BIG DATA PROCESS, MACHINE LEARNING AND ERM APPROACH

TO BETTER MONITOR A GROUP BENEFITS PORTFOLIO

V. Ranaivozanany, Qualified and IA-certified actuary, ERM CERA expert
C. Atchama, Qualified and IA-certified actuary
C. Laurans, Qualified and IA-certified actuary
CNP assurances

Corresponding author : Mrs Voahirana RANAIVOZANANY
e-mail : voahirana.ranaivozanany@cnp.fr
Mailing address : CNP assurances - Direction Technique Groupe – 4, place Raoul Dautry - 75015 PARIS.

This study proposes the use of technological and methodological advances to monitor group benefits plans. The standard monitoring method has reached its limits in the current context, where reactivity and precision are becoming essential.

Supervising and monitoring a group benefits portfolio is based on a comparison of claims expenses with premiums received to cover the claims.

Traditional monitoring consists in annually analysing claims-to-premiums ratios and, depending on the size of the portfolio, the analysis is performed by combining all or some of the customers (group plans). Under pressure from a highly competitive market, pricing margins are nonetheless small, and quickly identifying deviations in claims rates has become a priority. In addition, to be competitive, the monitoring process differentiates between customers who weaken the portfolio and those who improve it. Furthermore, this monitoring would be more efficient if price elasticity was considered according to customer profiles.

Established with a prototype, the process recommended by the study is based on a "big data" type approach, constructed using “machine learning” methods [1]. “Big data” seeks to make the most of the huge volume of data available in the portfolio, and open data can also be included to improve monitoring. Compared with conventional statistical methods, “machine learning” algorithms can process huge volumes of data. They also have predictive and prescriptive qualities.

A descriptive analysis was used to better comprehend the portfolio claims ratio. Beyond clearly identified profitability recovery actions, accurate knowledge of the profiles of beneficial or disadvantageous customers has enabled a predictive analysis to be performed on the profitability of new business. A prescriptive analysis is still currently under being studied to enhance the recovery strategy by taking price elasticity into account.

The indicators we analyse could be more responsive it they were automated and regularly updated. Established with fine granularity, they make for more accurate control than conventional indicators. In a prospective way, they also let one anticipate and refine the monitoring.

The last part of the study shows the value of integrating indicators and monitoring principles established in the insurance company's ERM concept. Such an approach will produce KRIs (Key

Risk Indicators) and KPIs (Key Performance Indicators) that are more relevant to insurance risk. It will also help us better integrate portfolio monitoring with the overall strategy of the company. A notion of risk tolerance associated with portfolio hedging is introduced. This tolerance to risk is translated into operational limitations for established indicators. We propose a reporting table to monitor indicators on an ongoing basis.

This study aims to underline the advantages of such an approach. It needs to be completed and adapted to the reality and needs of portfolios.

******

Keywords: big data, machine learning, ERM, group benefits, supervision, claims ratio, recovery, profile, subscribing, descriptive, predictive, prescriptive, elasticity, price, KPI, KRI, risk tolerance, indicator, reporting.

******

The study and the views expressed here are solely those of the author.

The results of the study described in this article have been changed to protect confidentiality.

CONTENTS

1. Introduction

1.1 Presentation of the study

This study proposes the use of technological and methodological advances to monitor group benefits plans. The standard monitoring method has reached its limits in the current context, where reactivity and precision are becoming essential.

Supervising and monitoring a group benefits portfolio is based on a comparison of claims expenses with premiums received to cover the claims. Traditional monitoring consists in annually analysing claims-to-premiums ratios. Depending on the size of the portfolio, the analysis is performed by combining all or some of the customers (group plans). More detailed studies are then conducted if specific trends emerge. This standard monitoring method has reached its limits in the current context, where reactivity and precision are becoming essential.

The first part of this article presents the study conducted to improve the supervision of a group benefits portfolio. A “big data” type approach is explored to capitalize on the huge volume of data available on the portfolio and to incorporate open data. The work is based on statistical or “machine learning” methods. Three stages are planned: a descriptive analysis to account for the portfolio's claims ratio, a predictive analysis to assess the potential profitability of new contracts, and a prescriptive analysis to steer the portfolio's policy recovery.

The analytical vectors used in the study can serve as performance indicators in addition to standard indicators. Their granularity is finer, and they can be updated several times in the year if automated. The second part of this article proposes an ERM approach to better integrate these new indicators with the overall strategy of the company.

1.2 The prototype portfolio and its issues

The study is based on a group benefits plan proposed for a particular business segment. In total, the portfolio comprises 15,000 customer companies throughout France. These companies vary in size, from a few employees or more than 500.

The contract offers three guarantees: capital in the event of death, an annuity in the event of non-accidental incapacity/disability, and an annuity in the event of accidental incapacity/disability. The company can choose some or all of these guarantees. Employee subscription is then mandatory.

Prices are revised annually and the contract can be cancelled annually. The pricing process depends on the company's staffing level: a standard grid is used for companies with fewer than 30 employees, and custom price schedules are prepared for larger companies.

