Some Applications of Big Data Analytics Techniques to the Insurance Business

Franco Moriconi
University of Perugia

I. Analyzing telematics Car Driving Data

Aim: car insurance pricing (including Usage-Based Insurance, UBI and Pay-As-You-Drive, PAYD)

Typical information (at individual driver level): location (GPS), speed, acceleration/braking, left-/right-turns, number of trips, total distance, total duration, daytime of trips, ...

? *How these high-dimensional and high-frequency data can be converted into useful covariate information?*

$\longrightarrow$

Classification of different driving styles, based on *unsupervised learning* (cluster analysis, pattern recognition)

As a first exercise we present:
$\oplus$ Wüthrich M.V. (2017). *Covariate selection from telematics car driving data.* European Actuarial Journal 7(1):89-108

For using Fourier analysis for pattern recognition see also:

⊕ Weidner W., Transchel F.W.G., Weidner R. (2016). *Classification of scale-sensitive telematics observables for riskindividual pricing.* European Actuarial Journal 6(1):3-24

Data

- telematics data from 1753 individual car drivers
- For each car driver 200 individual trips are recorded (GPS locations $(x_t, y_t)_{t\geq 0}$ second by second)

⊙ No information about daytime, type of car, etc. No information about incurred claims

! Since we have no car accident information, supervised learning methods cannot be applied!

Speed & acceleration

For each time interval $(t-1, t]$ one computes the *average speed*:

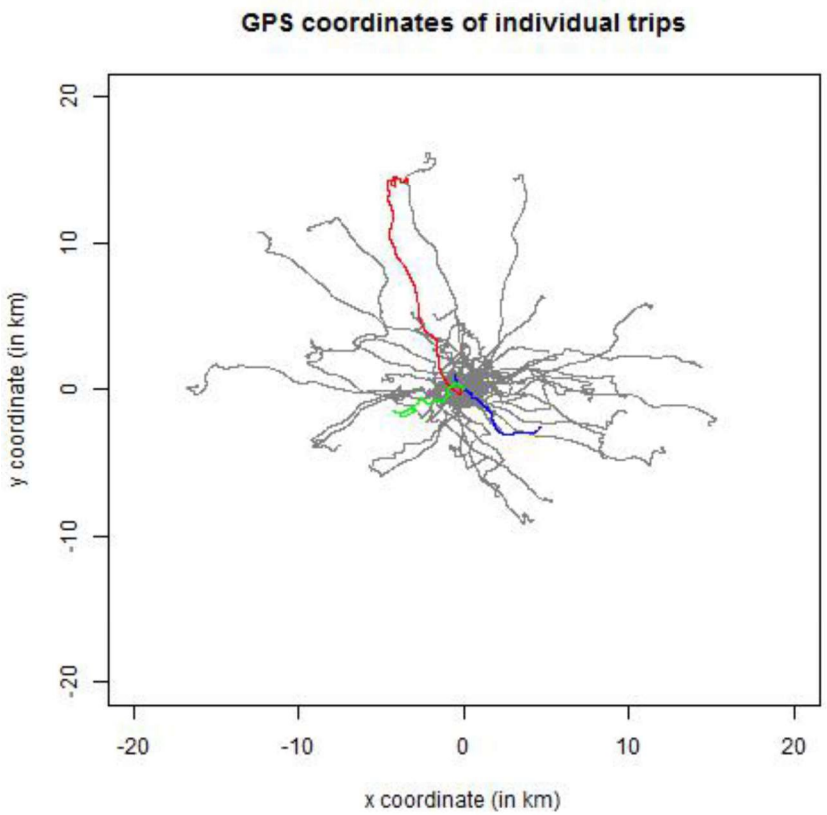
$$v_t = \frac{\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}}{t - (t-1)}, \quad t \geq 1\,,$$

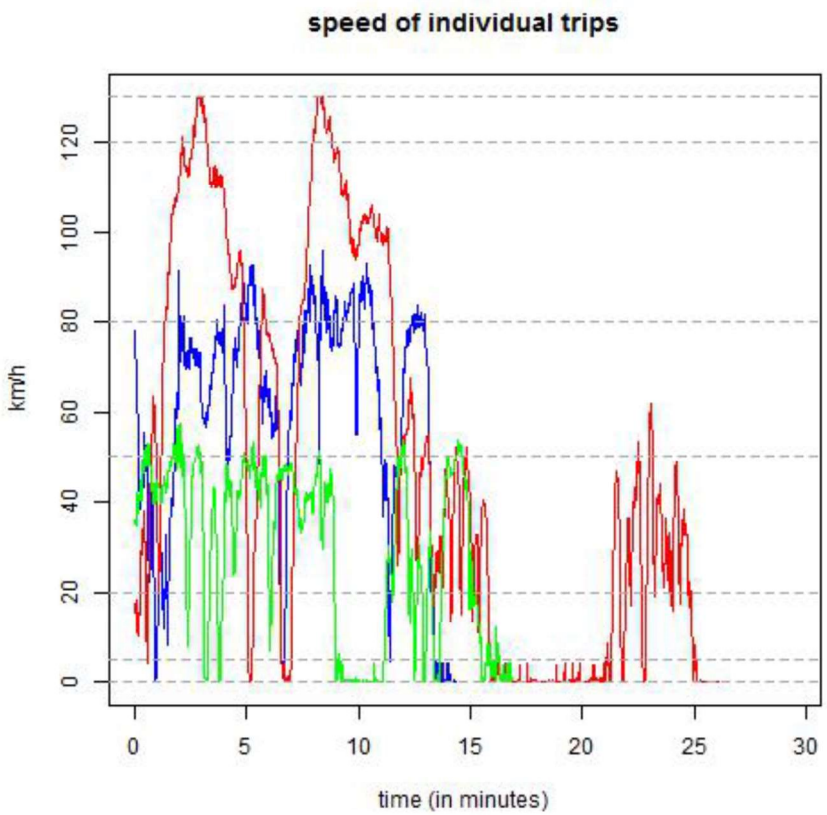
and the *average acceleration/bracking* (more precisely, the *change in average speed*):

