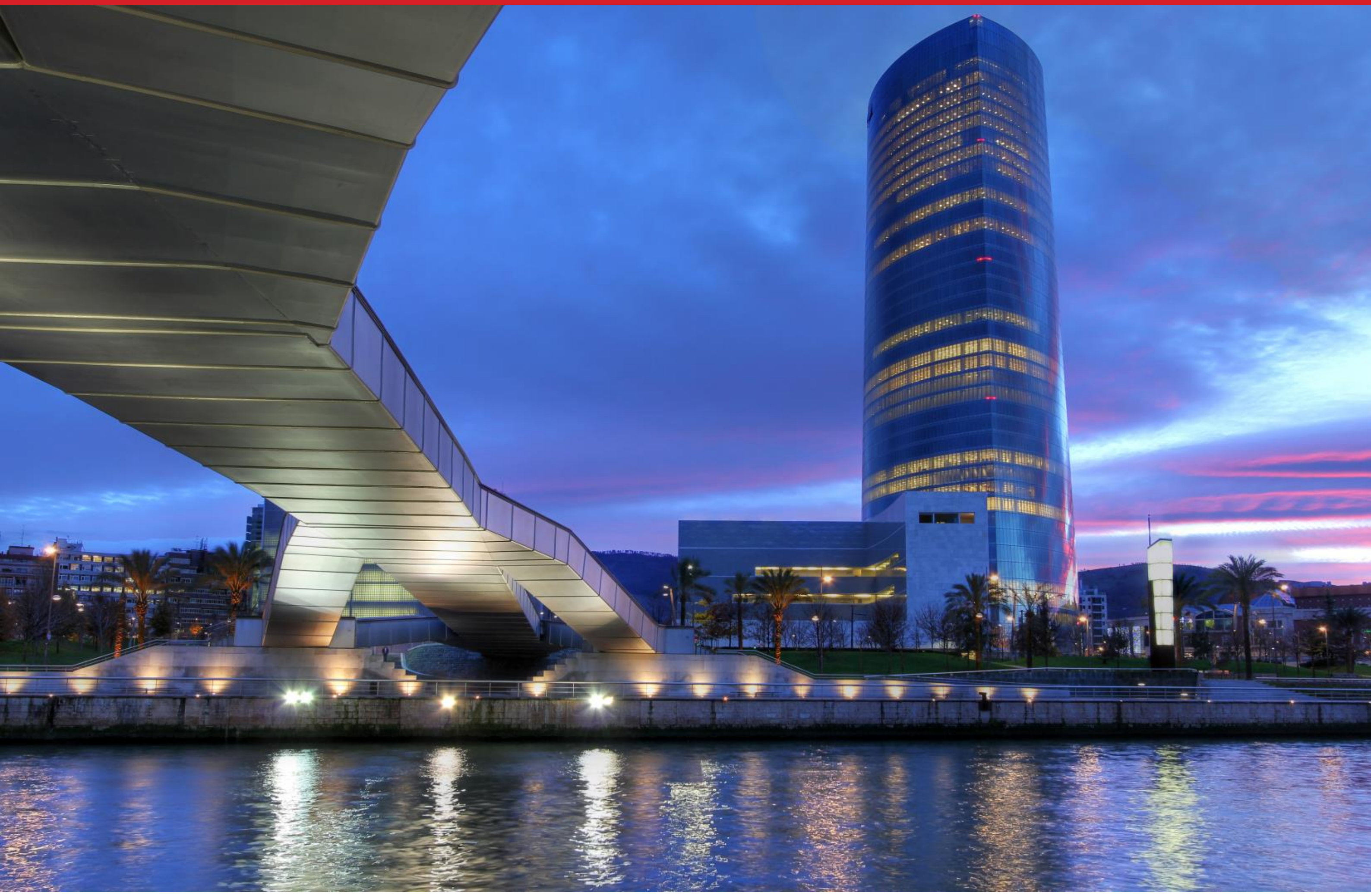
User guide

Technical Tariff Pricing tool

Release 1.0.0



EXECUTIVE SUMMARY

Nowadays, insurance companies must put forward products at competitive prices that must be continuously adapted in order to keep the pace with the rapidly changing environment of finance and financial markets. In order to respond to such needs Reacfin's team developed a pricing tool that allows for the design of a technical tariff in the context of non-life insurance.