Portfolio supervision is based on monitoring claims-to-premiums ratios. The ratios are calculated in total but also separately for the three risks. Monitoring however is done on a consolidated

basis for the 15,000 customer companies, as profitability is assessed by the insurer on a global level.

The claims-to-premiums ratio level stipulates a moderate claims ratio for death cover. However, this is not enough to compensate for the greater incapacity/disability cover, in order to achieve a satisfactory claims ratio level overall.

Furthermore, changes in ratios indicate a rise in the claims ratio since 2009 despite the recovery actions taken. In certain years, price rises have not resulted in an improved claims ratio for the portfolio. At times, additional weakening has even been observed.

******

2. Refining knowledge of the portfolio with a "Big data and Machine learning" approach

2.1 Extracting all the richness of the data

2.1.1 Working with raw data

Monitoring on a consolidated basis introduces compensation between the claims ratio levels of contracts. However, in order to be competitive, the supervision process needs to differentiate between customers liable to weaken the results of the overall portfolio and those that on the contrary improve it. It is important to introduce a monitoring process that considers customer companies individually.

A “big data” type approach is explored primarily to capitalize on the huge volume of data available on the portfolio.

The analysis is based on data internal to the insurance company with regard to customer companies, in its raw state and in particular not pre-processed or synthesized. The data has been enriched with a variable to assess the companies' claims ratio levels. External (or open) data on the business segment has been introduced to refine supervision. Ultimately, the database used includes a little over 15,000 lines and nearly 40 variables. The aim is to exploit available data hitherto underexploited.

<u>Internal data</u>

Internal data has been organized to obtain a condensed and usable form to characterize the customer companies. Of a very diverse nature, the data is taken from several databases not designed to be used in combination. To illustrate this: the company data is of an administrative nature, including the characteristics of employees, details of contracts taken out by the companies, flows of premiums received and benefits paid out, etc. The data retrieved from the various databases then had to be linked to the customer companies. The portfolio data was synthesized in approximately 15.000 lines, described by nearly 50 variables.

A second level of reprocessing was performed on the data thus structured. 32 of the 50 variables were then selected for the study. The variables that, on the basis of expert opinion, could not be connected to claims ratios were removed (example: Full name). The qualification of the data highlighted unreliable variables. We chose to select only one variable in the event of strong correlation (correlation study based on a principal component analysis).

A variable was then created to characterize the price performance of each of the customers. The variable is based on a comparison of benefits paid out and risk premiums received as cover.

External data

Complementing the internal data, external data characterizing the companies was considered (data on French companies).

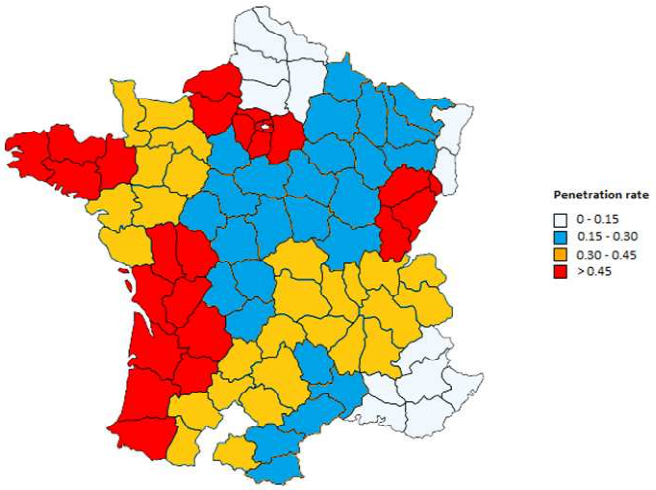
A link was then established. As the external ("open") data in question is structured, this only involved a limited amount of reprocessing.

2.1.2 Graphic representation of the results as a visual aid

For ease of reference, in the context of volumetric analysis, graphic representations were put in place.

As geographical location is a major element, a tool was developed to visually map out the statistics and results on a map of France. This positions the different established indicators.

For illustrative purposes, the map below shows the penetration rates of the insurance company in the business segment.



Penetration rate per region for the companies

For remember, the results of the study described in this article have been changed to protect confidentiality.

2.1.3 First type of values added: Knowledge of the portfolio with a fine granularity

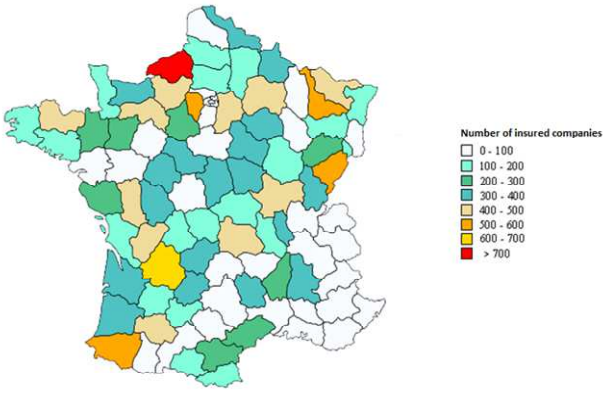
Cover profile

The penetration of the insurance company is shown on a map of France (number of companies insured). It appears not to be well established in the south-east quarter of France. No other significant trend was observed.