$$a_t = \frac{v_t - v_{t-1}}{t - (t-1)}, \quad t \geq 2\,.$$

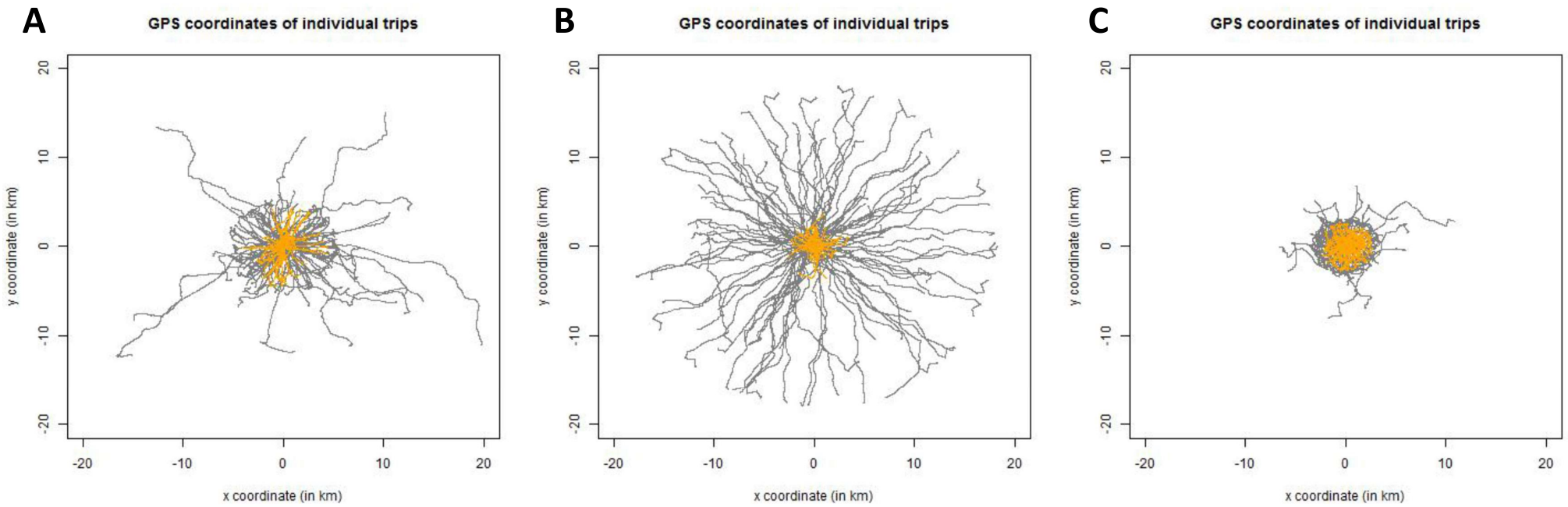
Example. 200 individual trips of the same car driver

(For confidentiality reasons all trips are initialized at location $(0, 0)$, randomly rotated and reflected at the cohordinate axes)





Example. 200 individual trips of three different car drivers



Driving habits $\longrightarrow$

Simple empirical statistics: Total distance, Average distance per trip, Total time, Avg. time per trip, Avg. speed, ...

Speed buckets (in km/h)

$[0]$	car stands still
$(0, 5]$	acceleration or braking phase (from/to speed 0)
$(5, 20]$	low speeds
$(20, 50]$	urban area speeds
$(50, 80]$	rural area speeds
$(80, 130]$	highway speeds (truncated above 130 km/h)

$\longrightarrow$ speed bucket distributions (in driving time)

Analizing driving styles

v-a heatmaps

x-axis: average speed (in km/h)

