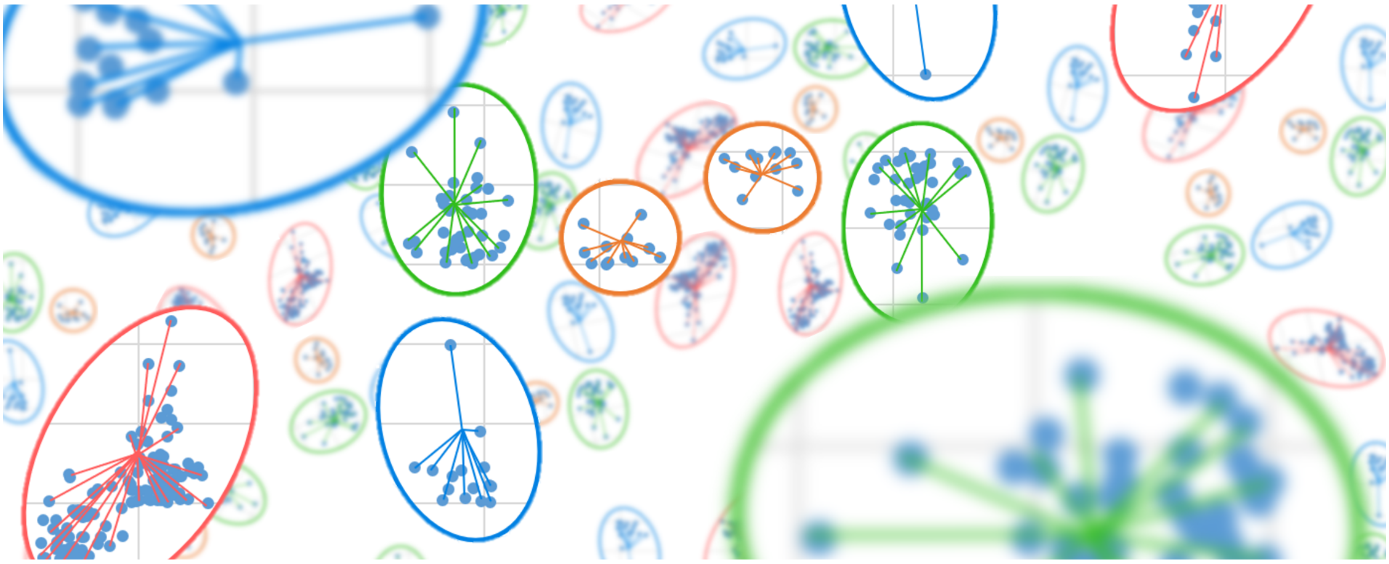
MILLIMAN WHITE PAPER

Research on non-life pricing procedures on encrypted and anonymous data under the GDPR

Thomas Poinsignon, IA
Antoine Ly, IA



The recent increases of data generated, stored and analysed by insurers to establish their pricing and underwriting policies has led to the emergence of new needs—both from a regulatory point of view, with the recent implementation of the EU General Data Protection Regulation (*GDPR*), and with a view to offering new services on the market (e.g., protection against *cyber risk*).

The work carried out in this paper is thus devoted to the development and analysis of actuarial methods within the *default security* framework—a principle of the GDPR imposed on companies using personal data.

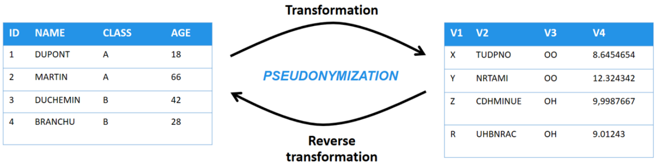
The objective is to extend the elementary mathematical concepts and models used when developing classical non-life insurance pricing models (simple linear regression and generalised linear models) to their use on secure data in accordance with regulatory requirements.

Anonymisation and pseudonymisation concepts at the core of privacy

The GDPR sets the practices insurers must respect with regard, among other things, to the data they have in their possession. In particular, we observe the *'principles of data protection from the design stage* and *security by default'*[1], which aim to clarify and formalise the constraints introduced, defining the concepts of *anonymised* and *pseudonymised* data.