The tool design aims to strike a balance between the user friendliness, by automating most of the processes and facilitating the user interaction, and the underlying algorithm complexity based on sound actuarial techniques.

In a nutshell, the tool employs a stepwise regression algorithm based on a Generalized Linear Model (GLM). The goal is to establish which among the explanatory variables has a correlation to the loss and, additionally, to provide the magnitude of this correlation.

Reacfin
Know-How to Risk

TVA: BE 0862.986.729

BNP Paribas 001-4174957-56

RPM Nivelles

Tel: +32 (0)10 84 07 50

info@reacfin.com

www.reacfin.com

Reacfin s.a./n.v.

TVA: BE 0862.986.729

BNP Paribas 001-4174957-56

RPM Nivelles

Tel: +32 (0)10 84 07 50

info@reacfin.com

www.reacfin.com

Reacfin s.a./n.v.

1 INTRODUCTION

This user guide accompanies the Pricing Tool release version 1.0.0. It provides a description of the tool interface and a walk-through via the predefined available demo version.

We assume that the user is familiar with the actuarial methods employed in the context of non-life insurance tariffs and, consequently, this document focuses on the tool utilization in terms of features and functionality. However, the interested reader should refer to [1] for a more elaborate description of such actuarial techniques.

In general, when developing a new commercial tariff, the first step consists in the calculation of the technical tariff. By definition, the technical tariff corresponds to the amount the insurer should charge in the case of a theoretical zero sum game (neither profit nor loss). The purpose of this tool is to facilitate the calculation of such tariff by providing, on the one hand, an easy to use interface and, on the other hand, by using, on the other hand, powerful statistical algorithms in a seamless way. In a nutshell, for the explanatory variable selection the tool relies on a stepwise regression algorithm [2] based on Generalized Linear Models (GLM) [3].

2 PRICING TOOL PROCESS SYNOPSIS AND NOTATIONS

Figure 1 depicts the overall process synopsis. It is noticeable that the application process is divided into successive phases that follow in a linear manner, meaning that later phases are available after the completion of the previous ones. For each phase a corresponding interface is provided in order to allow for user interaction.

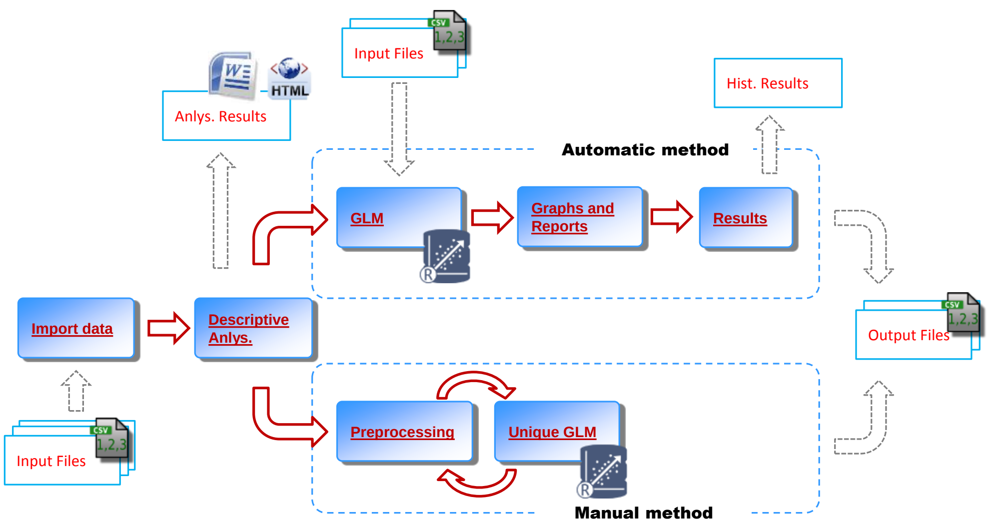


Figure 1. Process synopsis.