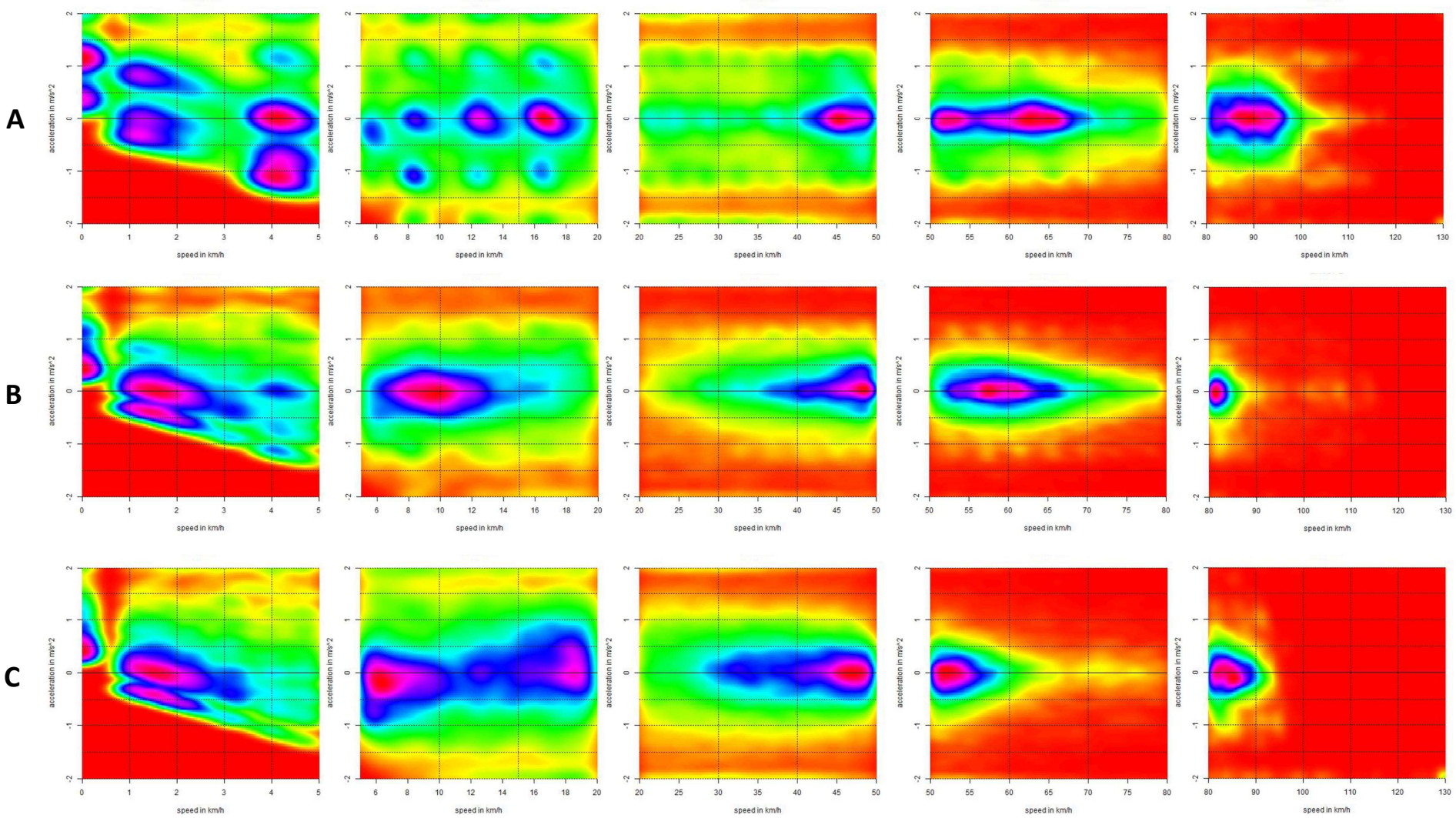
y-axis: corresponding acceleration/braking pattern (in m/s^2)

v-a heatmaps by speed bucket: in each speed bucket, the resulting empirical density (normalized to 1)

Example for the speed bucket $(5, 20]$:

For each driver, all observations (v_t, a_t) with $v_t \in (5, 20]$ are collected. By normalizing by the number of observations falling into the speed bucket, a two-dimensional (discrete) distribution on the rectangle $R = (5, 20] \times (-2, 2]$ is obtained

v-a heatmaps for the car drivers A, B, C



Classifying the v-a heatmaps

Goal: identify the car drivers that have similar v-a heatmaps

Similar v-a heatmaps $\Longrightarrow$ similar driving styles $\Longrightarrow$ same categorical class for insurance pricing (categorical covariates)

Non supervised learning tools

E.g. Cluster analysis methods

- define a *dissimilarity measure*
- use the *K-means Clustering Algorithm* to allocate different driving styles to different categories

  For a given number K of categorical classes (clusters), find the classifier that minimizes the aggregate within-cluster dissimilarity measure.

Subsequent papers

$\oplus$ Gao G., Wüthrich M.V. (2017). *Feature extraction from telematics car driving heatmaps*. To appear in European Actuarial Journal.

$\longrightarrow$

Two-dimension reduction techniques applied to v-a heatmaps: principal components analysis (via singular value decomposition) and bottleneck neural networks.
Both techniques serve to represent the high dimensional v-a heatmaps by two continuous covariates. The bottleneck neural network can also be seen as a non-linear PCA

$\oplus$ Gao G., Meng S., Wüthrich M.V. (2018). *Claims frequency modeling using telematics car driving data*. To appear in Scandinavian Actuarial Journal.

$\longrightarrow$

Predictive power of covariates from telematics data is investigated (having additional information about the car and driver and having insurance claims data). Claim frequency modelled using a Poisson regression model

II. Individual Claims Reserving Using CARTs
(Classification And Regression Trees)

In non-life insurance the ***claims reserve*** R is the money amount the insurer should hold in order to pay all open claims, i.e. claims currently incurred but not yet settled.

Typical figures for relevant Lines of Business (e.g. R.C. Auto): Reserve $\sim 10^8 \div 10^9$ €, # open claims $\sim 10^5 \div 10^6$

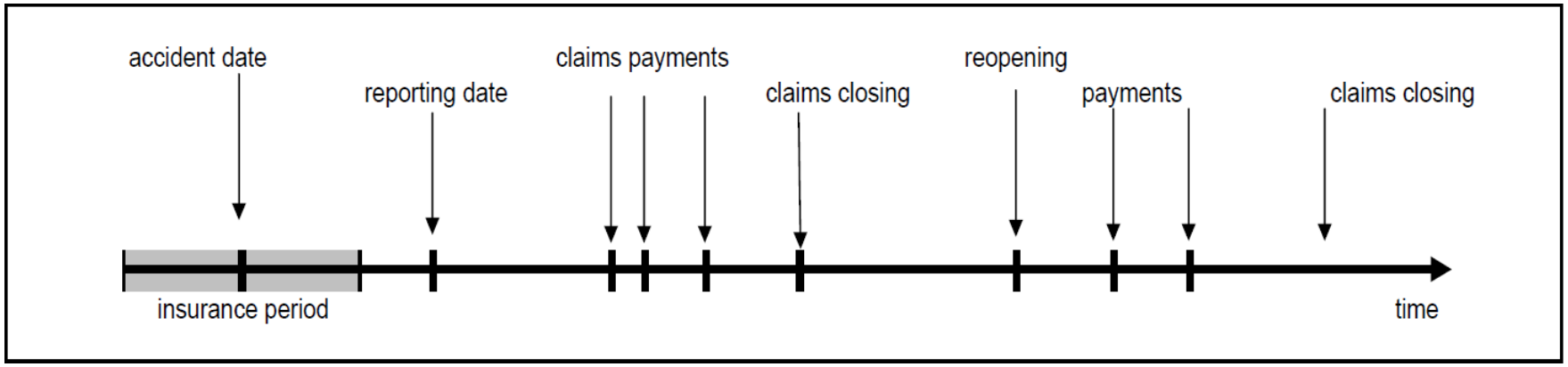
The *claims development process*, from the accident date to the settlement date, can take many years.