'Pseudonymisation' consists of making data partially anonymous. It may be difficult but it can still be traced back and attributed to an individual (Figure 1).

FIGURE 1: DIAGRAM OF PSEUDOYMISATION PROCESS

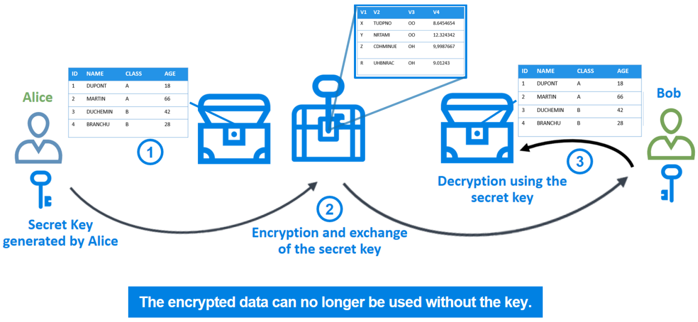


ID	NAME	CLASS	AGE
1	DUPONT	A	18
2	MARTIN	A	66
3	DUCHEMIN	B	42
4	BRANCHU	B	28

V1	V2	V3	V4
X	TUDPNO	OO	8.6454654
Y	NRTAMI	OO	12.324342
Z	CDHMINUE	OH	9,9987667
R	UHBNRAC	OH	9.01243

To achieve this, several methods exist, including secret key encryption methods (shown in Figure 2). We examined[2] the possibility of performing a simple linear regression on pseudonymised data by encryption, without ever having to decipher it (see Figure 3).

FIGURE 2: SECRET KEY ENCRYPTION PROCEDURE


Alice
1
Secret Key generated by Alice
2
Encryption and exchange of the secret key
3
Decryption using the secret key
Bob
The encrypted data can no longer be used without the key.

ID	NAME	CLASS	AGE
1	DUPONT	A	18
2	MARTIN	A	66
3	DUCHEMIN	B	42
4	BRANCHU	B	28

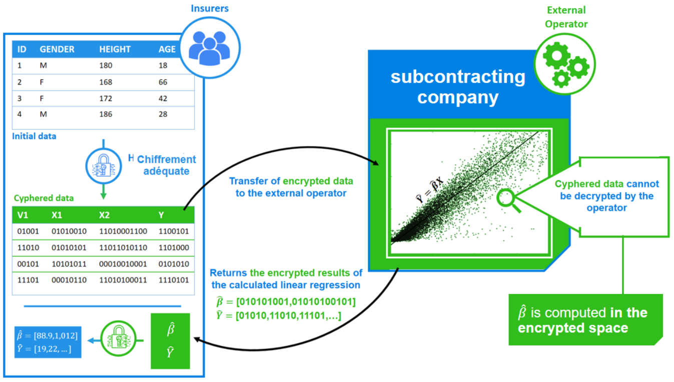
V1	V2	V3	V4
X	TUDPNO	OO	8.6454654
Y	NRTAMI	OO	12.324342
Z	CDHMINUE	OH	9,9987667
R	UHBNRAC	OH	9.01243

ID	NAME	CLASS	AGE
1	DUPONT	A	18
2	MARTIN	A	66
3	DUCHEMIN	B	42
4	BRANCHU	B	28

[1] General Data Protection Regulation (2016), Art. 25.

[2] Poinsignon, T. (2018). Processus de tarification non-vie sur des données chiffrées et anonymisées.

FIGURE 3: PERFORMING A LINEAR REGRESSION ON PSEUDONYMISED DATA



ID	GENDER	HEIGHT	AGE
1	M	180	18
2	F	168	66
3	F	172	42
4	M	186	28

V1	X1	X2	Y
01001	01010010	11010001100	1100101
11010	01010101	11011010110	1101000
00101	10101011	00010010001	0101010
11101	00010110	11010100011	1110101

Insurers
Initial data
H Chiffrement adéquate
Cyphered data
$\hat{\beta} = [88.9,1,012]$
$\hat{Y} = [19,22,...]$
$\hat{\beta}$
$\hat{Y}$
Transfer of encrypted data to the external operator
Returns the encrypted results of the calculated linear regression
$\hat{\beta} = [010101001,01010100101]$
$\hat{Y} = [01010,11010,11101,...]$
External Operator
subcontracting company
$\hat{Y} = \hat{\beta}X$
Cyphered data cannot be decrypted by the operator
$\hat{\beta}$ is computed in the encrypted space

This method could find its place within the cloud-computing framework, by delivering a highly secure process to delegate computations and statistical analysis from an insurer, for instance to an external service provider.