The tool employs two distinct methods for the calculation of the tariff, i.e. the manual and automatic methods, respectively. As the name suggests, the automatic method allows for a calculation with as little as possible user interaction, whereas the manual method enables us to fine tune the process. More precisely:

- Automatic method: provides the choice of the explanatory variables that will be included in the analysis from the pool of all variables available in the input data file. Consequently, based on the choice, the tool will automatically determine the significant variables and the corresponding relativities by applying the GLM analysis.
- Manual method: provides, beside the choice of the explanatory variable, also the possibility to preprocess these variables (e.g. applying filters, merging modalities,…) before applying a single GLM analysis.

Hereunder, we denote the main concepts that will be used for the remainder of the document:

- REi : Risk exposure of policyholder i
- $\mathrm{Off_i}$: Offset for policyholder i
- $\mathrm{N_i}$: Observed number of claims of policyholder i
- Ki : Scaling factor of policyholder i
- Pi: Premium paid by policyholder i
- S_i: Total observed claim amount of policyholder *i*
- $C_{i,k}$: k^{th} individual claim amount of policyholder *i*
- β_i : Coefficients of the explanatory variables in the GLM
- $\exp(\beta_i)$: Denoted as “relativities” in this document

The pure premium for the policyholder i can be expressed as:

$$Pure\ Premium_i = E[N_i] \cdot E[C_i] \ \cdot \ \alpha$$

Where we denote by:

- $E[N_i]$ the expected claim frequency for the policyholder i
- $E[C_i]$ the expected mean cost for each claim of policyholder i
- α the scaling factor used for the atypical treatment (see Section 7.2).

The expected claim frequency ($E[N_i]$) and the expected mean cost ($E[C_i]$) must be estimated separately. Consequently, the automatic method will be split into the corresponding frequency analysis and the severity analysis.

3 IMPORT DATA PHASE

Figure 2 depicts the first interface of the tool, namely the Import Data. As the name suggests, its main purpose is to allow the user to import the data file that serves as basis for the calculation of the final tariff. In order to do so, the path of the input file should be provided in the corresponding case. We underline that the appropriate regional settings must be applied in order to accommodate for the correct data separator in the csv file. Additionally, the path of the output folder should be provided as well.

It is noticeable that for the online-demo version of the tool there is no possibility to select an input file since a demo data set is already loaded. Consequently, for the demo version the user is not able to upload his own data.

We underline that the input file should have a specific format. The complete specification of the input file is provided in Section 10.1.