At any date during the cost development an *ultimate cost* U should be estimated and the corresponding reserve should be obtained as:

$$R = U - P$$

where P is the amount paid until the current date.

Typical development of a non-life insurance claim



- Other relevant events: lawyer involved, ...
- Additional complexity for Italian *Motor Third Party Liability* (R.C. Auto): *Convenzione tra Assicuratori per il Risarcimento Diretto* $\longrightarrow$ Card/NoCard/...

Traditional Claims Reserving

An individual reserve, the **case reserve**, is associated to each claim, at least at beginning of development.

However, usually the claims reserve (or the claims ultimate cost) is estimated using **aggregated data**:

- claims payments are organized by accident year $i = 1, 2, \ldots, I$ and development year $j = 1, 2, \ldots, J \leq I$
- matrix representation: $C_{i,j}$ denotes the aggregated payments in cell (i, j), i.e. the sum of payments made in accident year i and development year j for all open claims
- observational data: the top-left triangle of the $C_{i,j}$ currently observed is computed
- prediction: the lower-right triangle of future $C_{i,j}$ is estimated using an appropriate statistical model

Multidimensional time serie forecasting $\longrightarrow$ claims reserving models (e.g. chain-ladder)

Individual Claims Reserving

Using machine learning techniques (e.g. CARTs) it is possible to:

- estimate R for each individual claim
- forecast relevant events in the individual cost development process $\longrightarrow$ **Claim Watching**:
  - $\star$ *for which claims the probability that a lawyer will be involved in one year is greater than* $\alpha\%$ *?*
  - $\star$ *for which claims the estimated ultimate cost differs from the case reserve for more than* $\beta\%$*?*

$\longrightarrow$ An example on Italian Motor Third Party Liability (R.C. Auto)

Basic references:

$\oplus$ Wüthrich, M.V. (2016). *Machine learning in individual claims reserving.* SSRN Manuscript ID 2867897.

$\oplus$ D'Agostino L. *et al.* (2018). *Machine learning per la riserva sinistri individuale. Un'applicazione R.C. Auto degli alberi di classificazione e regressione.* Alef Technical Report – 18/02 – Roma.

An example on Italian Motor Third Party Liability (R.C. Auto)

Data. historical information on the development process at individual claim level, with maximum time length and maximum details (also qualitative)

At time t, claims data are organized by *accident year* i, *reporting delay* j and *time lag* $\ell = t - i$

General model. A regression model is defined (and estimated) for each time lag ℓ:

$$\mathbf{E}\left[W_{t+1}|\mathcal{F}_t\right] = \mu_\ell(\boldsymbol{x}_t)$$

- W_{t+1}: response variable at time $t+1$ (to be predicted at time t)
- $\mathcal{F}_t$: information available at time t
- $\boldsymbol{x}_t$: vector of covariates (explicative variables, "feature") at time t
- μ_ℓ: regression function (depending only on ℓ)

CART model. In general the regression function μ_ℓ is fully non-parametric and can have any form. We calibrate μ_ℓ using classification and regression trees.

Since a response variable is used, we perform supervised learning

A *frequency-severity* model is considered (compound model):

Frequency model: a model for the events

Severity model: a model for the expected payments, conditionally on the payment is made

- Frequency model: qualitative variables involved → *classification tree approach*
  - response variable - 4 events: *ci sarà un pagamento Card?, ci sarà un pagamento NoCard?, il sinistro andrà "in causa"?, verrà chiuso?* → $2^4 = 16$ possible outcomes
  - explanatory variables - *c'è/non c'è una riserva Card, c'è/non c'è una riserva NoCard, c'è stato/non c'è stato un pagamento Card, c'è stato/non c'è stato un pagamento NoCard, è/non è aperto, è/non è in causa, ritardo di denuncia, ...*
- Severity model: only quantitative variables → *regression tree approach*
  - response variables - *se ci sarà un pagamento Card, quale importo atteso?, se ci sarà un pagamento NoCard, quale importo atteso?*
  - explanatory variables - *valore della* case reserve *appostata per Card e per NoCard, valore del pagato cumulato per Card e per NoCard, il sinistro è/non è in causa, ritardo di denuncia, ...*

Estimating the regression function

A non-parametric estimate of the regression function μ_ℓ is obtained by CARTs

The estimated function $\widehat{\mu}_\ell$ is piecewise constant on an appropriate partition of the feature space:

$$\mathcal{P}_\ell := \left\{ \mathcal{R}_\ell^{(1)}, \ldots, \mathcal{R}_\ell^{(n_\ell)} \right\}$$

That is there exist n_ℓ constants $\overline{\mu}_\ell^{(r)}$ such that:

$$\widehat{\mu}_\ell \left(\boldsymbol{x}_t \right) = \sum_{r=1}^{n_\ell} \overline{\mu}_\ell^{(r)} \, \mathbf{1}_{\left\{ \boldsymbol{x}_t \in \mathcal{R}_\ell^{(r)} \right\}}$$