However, to be able to perform such calculations over cyphered data we had to use and implement singular encryption schemes. Such methods are said to be *homomorphic* because they allow us to define equivalent operators ($*_1, *_2$) respectively for $+$ and $\times$ in the encrypted space, enabling computation of linear forms in this space while preserving consistency:

$$\ddot{y} = \ddot{a} *_2 \ddot{x} *_1 \ddot{b} \Leftrightarrow y = a \times x + b$$

Where $\ddot{a}$ is the encrypted value of any a within the cypher space.

Therefore, we focused on two schemes: first of all the *Efficient Integer Vector Homomorphic Encryption* scheme,[3] which we implemented in Python and whose theoretical aspect has the main advantage of being relatively simple to handle. By applying this scheme to our moderate-sized data, and with some concessions on upstream data processing, we were able to perform our linear regression without decrypting the data during the process.

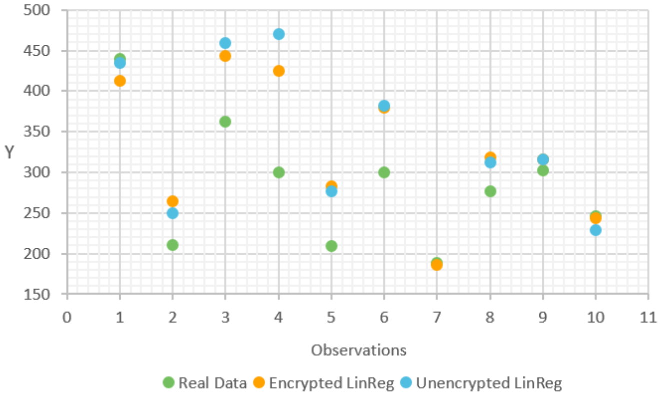
However, these concessions seem us to be overly constraining in a concrete application framework (for example, the secure delegation of calculations from an insurer to an external service provider in *cloud computing*).This is why we decided to reiterate this methodology but using a more robust—and more complex—encryption scheme in R, the *Fan and Vercauteren scheme*.[4]

From this model, we were able to obtain results equivalent to the previous scheme (see the different estimations of Y according to the space in which the linear regression occurred, shown in Figure 4 below) but without having to make any changes to our data beforehand. To achieve this, we proceeded differently. With the *Efficient Integer Vector Homomorphic Encryption* scheme we simply calculated the estimate of the coefficient vector of the regression $\hat{\beta}$ by the formula of ordinary least squares ($\hat{\beta} = (X^T X)^{-1} X^T Y$, with Y the response data vector and X the covariate matrices) in the encrypted space. Here, with the *Fan and Vercauteren* scheme, we have chosen instead to obtain an estimate of $\hat{\beta}$ by performing a gradient descent (GD) in the encrypted space.

Moreover, this gradient descent in the encrypted space converges well towards the value of $\hat{\beta}$ if the number of iterations is large enough.

Nevertheless, the significant computation time induced by operations in the encrypted space, as well as the use of 'only' pseudonymised data (the GDPR remains restrictive on the use of such data and getting off these constraints requires anonymising the database), has led us to consider an alternative approach.

FIGURE 4: RESULT OF LINEAR REGRESSION WITHIN DIFFERENT SPACE



The significant computation time of operations in the encrypted space and its overall complexity led us to consider alternative methods based on anonymization, as recommended by the GDPR, for production purposes.

While we previously looked at procedures that would allow an insurer to externalise its most heavy tasks to a service provider in complete security, we want to focus now instead on a local solution based on anonymisation, which should make non-life pricing more easily compliant with the GDPR.