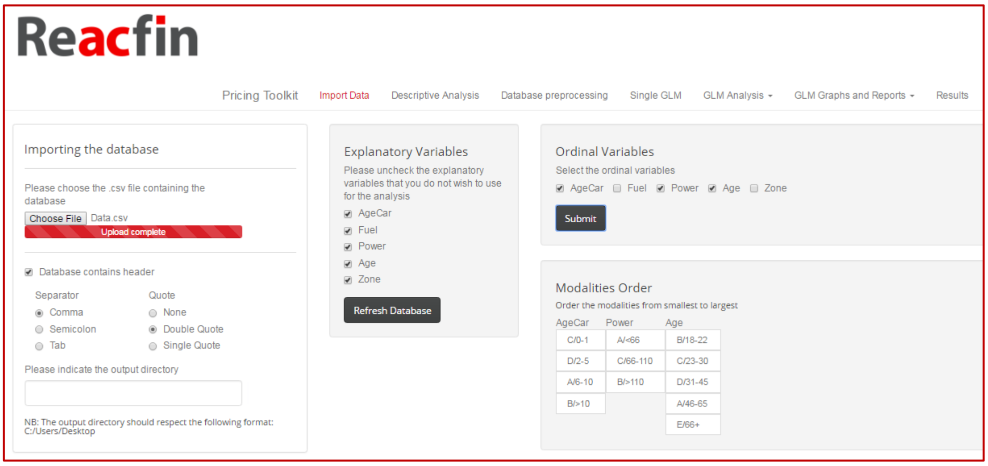


Figure 2. Import Data interface.

Based on the available input file information, a list of all available explanatory variables is displayed on the interface. This will allow the user to choose only a subset of variables that will be further employed in the tariff calculation process. After performing the selection, the data must be refreshed (Refresh Database).

The next step consists in the ordinal variable selection. In order to do so, the user must have prior knowledge of the input data and identify the ordinal ones from the list of selected explanatory variables. We underline that all the input explanatory variables are categorical.

The final step in the Import Data phase consists in choosing the modality order for the ordinal variables. It is noticeable that the tool suggests the modality order (see Figure 3). However, if necessary, the user can change it by drag and drop. The reordering of the ordinal variables will impact the next steps of the process.

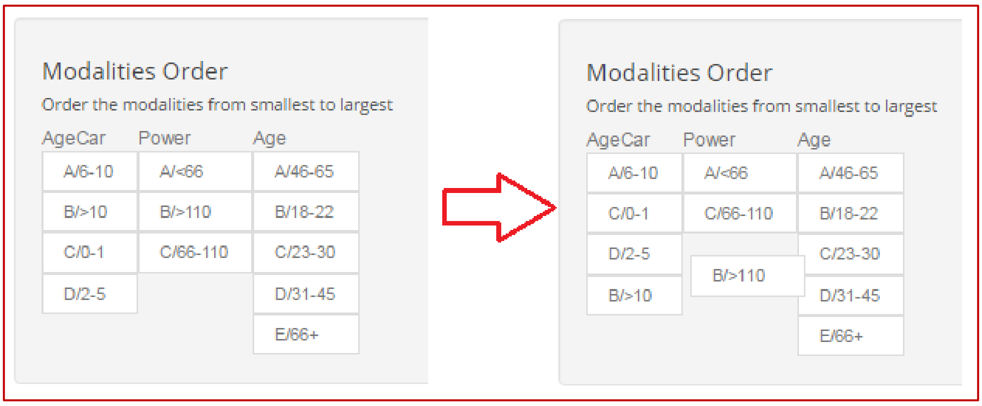


Figure 3. Modality order selection.

It is noticeable that the tool is tagging the modalities of the ordinal variables with a char prefix (e.g. A/, B/, C/, etc). By default, the tool will order the classes according to the risk exposure, meaning that the modalities with greater representation in the input data will be prefixed with “A/”. In order to organise the variables given their logic order the user can drag and drop the corresponding modalities in the desired sequence. For instance, in the example depicted in Figure 3 the AgeCar variable is ordered automatically by the tool according to the risk exposure (i.e. A/6-10). However, the user will reorder them (i.e. C/0-1, D/2-5, A/6-10, B/>10).

The user must enable the selection by clicking the “Submit” button before going to the next phase.

4 DESCRIPTIVE ANALYSIS PHASE

The Descriptive Analysis interface enables the user to visualise basic statistics performed on the explanatory variables available in the input data. It should be noted that the analysis can be performed on the whole set of input variables and not only on the subset of the selected ones during the Import Data phase (see Section 3). In order to populate the drop down list the user must reload the input data via the “Import variables” button. Thus, all the available variables will be then selectable via a dropdown menu.