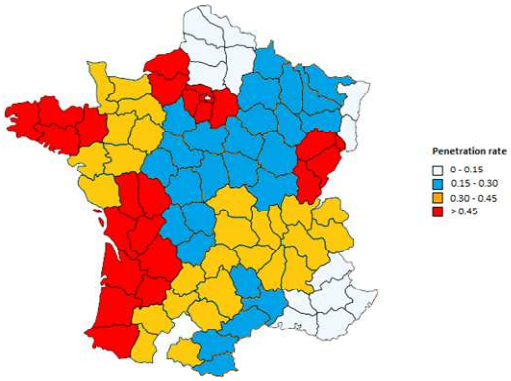
The analysis was completed with a map of penetration rates measured by comparing the numbers of insured companies with the numbers of companies in the business segment for each region ("open data"). Limited penetration rates emerge for the south-east quarter of France. The rates are often comparable for neighbouring regions.

The map of penetration rates lets one assess the potential for market exploration in the different regions. Beyond that, it measures the match between commercial policy and achievements if targets are set for each geographical segment.

Number of insured companies



Penetration rate



For remember, the results of the study described in this article have been changed to protect confidentiality.

Price performance according to department

The variable created to characterize price performance is retranscribed to compare the price performance of the different departments. The death risk is good in all departments, with a few exceptions. For non-accidental occupational incapacity, there are contrasts in the quality of subscription between departments. For accidental occupational incapacity, the quality of subscription is uniformly worse.

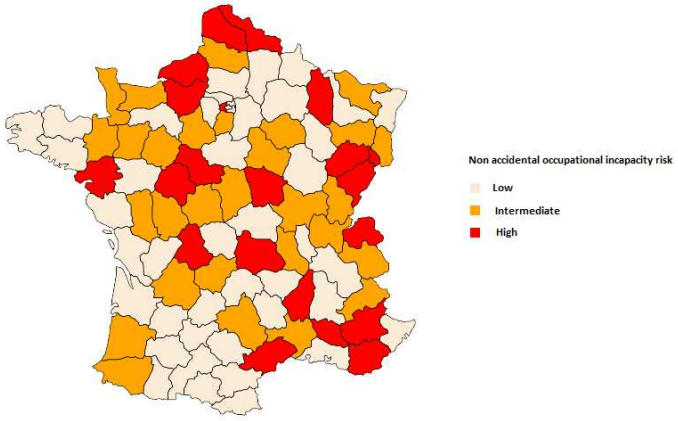


Illustration: non-accidental occupational incapacity

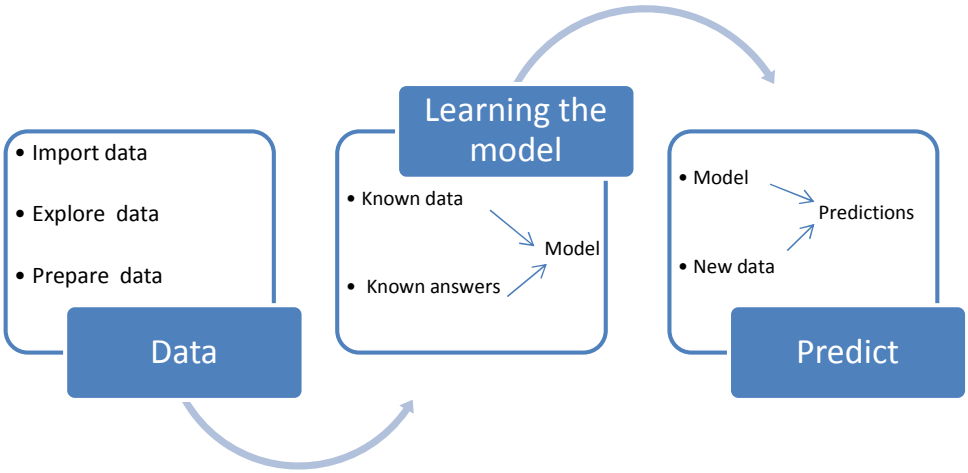
For remember, the results of the study described in this article have been changed to protect confidentiality.

2.2 Applying "Machine learning" to capitalize on big volumes of experience data

The work is based on statistical or “machine learning” methods. Compared with conventional statistical methods, “machine learning” algorithms can process huge volumes of data. They also have predictive and prescriptive qualities. One of their purposes consists in assigning analysed individuals to predefined classes.

2.2.1 Three complementary methods to confirm the results

The “machine learning” principle models behaviour according to past observations. These non-parametric methods let one keep as much flexibility as possible in the models.



“Machine learning” is used in the study for several purposes: choosing a better segmentation of the policyholder portfolio, predicting the quality of new prospective customers, predicting the reaction of customer companies according to various price recovery policy scenarios (assessment of price elasticity – in reflection).