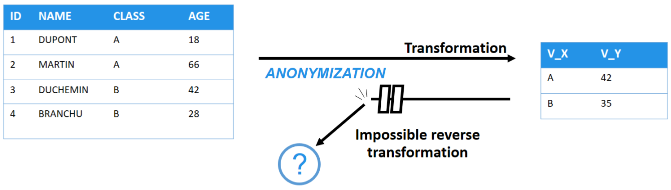
We decided, based on *anonymised* data, to establish a frequency/cost model for a motor insurance pricing and to compare the estimated premium amounts with those obtained using the same model but calibrated, in a usual way, on the individual policies in the portfolio (i.e., non-anonymised).

Data is considered to be anonymized, according to the GDPR, if it is strictly impossible to identify individuals. It is therefore an irreversible and delicate procedure if we want to keep as much information as possible about our data after its anonymisation.

[3] Yu, A. et al. (2015). Efficient Integer Vector Homomorphic Encryption.

[4] Fan, J. & Vercauteren, F. (2012). Somewhat Practical Homomorphic Encryption.

FIGURE 5: DIAGRAM OF ANONYMISATION PROCESS



ID	NAME	CLASS	AGE
1	DUPONT	A	18
2	MARTIN	A	66
3	DUCHEMIN	B	42
4	BRANCHU	B	28

V_X	V_Y
A	42
B	35

One more constraint concerning an anonymisation procedure for the GDPR is the ability to ensure that none of the rows from the anonymised data set refer to any specific individual from the initial data. In other case, this should still be considered as a pseudonymised data set.

Many different methods may be used to anonymise the data set, such as encrypting it and deleting the secret key, though this would make it impossible to use forever. Another common practice is based on adding noise to quantitative data or bucketising it. These last procedures are all generalisation-focused ones. Even if they tend to work and respect the rules of an anonymisation process at first glance, everyone should be aware of the limits and sometimes risks they induce.

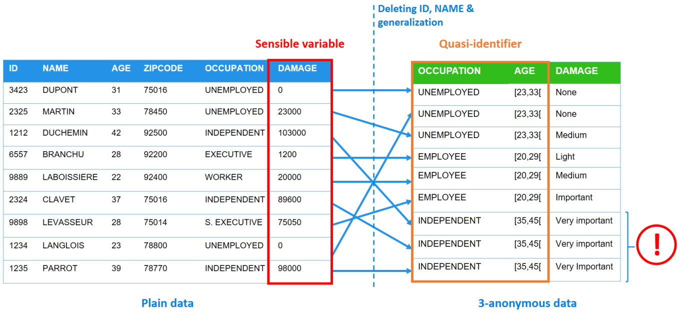
AI-driven anonymisation beyond generalisation

To highlight these limits and risks, let's look at the *k-anonymisation* procedure, which is an anonymisation method based on generalisation that can be assimilated as a very global approach of the bucketisation.

The idea with this procedure is to form groups of *k* observations that will share the same modalities for the explicit variables (the *quasi-identifiers*, which will be the primary key for the data set) within each set, in order to protect the sensible variable (see Figure 6). In order to achieve this, we may make changes to the modalities of either the explicit or sensible variables.

But as easy—or at least as simple—as it seems, this procedure requires a lot of computations to ensure that the choice of the explicit variables and their new generalised modalities are optimal to guarantee the final anonymity of the data set.