CART algorithm collects in the same (hyper)rectangle $\mathcal{R}_\ell^{(r)}$ observations which are in some sense similar between each other. The rectangles are the *explanatory classes*

Binary split tree growing algorithm

- The feature space is successively partitioned into rectangles by solving *standardized binary split* questions
- Each binary split is obtained by minimizing a fixed *impurity measure* (typically, Gini index for classification trees, sum of squared errors for regression trees)
- In a first step a large tree is grown
- In a second step the appropriate tree size is determined by *cross-validation* (K-fold validation) and the initially large binary tree is "pruned" to that size: *cost-complexity pruning* (1-SD rule used)
- The rectangles $\mathcal{R}_\ell^{(r)}$ are the "leaves" of the classification/regression tree.

Implemented in R as `rpart` (with `method='class'` or `method='anova'`)

References:

⊕ Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and Regression Trees.* Wadsworth Statistics/Probability Series.

⊕ Hastie T., Tibshirani R., Friedman J. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer Series in Statistics.

Example of response variable (for the frequency model)

To apply `rpart` we need to make the response one-dimensional:

$$W_{t+1} := YNC_{t+1} + 2 \cdot YCA_{t+1} + 4 \cdot Z_{t+1} + 8 \cdot L_{t+1}$$

defined by the event indicators (0/1):

- YNC_{t+1}: "c'è un pagamento No Card";
- YCA_{t+1}: "c'è un pagamento Card";
- Z_{t+1}: "il sinistro è chiuso";
- L_{t+1}: "il sinistro è in causa".

Our aim is obtaining a *probability distribution* for the 16-valued random variable W_{t+1} based on the vector $\boldsymbol{x}_t$ of covariates observed until time t

Explicit representation of the response variable W

YNC	YCA	Z	L	W	stato
0	0	0	0	0	ASP0: aperto senza pagamenti e non in causa
1	0	0	0	1	APN0: aperto con pagamento No Card e non in causa
0	1	0	0	2	APC0: aperto con pagamento Card e non in causa
1	1	0	0	3	APP0: aperto con pagamento Card e No Card e non in causa
0	0	1	0	4	CSS0: chiuso senza seguito e non in causa
1	0	1	0	5	CPN0: chiuso con pagamento No Card e non in causa
0	1	1	0	6	CPC0: chiuso con pagamento Card e non in causa
1	1	1	0	7	CPP0: chiuso con pagamento Card e No Card e non in causa
0	0	0	1	8	ASPL: aperto senza pagamenti e in causa
1	0	0	1	9	APNL: aperto con pagamento No Card e in causa
0	1	0	1	10	APCL: aperto con pagamento Card e in causa
1	1	0	1	11	APPL: aperto con pagamento Card e No Card e in causa
0	0	1	1	12	CSSL: chiuso senza seguito e in causa
1	0	1	1	13	CPNL: chiuso con pagamento No Card e in causa
0	1	1	1	14	CPCL: chiuso con pagamento Card e in causa
1	1	1	1	15	CPPL: chiuso con pagamento Card e No Card e in causa

Example from rpart

All observed claims with lag $\ell = 0$ are considered (284.336 claims). Pruned tree:

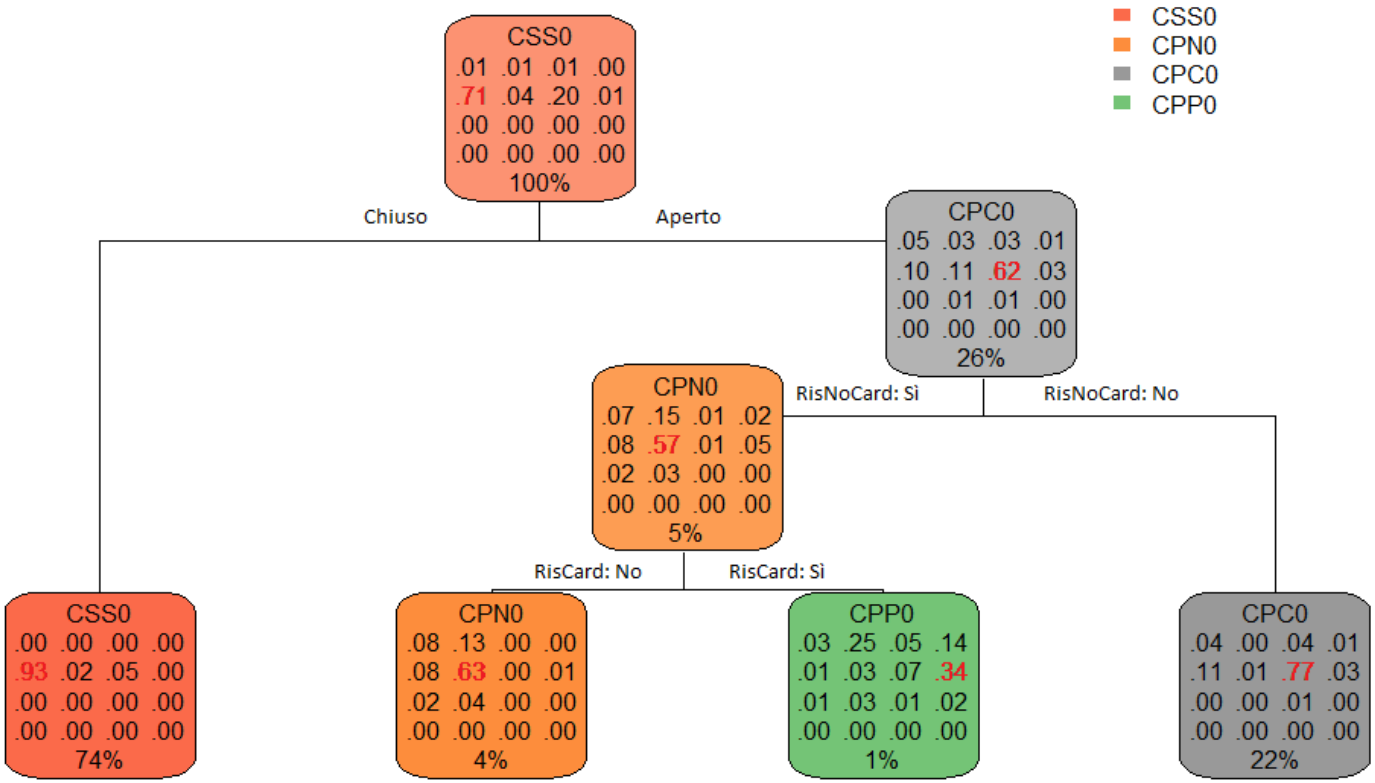


Figura 1 – Classification tree for lag $\ell = 0$.

The **rpart** algorithm indicates that the best prediction is obtained by partitioning the set of all claims with $\ell = 0$ in 4 explanatory classes. For each class a probability distribution of W is estimated

nodo	nobs	caratteristiche della foglia	CM	distribuzione di probabilità della risposta							
1	284.336	sinistri omogenei	CSS0	ASP0	APN0	APC0	APP0	CSS0	CPN0	CPC0	CPP0
				0.014	0.0087	0.0084	0.0021	0.71	0.042	0.2	0.0081
				ASPL	APNL	APCL	APP0	CSSL	CPNL	CPCL	CPPL
				0.002	0.0021	0.0017	0.00026	0	0.00019	0.00021	0.00004
2	209.046	sinistri chiusi	CSS0	ASP0	APN0	APC0	APP0	CSS0	CPN0	CPC0	CPP0
				0.024	0.001	0.00076	0.00006	0.93	0.017	0.046	0.00035
				ASPL	APNL	APCL	APP0	CSSL	CPNL	CPCL	CPPL
				0.00094	0.00037	0.00032	0.00001	0	0.00004	0.00004	0
12	12.659	sinistri aperti solo per la partita di danno NoCard	CPN0	ASP0	APN0	APC0	APP0	CSS0	CPN0	CPC0	CPP0
				0.076	0.13	0.0011	0.0044	0.083	0.63	0.0023	0.011
				ASPL	APNL	APCL	APP0	CSSL	CPNL	CPCL	CPPL
				0.017	0.035	0.00024	0.00063	0	0.0035	0	0.00016
13	1.472	sinistri aperti per entrambe le partite di danno	CPP0	ASP0	APN0	APC0	APP0	CSS0	CPN0	CPC0	CPP0
				0.034	0.25	0.052	0.14	0.0054	0.032	0.071	0.34
				ASPL	APNL	APCL	APP0	CSSL	CPNL	CPCL	CPPL
				0.013	0.028	0.0088	0.021	0	0	0	0.0014
7	61.159	sinistri aperti ma non per la partita di danno NoCard	CPC0	ASP0	APN0	APC0	APP0	CSS0	CPN0	CPC0	CPP0
				0.041	0.0035	0.035	0.0052	0.11	0.0052	0.77	0.026
				ASPL	APNL	APCL	APP0	CSSL	CPNL	CPCL	CPPL
				0.0023	0.00036	0.0063	0.00054	0	0.00002	0.00085	0.00013

Tabella 1 – Distribuzioni di probabilità stimate dall'algoritmo `rpart`. In rosso sono indicati lo stato modale e la sua probabilità

Subsequent steps

To complete the model:

- Multiperiod forecasting
  The frequency model must be extended to multiperiod predictions. A simulation approach is used
- Application of the severity model
  Conditional prediction of payd amounts is performed by regression trees

Example. Cost development for a single claim

Let us consider a specified individual claim, having the following feature:

- $i = I$, sinistro accaduto nell'anno I (l'anno corrente);
- $j = 0$, sinistro con reporting delay nullo, cioè l'anno di denuncia coincide con quello di accadimento;
- $Z0 = 0$, sinistro aperto;
- $L0 = 0$, sinistro non in causa;
- $YCA0 = 0$, sinistro senza pagamenti per la partita di danno Card nell'anno I;
- $YNC0 = 0$, sinistro senza pagamenti per la partita di danno No Card nell'anno I;
- $RCA0 = 1$, sinistro per cui è stata appostata riserva Card nell'anno I;
- $caseRCA0 = 5.000$, è stata appostata riserva Card di 5.000 € nell'anno I;
- $RNC0 = 1$, sinistro per cui è stata appostata riserva No Card nell'anno I;
- $caseNC0 = 55.000$, è stata appostata riserva No Card di 55.000 € nell'anno I.