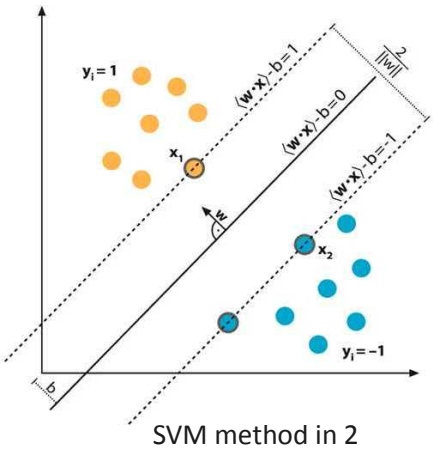
Given the high number of discriminating variables, we chose to use the SVM (Support Vector Machine), “neural network” and “forest of trees” methods, which are well suited to the context. This is because more conventional statistical approaches (linear regression, logistical regression) are too limited to be used in this context.

We used all three methods to leverage their complementarity and to confirm the reliability of the results. Because if different principles are followed, they could generate different results. Convergence of the resulting conclusions is indicative of a high quality of classification.

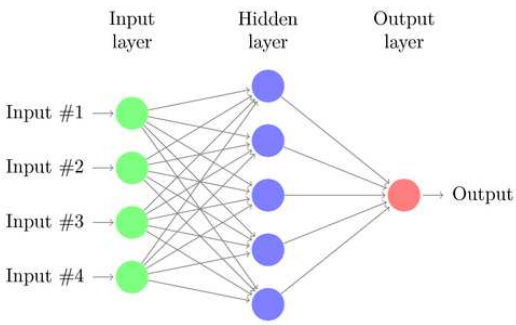
In broad terms, the SVM method schematically isolated each class in space by creating borders.

A neural network is defined to characterize hidden layers and transition functions. In concrete terms, data enters the neural network and is translated for use on output, which enables a class to be assigned to it.

The forest of trees method consists in determining N decision trees from different samples taken from the same data set. Each tree assigns a data item to a class. The class we select is the one having the greatest possible number of people of individual trees.



SVM method in 2 dimensions [5]



Neural network Method [3]

2.2.2 Summary of the analysis approach under consideration

Three analytical stages are planned: a descriptive analysis to account for the portfolio's claims ratio, a predictive analysis to assess the potential profitability of new contracts, and a prescriptive analysis to steer the portfolio's policy recovery. At the time of writing, prescriptive analysis is under design.

The indicators can be updated at any time of the year and can be made more responsive through automation. This is because the annual nature of conventional indicators limits the responsiveness of supervision. Whereas, under pressure from a highly competitive market, pricing margins are nonetheless small, and quickly identifying deviations in claims rates has become a priority.

The table below succinctly describes the proposed approach [4].

Searching for avenues of recovery

	DESCRIPTIVE	PREDICTIVE	PRESCRIPTIVE
	What happened?	**What could happen?**	**What should happen?**
What we need to do	Understand why the portfolio shows a high claims ratio	Estimate the potential profitability of new subscriptions	Steer portfolio recovery policy
What we need to know	The company claims ratio profile The level of claims ratios in each region	Recognize companies that improve and/or weaken the portfolio	How will companies react to recovery policy decisions
What Analytics can contribute	Segmentation of the portfolio according to customers' levels of claims ratios Reporting of claims ratios and penetration rates per department or region, etc. (reflected on a map of France)	Predictive model – Estimate the probable claims ratio of a newly subscribed company	According to the adopted policy, simulate the behaviour of companies Optimization – What would the best recovery policy be
What makes this analysis possible	Bulk processing and standardization of data Principal component analysis, ascending hierarchical classification validated by Machine Learning	Machine Learning with current data	Machine Learning with historical data