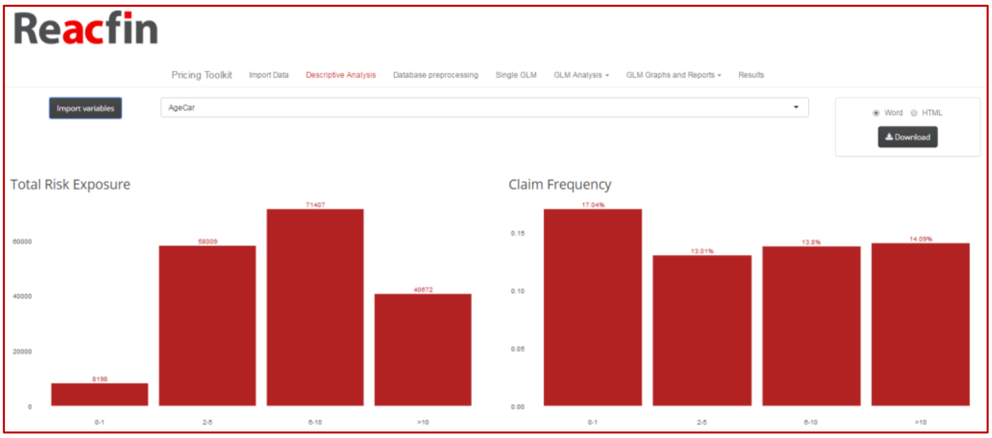


Figure 4. Descriptive analysis interface

As a result of the analysis the following types of histograms will be displayed:

- Total Risk Exposure : $\sum \mathrm{ER_i}$
- Claim Frequency : $\sum \mathrm{N_i} / \sum \mathrm{ER_i}$
- Severity : corresponds to the mean cost $\sum \mathrm{S_i} / \sum \mathrm{N_i}$
- Severity Rate : $\sum (\frac{\mathrm{S_i}}{\mathrm{K_i}}) / \sum \mathrm{N_i}$
- Average Premium: $\sum \mathrm{P_i}$. (it corresponds to the premium available in column 8 from the input file – see Section 10.1)
- Average Scaling : $\sum \mathrm{K_i}$
- Loss Ratio : $\sum \mathrm{S_i} / \sum \mathrm{P_i}$

For each of the resulting histograms the bins will correspond to the variable modalities. Additionally, the results summary table is provided for each variable. It is noticeable that the histograms’ modality order follows the order given by the user during the Import Data phase. This is important since a well-chosen order (as described in Section 3) will allow for an easier interpretation of the results, for instance the identification of trends.

We point out that the graphical results of the analysis (the histograms) are downloadable as a Word or html file, whereas the results table is exported as a csv file only.

5 DATABASE PREPROCESSING PHASE

The Data preprocessing interface allows for the fine tuning of the underlying data. As presented in the synopsis the preprocessing phase is employed iteratively together with the single GLM analysis.

In a nutshell, during this phase, we can apply filters and merge modalities at the level of the explanatory variables. Additionally, we can cap the value and scale the response variables.