FIGURE 6: DIAGRAM OF K-ANONYMISATION PROCESS



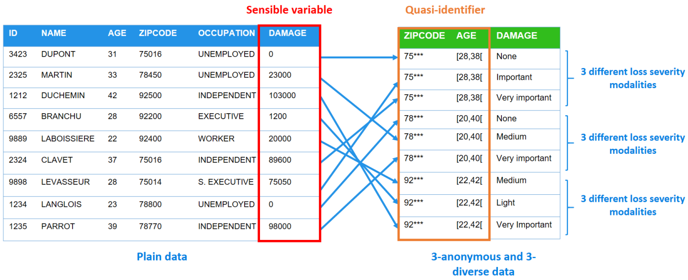
ID	NAME	AGE	ZIPCODE	OCCUPATION	DAMAGE
3423	DUPONT	31	75016	UNEMPLOYED	0
2325	MARTIN	33	78450	UNEMPLOYED	23000
1212	DUCHEMIN	42	92500	INDEPENDENT	103000
6557	BRANCHU	28	92200	EXECUTIVE	1200
9889	LABOISSIERE	22	92400	WORKER	20000
2324	CLAVET	37	75016	INDEPENDENT	89600
9898	LEVASSEUR	28	75014	S. EXECUTIVE	75050
1234	LANGLOIS	23	78800	UNEMPLOYED	0
1235	PARROT	39	78770	INDEPENDENT	98000

OCCUPATION	AGE	DAMAGE
UNEMPLOYED	[23,33[	None
UNEMPLOYED	[23,33[	None
UNEMPLOYED	[23,33[	Medium
EMPLOYEE	[20,29[	Light
EMPLOYEE	[20,29[	Medium
EMPLOYEE	[20,29[	Important
INDEPENDENT	[35,45[	Very important
INDEPENDENT	[35,45[	Very important
INDEPENDENT	[35,45[	Very Important

In fact, in production one uses heuristics to find it in a decent amount of time. This approximation may lead to potential information leak as can be observed in Figure 6. Obviously in this case the green table in Figure 6 cannot be considered as anonymised as it is vulnerable to *homogeneity attacks*, where anyone would be able to know that a specific individual who is either an independent or between 35 and 44 had a very important crash because each observation of the third group shares the same modality for the sensible variable. Plus this table is also vulnerable to *third-party* attacks, so if one knows that any of the clients had an accident and is unemployed and 25, you can be sure that he had medium damage.

In order to impede these weaknesses from happening, multiple constraints can be added to the k-anonymisation, such as *i-diversity*, which ensures that the sensible variable counts at least *i* different modalities between the *k* observation of each group (see Figure 7).

FIGURE 7: I-DIVERSITY CONSTRAINT OVER K-ANONYMISATION



ID	NAME	AGE	ZIPCODE	OCCUPATION	DAMAGE
3423	DUPONT	31	75016	UNEMPLOYED	0
2325	MARTIN	33	78450	UNEMPLOYED	23000
1212	DUCHEMIN	42	92500	INDEPENDENT	103000
6557	BRANCHU	28	92200	EXECUTIVE	1200
9889	LABOISSIERE	22	92400	WORKER	20000
2324	CLAVET	37	75016	INDEPENDENT	89600
9898	LEVASSEUR	28	75014	S. EXECUTIVE	75050
1234	LANGLOIS	23	78800	UNEMPLOYED	0
1235	PARROT	39	78770	INDEPENDENT	98000

ZIPCODE	AGE	DAMAGE
75***	[28,38[	None
75***	[28,38[	Important
75***	[28,38[	Very important
78***	[20,40[	None
78***	[20,40[	Medium
78***	[20,40[	Very important
92***	[22,42[	Medium
92***	[22,42[	Light
92***	[22,42[	Very Important

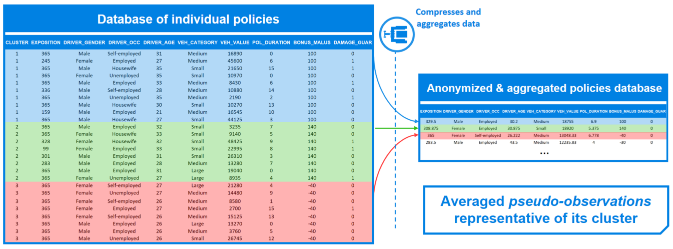
Sensible variable
Quasi-identifier
3 different loss severity modalities
3 different loss severity modalities
3 different loss severity modalities
Plain data
3-anonymous and 3-diverse data

However, this still can be vulnerable to homogeneity attacks[5] (for instance, in Figure 7 one can deduce from the green table that any individual living in a 92*** region who is between 22 and 41 had an accident). Furthermore, generalisation methods are finally all based on decorrelating explanatory variables (quasi-identifiers) from the variable of interest (sensible variable). In our case, where we want to perform a frequency/cost model for a motor insurance pricing from the anonymised data set, this is definitely something we would like to limit.