Direction of improvement of added values

2.3 Second type of values added: Segmenting the portfolio to better supervise it

2.3.1 Segmentation done: Characterize companies that improve or weaken the portfolio

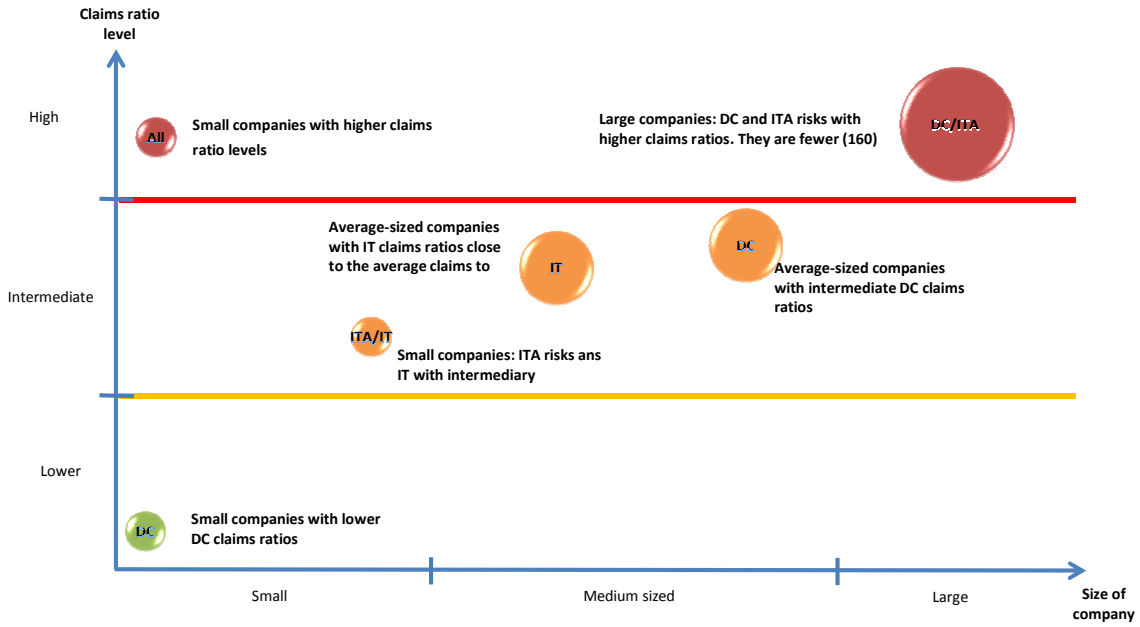
The policyholder portfolio was segmented to characterize companies likely to improve it and those that on the contrary are likely to weaken it. The segmentation was based on quantitative variables. Qualitative variables helped to characterize the established classes. The aim is to form as homogeneous groups as possible.

The principal component analysis we performed was not conclusive with regard to the number of customers to assign and the significant number of explanatory variables. An interpretation of the projection axes and resulting groups did not clearly bring out distinct classes. The segmentation of the portfolio was therefore approached by means of an ascending hierarchical classification.

Several potential classifications emerged from the dendrogram. A line of thought based on knowledge of the portfolio and with the aid of descriptive statistics helped us select 5 classifications that seemed to be meaningful. The quality of each of these 5 classifications was validated with "machine learning" methods. The validation principle is based on measuring the error rate of assigning each company to a given class.

2.3.2 Results

The completed segmentation identified the number of employees in the company as the most discriminating characteristic. Most of the messages are retranscribed in the graph below.



For remember, the results of the study described in this article have been changed to protect confidentiality.

The results revealed several avenues for improving supervision (management actions).

In a comparative manner, companies with more than 500 employees have the highest claims ratio levels. Companies with fewer than 30 employees have the lowest.

There are few companies with more than 500 employees (200 out of the 15,000) but impact on the economy of the portfolio through the number of employees and therefore the number of insured persons they represent. An individualized monitoring of these 200 companies can help contain the portfolio's claims ratio. A claims-to-premiums ratio dashboard can refine recovery actions (prioritization, arbitration, etc.).

Furthermore, custom prices are offered to companies with more than 30 employees but their pricing performance is worse than that of companies with fewer than 30 employees. A table listing commercial prices as opposed to technical prices would help understand whether price reductions are at the root of this poorer performance. A tracking table could monitor due observance of the reduced pricing budget over time.

Indirectly, confirmation of the pricing base or the necessity to adjust it will reveal conclusions about the analysis of price reductions.

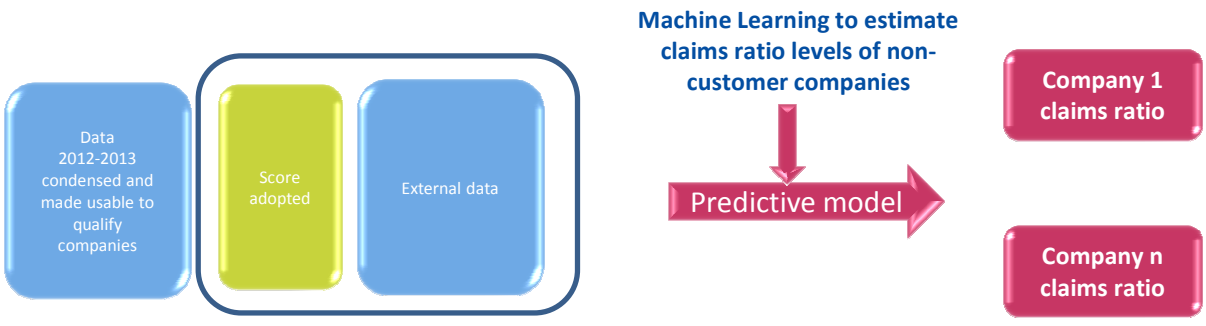
2.4 Third type of values added: Predicting the impact of actions taken

2.4.1 Help for market exploration

The characteristics of companies improving or weakening the portfolio were identified in the previous paragraph. It would be more efficient if the insurer capitalized on this new data to work out its policy with regard to new prospective customers.

By applying “machine learning” algorithms to the data integrating the performance scores of the insured companies, we get an idea of companies not yet in the portfolio that could improve or weaken it.

Each customer company was given a score characterizing its price performance. Modelling was done on a database comprising company identifies, external data on them and their performance scores.