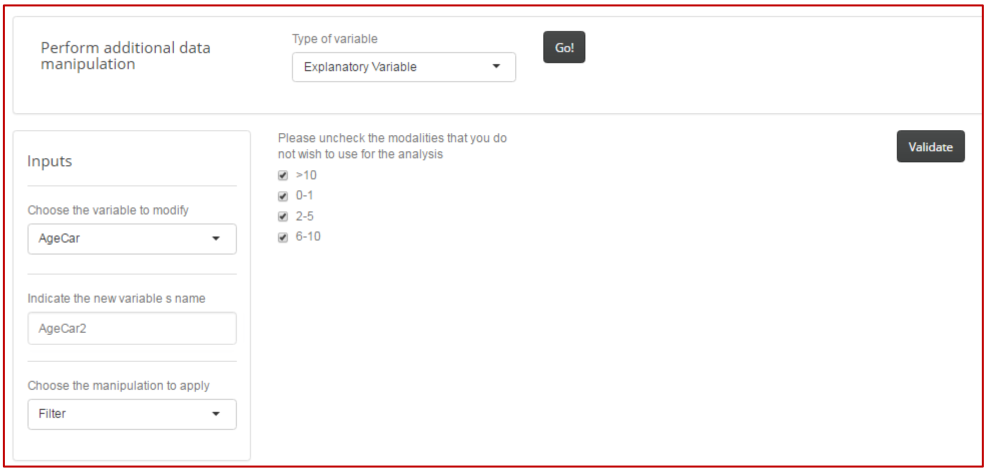


Figure 5. Data adjustment interface

As depicted in Figure 5 the user must choose from the corresponding dropdown menu either the explanatory variables or the response variables as a basis for the data adjustment. Next, the user must identify the variable and specify the name after adjustment. Finally, the type of adjustment must be selected.

It should be noted that after performing this step (by clicking the Validate button) both the original variable data and the data after adjustment will be available in the data set. The name of the new variable should be different from all the column names available in the original data file, otherwise an error message will be displayed.

By default, the response variables available for data adjustment will be the either the number of claims or the total claim amount (or any manipulation thereof).

In the following we will present in detail each of the data adjustments that the user can perform specifying if it is applicable to an explanatory or a response variable, respectively.

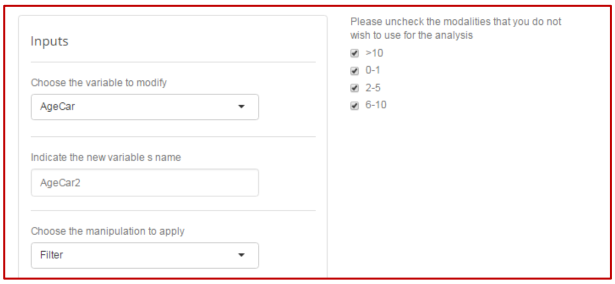


Figure 6. Data adjustment interface – Filter

<u>Filter</u> (applicable for explanatory variables). This adjustment will allow for the exclusion of one or several a specific modalities. For the selected variable the available modalities are displayed on the right side of the screen (see Figure 6).

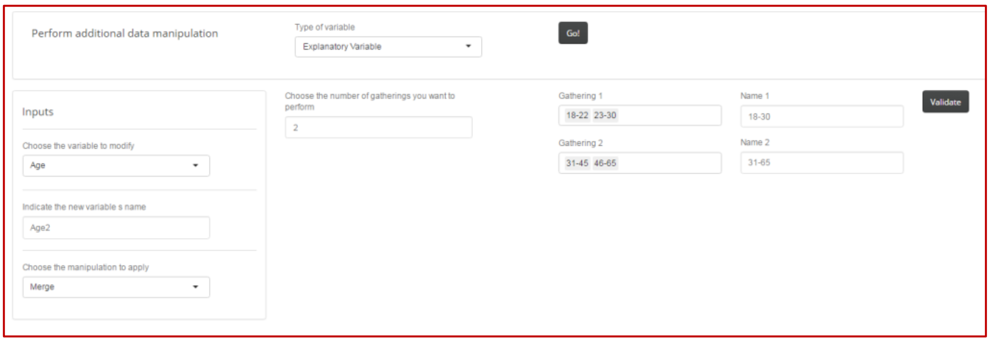


Figure 7. Data adjustment interface - Merge

<u>Merge</u> (applicable for explanatory variables). This adjustment will enable the merging of two or more modalities together. In this case, the user first has to choose the number of groups he wants to perform. Further, he shall define which modalities should be gathered and what should be the name(s) of the newly created modality (modalities).