Therefore, we have been looking at an alternative anonymisation methodology which would verify those constraints. Coming from the idea of the generalisation methods consisting in grouping observations (or policies in our case) from a horizontal point of view (by aggregating variables and modalities, etc.), we thought of a vertical approach based on line-by-line aggregation (see Figure 8), which will both enforce the guarantee of the anonymisation process and also the remaining variance in the final data set, thanks to a clustering algorithm.

[5] When all modalities for the sensitive variable are equivalent in a single group.

FIGURE 8: DIAGRAM OF LINE-BY-LINE AGGREGATION ANONYMISATION



Database of individual policies

CLUSTER	EXPOSITION	DRIVER_GENDER	DRIVER_OCC	DRIVER_AGE	VEH_CATEGORY	VEH_VALUE	POL_DURATION	BONUS_MALUS	DAMAGE_GUAR
1	365	Male	Self-employed	31	Medium	16890	0	100	0
1	245	Female	Employed	27	Medium	45600	6	100	1
1	365	Male	Housewife	35	Small	21650	15	100	1
1	365	Female	Unemployed	35	Small	10970	0	100	0
1	365	Male	Employed	33	Medium	8430	6	100	1
1	336	Male	Self-employed	28	Medium	10880	14	100	0
1	365	Male	Unemployed	35	Medium	2190	2	100	1
1	365	Male	Housewife	30	Small	10270	13	100	0
1	159	Male	Employed	21	Medium	16545	10	100	0
1	365	Male	Housewife	27	Small	44125	3	100	0
2	365	Male	Employed	32	Small	3235	7	140	0
2	365	Female	Housewife	33	Small	9140	5	140	0
2	328	Female	Housewife	32	Small	48425	9	140	1
2	99	Female	Employed	33	Small	22995	8	140	1
2	301	Male	Employed	31	Small	26310	3	140	0
2	283	Male	Employed	28	Medium	13280	7	140	0
2	365	Male	Employed	31	Large	19040	0	140	0
2	365	Female	Unemployed	27	Large	8935	4	140	1
3	365	Female	Self-employed	27	Large	21280	4	-40	0
3	365	Female	Unemployed	27	Medium	14480	9	-40	0
3	365	Female	Self-employed	26	Medium	8580	1	-40	0
3	365	Female	Employed	27	Medium	2700	15	-40	1
3	365	Female	Self-employed	26	Medium	15125	13	-40	0
3	365	Male	Employed	26	Large	13270	0	-40	0
3	365	Male	Employed	26	Medium	3760	5	-40	0
3	365	Male	Unemployed	26	Small	26745	12	-40	0

Anonymized & aggregated policies database

EXPOSITION	DRIVER_GENDER	DRIVER_OCC	DRIVER_AGE	VEH_CATEGORY	VEH_VALUE	POL_DURATION	BONUS_MALUS	DAMAGE_GUAR
329.5	Male	Employed	30.2	Medium	18755	6.9	100	0
308.875	Female	Employed	30.875	Small	18920	5.375	140	0
365	Female	Self-employed	26.222	Medium	13048.33	6.778	-40	0
283.5	Male	Employed	43.5	Medium	12235.83	4	-30	0

...

Averaged *pseudo-observations* representative of its cluster

Basically, we start by performing an unsupervised machine learning algorithm on our initial data set in order to cluster the data into *n* groups. Then for each cluster we compute a single *pseudo-observation*, which is an average of the observations (policies) of the group (for each quantitative variable we take the mean of the rows, while for the qualitative variables we consider the most represented modality). Finally the *n* pseudo-observations constitute the anonymised data set needed.

The obtained data set is then securely anonymised according to the GDPR, as every row is representative of at least two observations (policies) from the initial data set but it is impossible to exactly know which ones, while not any row is specific to a singular observation from the plain data set.

> A vertical approach based on line-by-line aggregation which will both enforce the guarantee of the anonymisation process and also the remaining variance, thanks to a clustering algorithm.