These initial opinions on potential customers are considerable working advantages for the sales teams, even if they are merely indicators. They identify prospective customers that could be profitable for the insurance company and/or those in principle in line with the insurer's subscription policy.

These opinions also serve as choice indicators for negotiations by adjusting proposals in a well-informed manner.

Within the framework of the study, the prediction model characterizes the claims ratio of each new prospective customer. Each company can be assigned to a group and it can be used above all to individually assess prospective customers, more particularly for invitations to tender.

A map showing the profitability of non-customer companies is under consideration. It will offer a more comprehensive view for all departments and influence strategy by geographically directing commercial canvassing, if desired.

2.4.2 Steering recovery policy

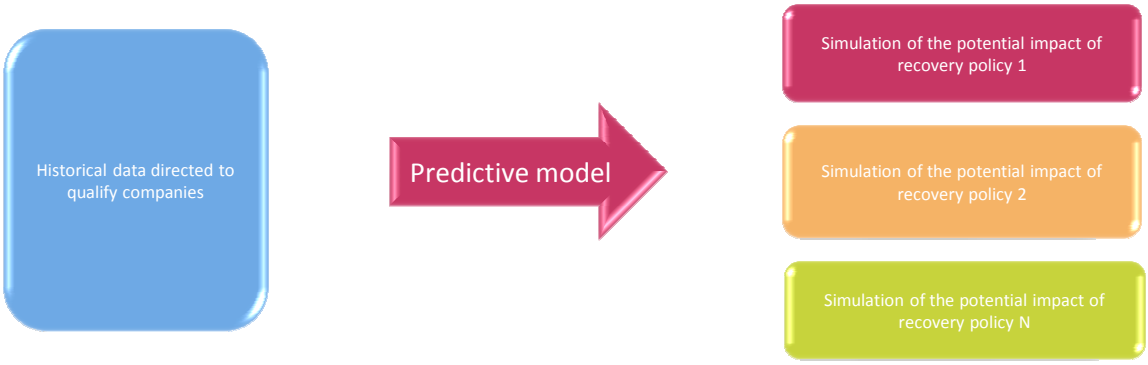
At this juncture, this part is at the thinking and design stage. The point at issue is that price increases in the past have not always had the expected effects.

Considering the 15,000 customer companies and even the 200 largest customer companies, the extent of individualized risk assessments is indeed still limited. Recovery actions did not take into account the circumstances of customers in a sufficiently personalized manner. A price increase could among other things have the adverse effect of triggering cancellation of the contract by a company having a good claims ratio.

An individual analysis of each of the 15,000 customer companies cannot be considered. “Machine learning” algorithms on the other hand can simulate the reactions of customer companies in price recovery scenarios. These algorithm can be based on historical data, which includes the reactions of companies to past price recovery campaigns.

We are therefore considering applying “machine learning” algorithms to adjust future marketing campaigns. The ultimate aim is to find the best commercial policy that optimizes the overall profitability of the portfolio.

Application of Machine Learning



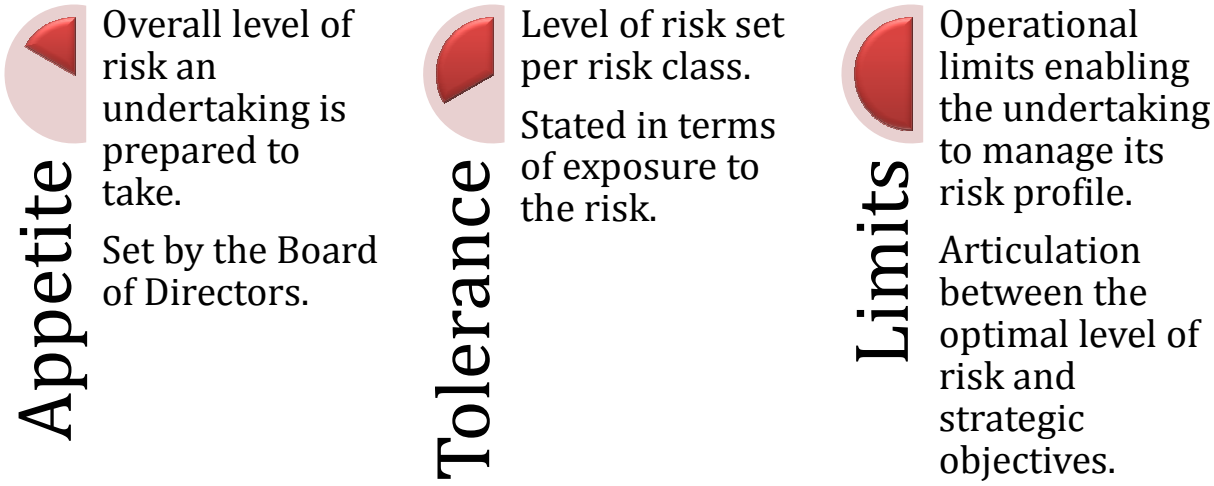
******

3. Improving supervision by incorporating this new knowledge into the ERM concept

3.1 ERM concept

Within the framework of an ERM approach, the insurance company sets an overall level of risk that it is prepared to take to meet its goals. This level called "risk appetite" is then adjusted according to the classes of risk it uses to make decision ("risk tolerance" adjustments).