Simulated sample paths for “partita di danno Card”

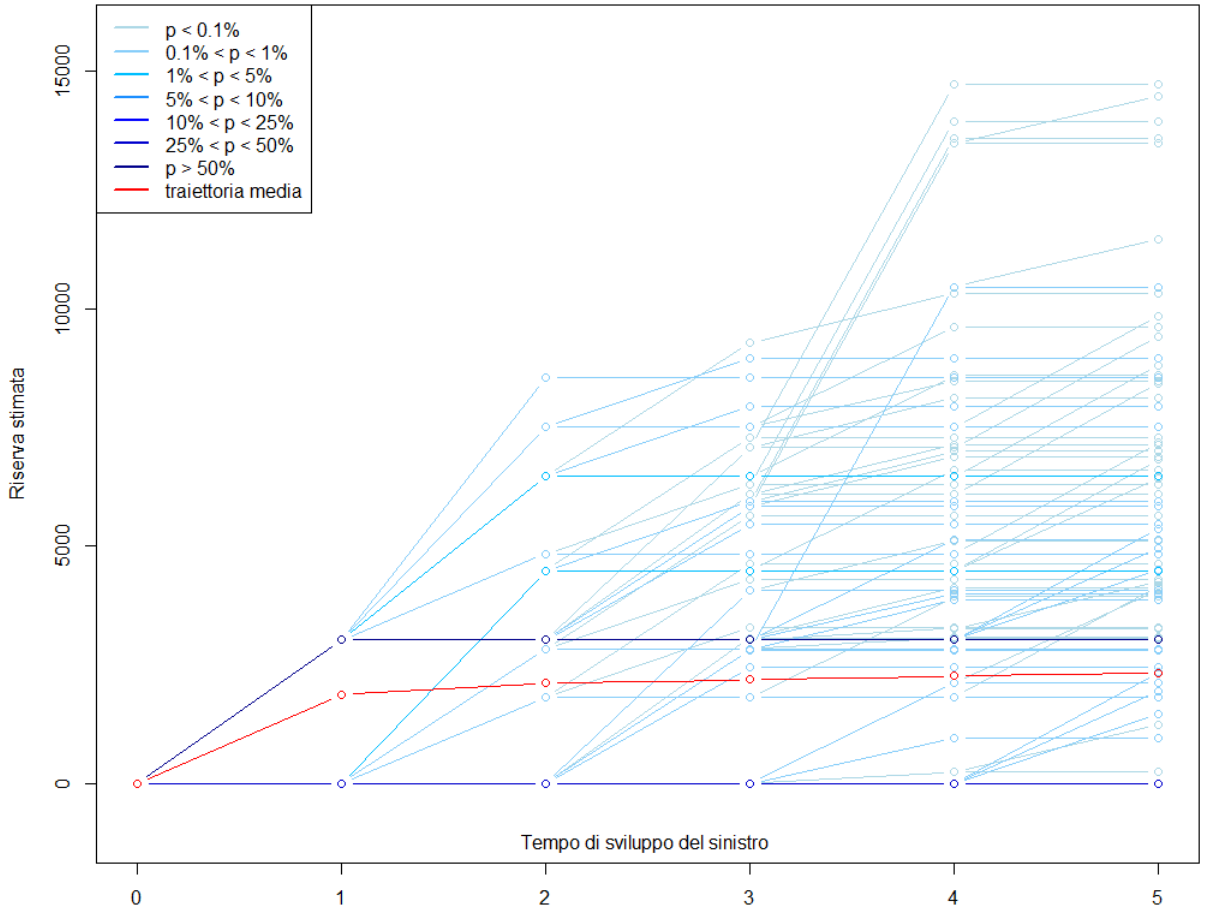


Figura 2 – Traiettorie del costo della partita di danno Card. Il modello suggerisce di rivedere la riserva Card da 5.000 € a 2.322 €

Simulated sample paths for “partita di danno No Card”