Once we have defined our methodology to anonymise our data, we need to precisely set the framework of the anonymised pricing procedure we intend to perform.

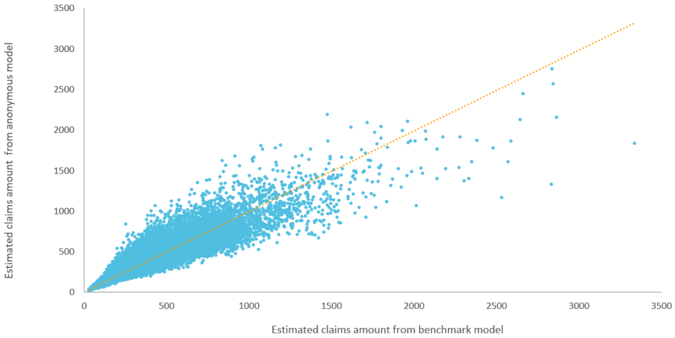
The data we used to realise our motor insurance pricing comes from a pricing game session (*100% Actuaires 2015*), which we classically split into two samples, one for a training purpose (60% of the whole data) and the other for testing. Then the methodology we implement will allow us to compare for a single model the impact of its anonymisation.

In fact, we will start by establishing a single frequency/cost generalised linear model (GLM) and then, on one hand, training and testing the model line by line like one would usually do (benchmark model), while on the other hand, training the model from the anonymised learning sample and then testing it on the test sample—line by line to be coherent in a production case (anonymised model). Finally we compare and look at the deviation between the claims amount estimated through the benchmark and the anonymised models, having as objective to obviously get both as close as possible.

Whereas our approach for anonymisation is driven by artificial intelligence (AI), we had to test several different machine learning algorithms to find the one(s) that would suit the best our pricing model among k-means clustering, hierarchical clustering, density-based clustering (OPTICS), affinity propagation, etc.

We observe that our benchmark model based on a negative binomial distribution (with log as link function) for frequency estimation and on a negative gaussian (with inverse as link function) for averaged cost, provides quite accurate results while slightly overestimating the overall pricing over the test sample. Then, from the whole pool of clustering algorithms we tested, two methods turn out to give some interesting results compared to the benchmark estimations. First, using a k-means algorithm (with k = 6,000) to cluster the learning sample prior to the pseudo-observations computations leads to a relative deviation (on average) of the estimated costs between the anonymised model and the benchmark model of only 4.56%, which can easily be seen graphically in Figure 9 as the points here are quite close to y = x.

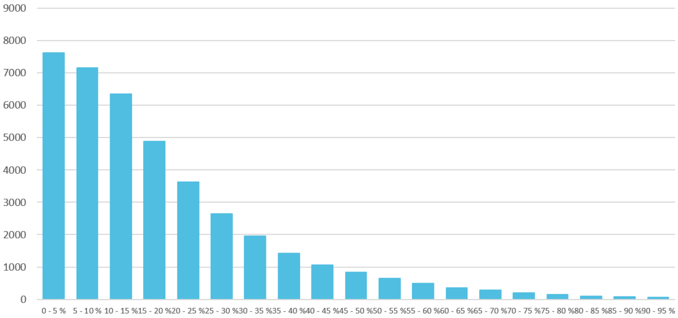
FIGURE 9: COMPARISON MODELS USING K-MEANS ALGORITHM



While using a k-means algorithm allows it to perform well on average (for estimating pricing of the whole portfolio), using a density base clustering such as OPTICS provides more accurate estimation indivdually, as Figure 10 suggests.

From this histogram, one can see that, using an OPTICS algorithm, about 60% of the test base has an individual deviation between the benchmark model and anonymised model of less than 15%.