To organize the company's day-to-day activities according to its goals, these "risk tolerances" are then translated into "operational limits" [2].



New knowledge about the portfolio presented earlier can improve supervision indicators in an ERM framework. These elements were identified through a data analysis at a detailed level and thanks to the power of machine learning algorithms for bulk data processing.

In what follows, three KRI and KPI indicators (Key Risk Indicators and Key Performance Indicators) taken from this analysis are illustrated.

Their incorporation into supervisory reporting on the company's KRIs and KPIs will improve compliance with operational limits consistent with the risk appetite defined by the company.

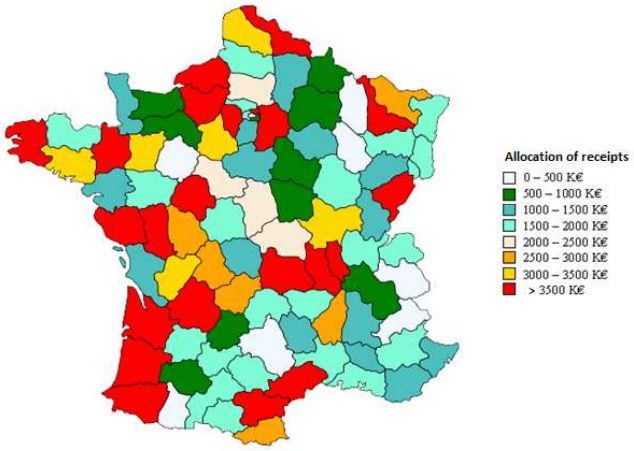
3.2 New KRIs and KPIs

3.2.1 KRI for monitoring risk exposure

A map of France of received premiums is proposed as a KRI to measure risk exposure and potentially to steer subscriptions or encourage thinking on the advisability of adopting risk mitigators.

Receipt levels were allocated to 8 classes (identical ranges). For remember, the results of the study described in this article have been changed to protect confidentiality.

Allocation of receipts



This map compares the insurer's risk exposure in different departments. A macro assessment is still valid even if the indicator is biased by price differentials. High levels of receipts indicate high exposure, and conversely. The level of exposure relates to the number of policyholders and to the intensity of the risk (expected costs).

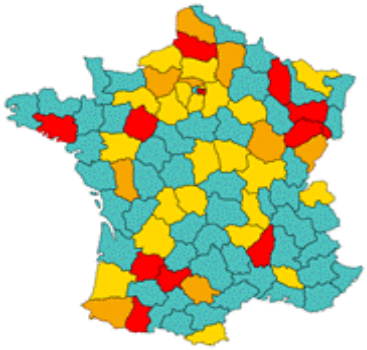
Risk pooling is better when the exposure levels are evenly distributed across all departments (risk diversification). A concentration of high exposures weakens the portfolio. Claims ratio levels for departments are points to watch when monitoring the portfolio.

This monitoring in 2013 brings out high exposures concentrated in 5 regions: Aquitaine, Rhône-Alpes, Languedoc-Roussillon, Poitou-Charentes and Haute-Normandie.

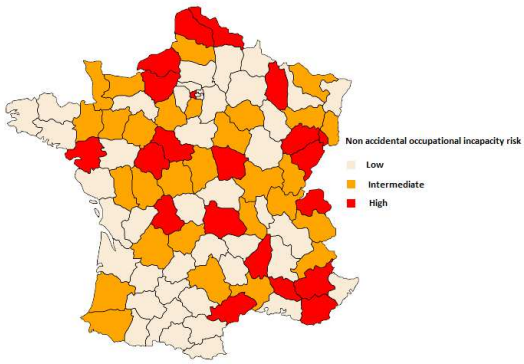
3.2.2 Pricing performance monitoring KRI

The maps of France showing levels of pricing performance are proposed as KPIs. Where relevant they identify departments that greatly undermine the portfolio. They also alert one of the departments that show significant worsening or improvement. Audits can be conducted if deviations are observed. In case of improvement, the department are also points to watch.

