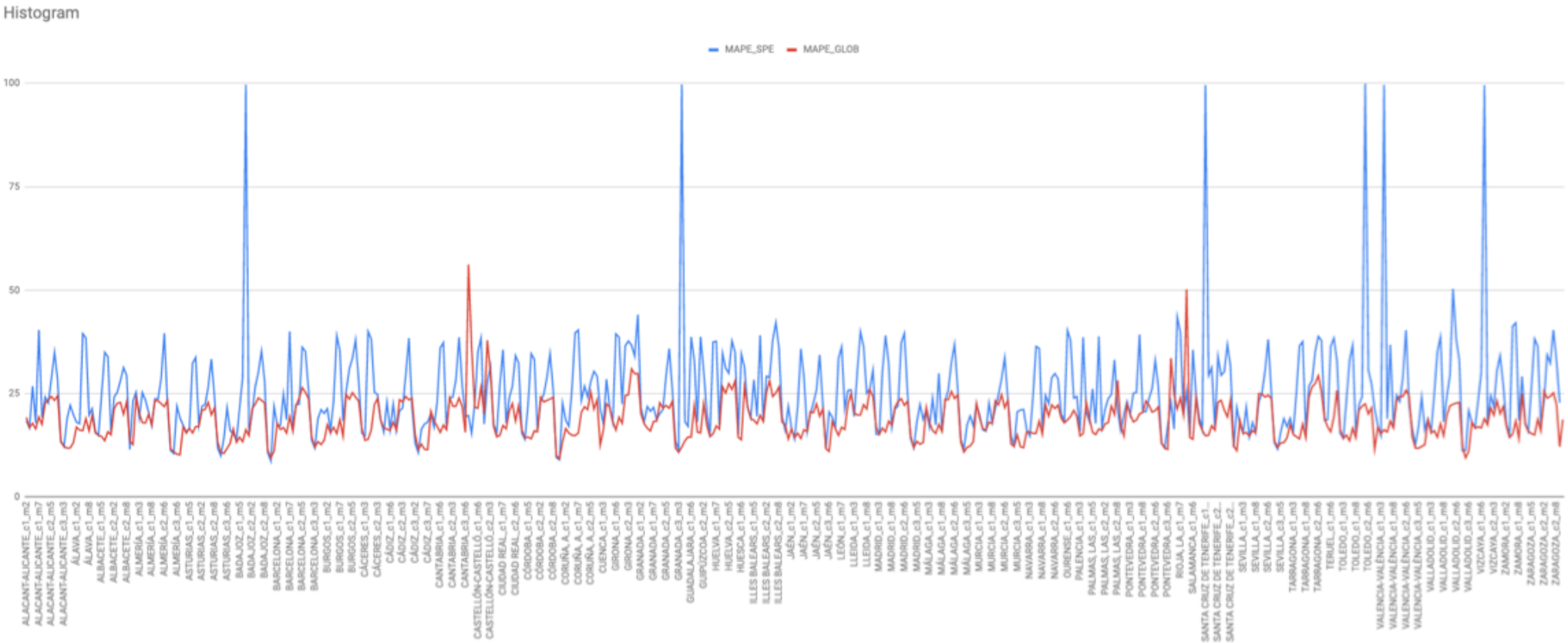
Results


Real price/estimated (€)
800
600
400
ZARAGOZA
SORIA
HUESCA
TERUEL
CÁCERES
CASTELLÓN/CASTELLÓ
TARRAGONA
SEGOVIA
ZAMORA
GUADALAJARA
CIUDAD REAL
PALENCIA
ÁVILA
MADRID
RIOJA, LA
ALBACETE
CUENCA
BADAJOZ
TOLEDO
SALAMANCA
BURGOS
NAVARRA
SANTA CRUZ DE TENERIFE
CÓRDOBA
VALLADOLID
VALENCIA/VALÈNCIA
GUIPÚZCOA
ÁLAVA
LLEIDA
JAÉN
BARCELONA
GRANADA
OURENSE
ALACANT/ALICANTE
GIRONA
ILLES BALEARS
LUGO
HUELVA
MÁLAGA
ASTURIAS
SEVILLA
LEÓN
CORUÑA, A
VIZCAYA
ALMERÍA
PALMAS, LAS
CANTABRIA
CÁDIZ
MURCIA
PONTEVEDRA
CEUTA
MELILLA
Province

Conclusions

- In general, the model works very well
- It works better with external open data
- There are some provinces in wich differences are still noticeables
  - The model is wrong: there are specifities not ”captured” by the model
  - Reality is wrong -> Risks are overrated or underrated: market would be profitable with lower prices or they should be increased
- Test using different strategies
  - One model per modality VS one model per modality and Company VS one model per modality, company and province
- While we are working in this project, data keep growing, and results improve

Conclusions



Conclusions

- Thanks to cloud infrastructures ->
  - Manage big volumes of data
  - Model data behaviour

?