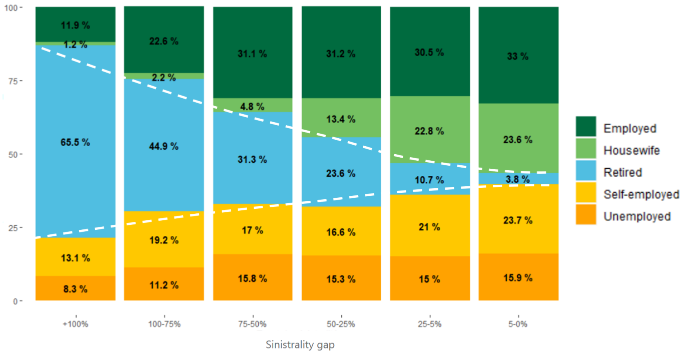
FIGURE 10: RELATIVE DEVIATION DISTRIBUTION, OPTICS



Although depending on the algorithms we chose they can produce decent results, we should still try to determine what is driving the pricing deviation between the models, notably to be able to apply corrections to the most badly predictive estimations from the anonymised model so they get closer to the benchmark results. To do this, one should simply have a look at variables modalities distribution from the test sample according to the measured deviation. In the example in Figure 11, we can for instance very clearly see that retired insurees are way more likely to see their premiums overestimated through the anonymised model than the others. Obviously, you should also look at more variables to refine the corrective analysis.

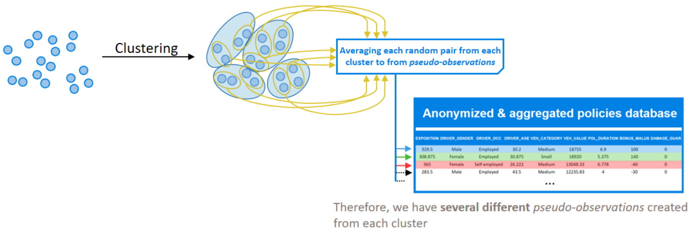
> These methods also reflect the emergence of new needs, particularly related to cyber risk, which is currently the focus of particular attention in the insurance sector.

FIGURE 11: ANALYSIS OF DEVIATION, DRIVER OCCUPATION



We presented here a working AI-driven anonymisation methodology to develop an anonymised motor insurance pricing procedure based on GLM statistical modelisation. However, more and more pricings are now also based upon other techniques like decision tree methods — Classification and Regression Tree (CART), random forest, etc. — and applying such anonymisation straight away will not give expected results. In such case one should consider tweaking the method to try to preserve even more variance in the anonymised data set, for instance by computing more pseudo-observations after the clustering phase (see Figure 12). For instance with this alternative methodology, we achieved to greatly reduce the pricing deviation for a tree based model (CART).

FIGURE 12: ALTERNATIVE TO PRESERVE MORE VARIANCE


Clustering
Averaging each random pair from each cluster to from *pseudo-observations*
Anonymized & aggregated policies database

EXPOSITION	DRIVER_GENDER	DRIVER_OCC	DRIVER_AGE	VEH_CATEGORY	VEH_VALUE	POL_DURATION	BONUS_MALUS	DAMAGE_GUAR
329.5	Male	Employed	30.2	Medium	18755	6.9	100	0
308.875	Female	Employed	30.875	Small	18920	5.375	140	0
365	Female	Self-employed	26.222	Medium	13048.33	6.778	-40	0
283.5	Male	Employed	43.5	Medium	12235.83	4	-30	0

...

Therefore, we have **several different** *pseudo-observations* created from each cluster

> In addition, these methods also find their place in the emergence of new needs, in particular cyber risk, which is currently the subject of attention by insurers

The challenge of maintaining data privacy is a recent technical issue highlighted by the advent of Big Data but also by the regulatory changes brought about by the GDPR in particular. The techniques presented here are avenues for consideration to take these constraints into account in the establishment of traditional actuarial techniques such as pricing issues, via a delegation of the insurer's calculations by encrypted cloud computing (pseudonymisation), but also locally with the insurer's facilities via an anonymised pricing technique.

Finally, as a research subject we are constantly working on developing new solutions to improve our methodologies and to be able to support as many pricing approaches as necessary in the future to easily build effective actuarial models within privacy standards.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACTS

Thomas POINSIGNON, IA
Consulting Actuary & Data scientist, **Paris**
thomas.poinsignon@milliman.com

Remi BELLINA, IA,
Consulting Actuary & Data Scientist**, Paris**
remi.bellina@milliman.com