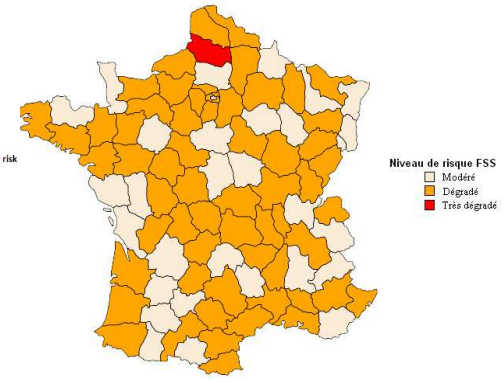
Death risk



Non-accidental occupational incapacity risk



Accidental occupational incapacity risk



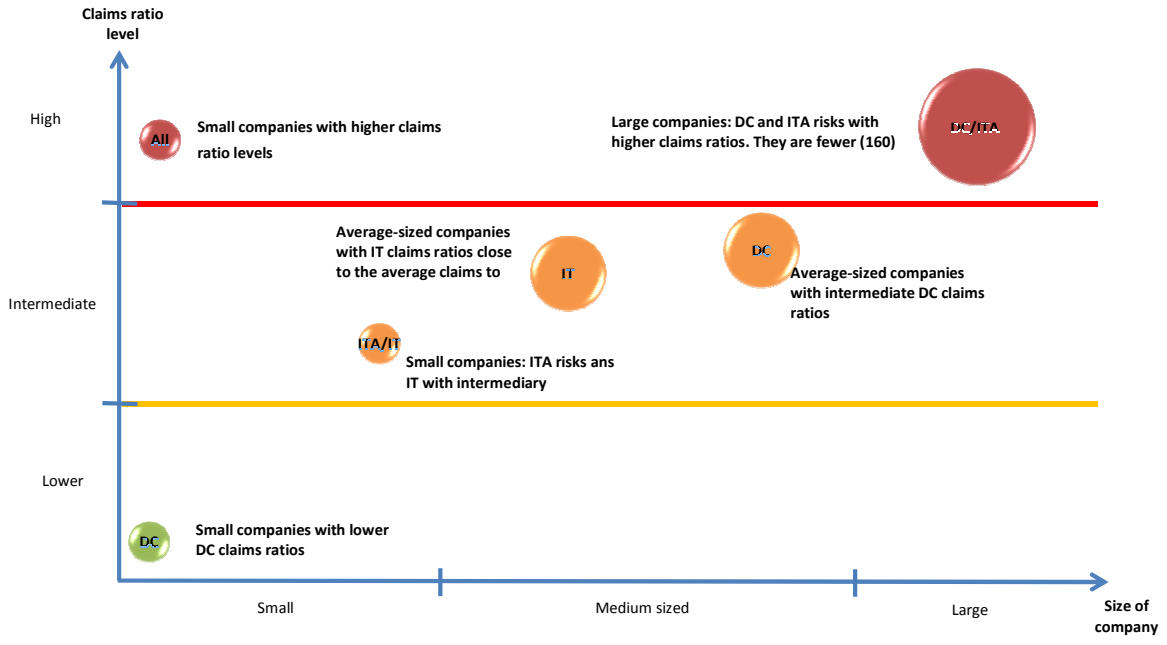
For the 2013 monitoring: the death price performed well with a few exceptions, fairly significant contrasts for non-accidental occupational incapacity, poor but fairly uniform performance fore accidental occupational incapacity.

For remember, the results of the study described in this article have been changed to protect confidentiality.

3.2.3 KRI for monitoring companies that improve or weaken the portfolio

If segmentation processes are automated, the characteristics of companies improving or weakening the portfolio can be evaluated on a regular basis.

The panorama describing the impact of different company profiles on the portfolio is proposed as a KRI. It helps one adjust portfolio supervision as soon as possible by swiftly bringing out the types of customer companies that may require action and the effects of actions taken previously.



3.3 Monitoring reports

The integration of reporting for these KRIs and KPIs can enrich the company's ERM dashboard.

The risk exposure monitoring KRI identifies a high level of cover in part of France. Compared with risk or contextual elements provided by the other indicators, this KRI enlightens the monitoring committee by providing a more accurate view. This may result in "management actions" to implement (steering subscriptions, adopting risk mitigators, etc.).

The price performance monitoring KPI identifies the recovery actions need in departments, where applicable. This will help the insurance company be more efficient and enable it to anticipate "management actions". Indeed, the type of monitoring usually done on a consolidated basis introduces a compensating phenomenon that precludes this rapidity.

The KRI monitoring companies that improve or weaken the portfolio identifies the profiles of customers for whom action is required, and brings out the effects of previous recovery actions.

******

4. Conclusion

This study offers a different outlook on the monitoring and supervision of a group benefits portfolio.

The machine learning algorithms use all the wealth of internal and external data concerning the policyholder portfolio. A very detailed monitoring process can be considered, as a complement to conventional methods. Recovery plans can moreover be drawn up by capitalizing on the effects of past actions.

The type of approach we have presented can moreover offer precise diagnoses that can be renewed whenever needed, several times a year if necessary. The actions to be taken are then identified quickly, as are their effects and the potential need to adjust them.

The greater responsiveness and precision are a considerable competitive advantage in the current highly competitive market.

5. References

[1] Bellina R., Delucinge S., Taillieu F. (2014) : Méthodes d'apprentissage statistique – « Machine Learning »

[2] Dubois D., Ranaivozanany V. (2014) : Diffusion d'une culture ERM au sein d'une organisation : une démarche complexe mais réalisable.

[3] http://www.texample.net/tikz/examples/neural-network/

[4] IBM Software - Thought Leadership White Paper (2013) : Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics.

[5] Université de Bonn : http://www.precision-crop-protection.uni-bonn.de/