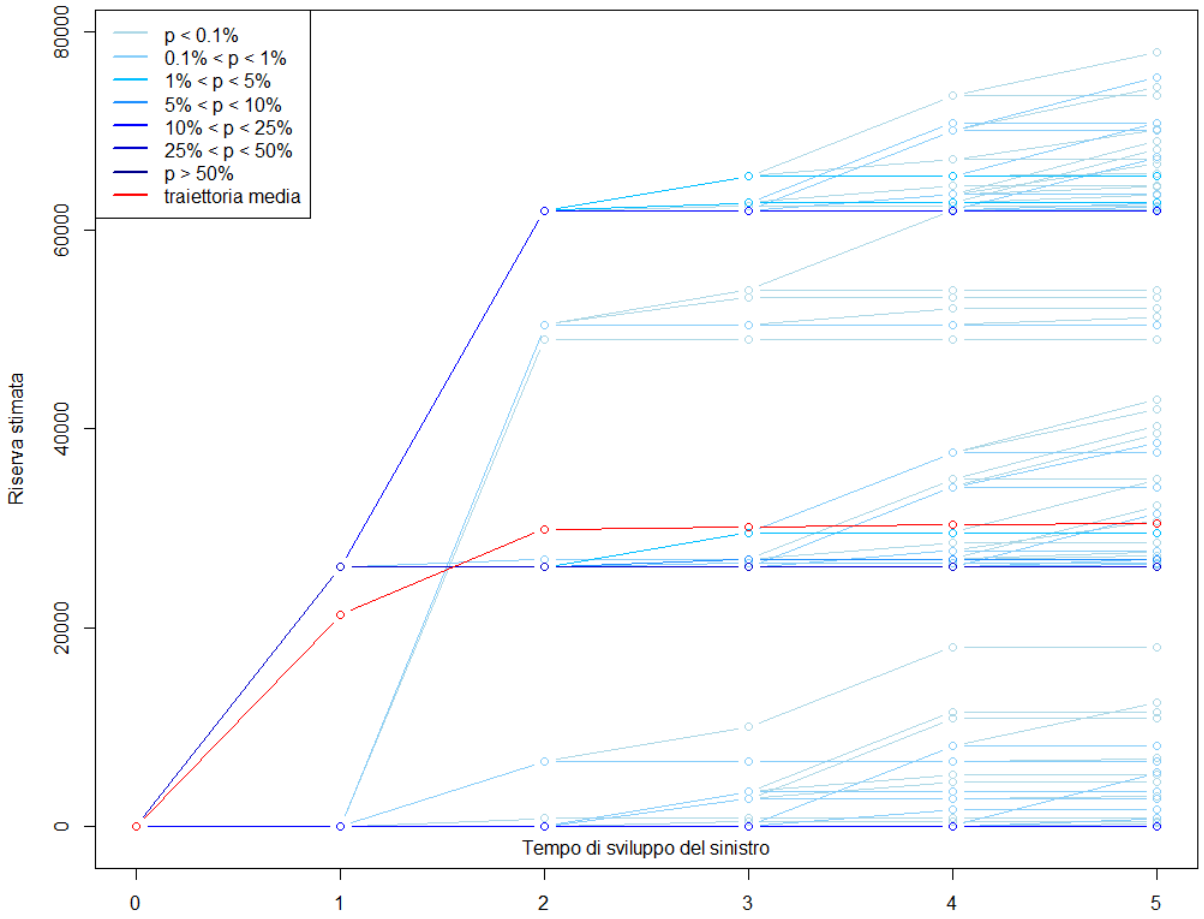


Figura 3 – Traiettorie del costo della partita di danno No Card. Il modello suggerisce di rivedere la riserva Card da 55.000 € a 30.468 €

Simulated sample paths for the total claim cost

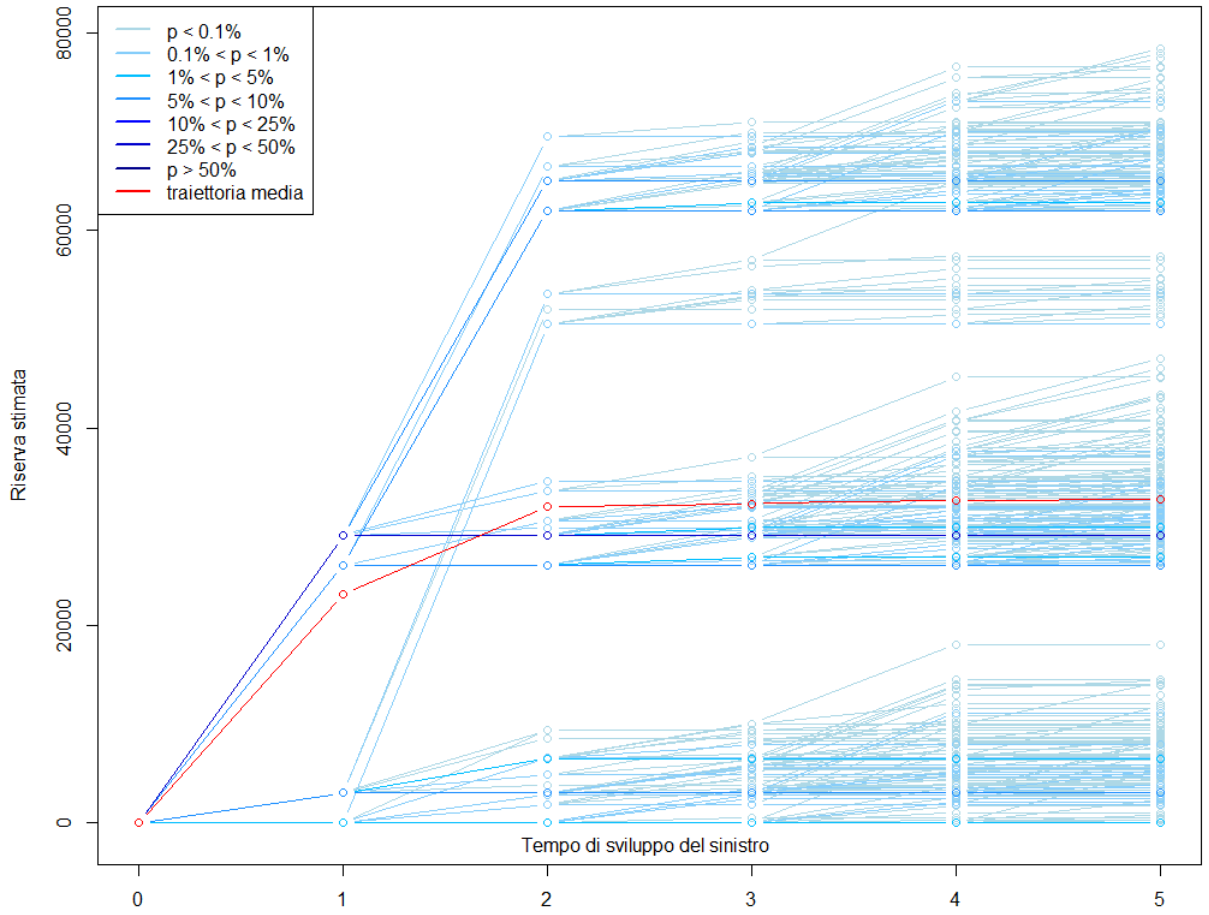


Figura 4 – Ttraiettorie del costo totale del sinistro. Il modello suggerisce di rivedere la riserva complessiva da 60.000 € a 32.790 €

III. Other machine learning applications specific to insurance

- ⋆ Extention/susbstitution of the *Generalized Linear Model* approaches to non-life insurance pricing
- ⋆ Refining traditional claims reserving approaches for heterogeneity and individual claims feature information using neural networks
- ⋆ Covariate selection for life insurance policies redemption (how policyholder behaviour depends on financial market variables)
- ⋆ In financial risk management (internal models), the probability distribution of the future market price of an asset-liability portfolio is required
  $\longrightarrow$ avoid heavy nested Monte Carlo simulations using Support Vector Machines for calibrating a market-dependent price function
- ⋆ ...