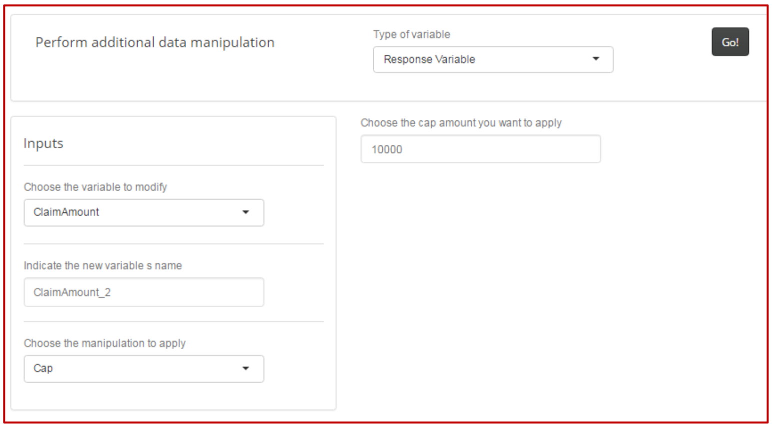


Figure 8. Data adjustment interface - Cap

<u>Cap</u> (applicable for response variables). This adjustment consists in applying a cap on the selected variable. For instance, given the example depicted in Figure 8, the user chooses to cap the value of the "ClaimAmount" variable at 10.000. A new variable named "ClaimAmount_2" will be created. Both variables will be equal except when the first one is greater than 10.000.

Reacfin

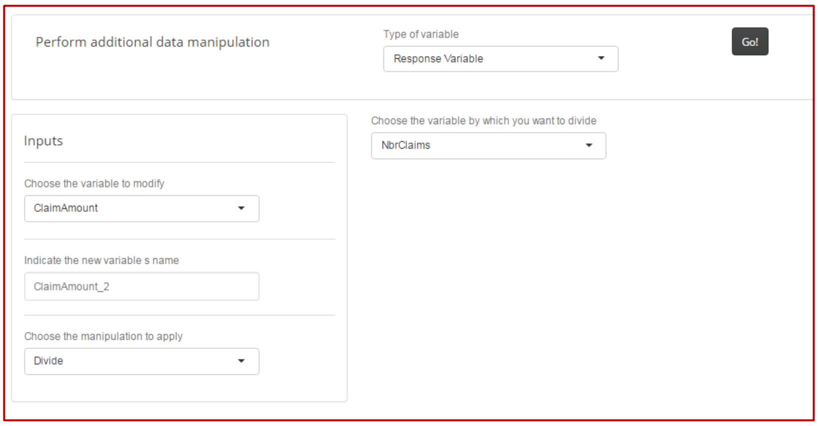


Figure 9. Data adjustment interface - Divide

Divide (available for response variables). This adjustment will allow for the scaling of a response variable by another explanatory variable. For instance, if the user wants to model the cost rate as response variable, then the cost variable will be divided by the sum insured. This action is applicable only to numerical variables.

6 SINGLE GLM ANALYSIS PHASE

During this phase, as the name suggests, the user can perform the single GLM analysis. The following parameters of the regression can be adjusted via the interface:

- Response variables that can be selected from the ones available in the initial data file or the ones created during the preprocessing phase
- Filter variables (optional) selected from the corresponding list created during the preprocessing phase
- Explanatory variables selected from the corresponding list based on the initial data file. Additionally, for each of them a list of possible modification is available.
- The offset, the weight and the scaling variables (optional) can be chosen from among all the variables of the database (initial data file or created variables).
- The p-value (default value 5%)
- The distribution (i.e. Poisson or Gamma)

As a result of the calculation, three tables are provided:

- A list of the rejected and significant variables and their corresponding p-values
- A table containing the estimate, the standard error, the z-value and the p-value of all the modalities of the significant variables
- A list of tables containing Wald's statistic results. Based on this statistic we can decide if modalities should be gathered or not. Thus, if the displayed value is lower than 3.84, the cell is coloured in red and the modalities should be gathered.

Reacfin

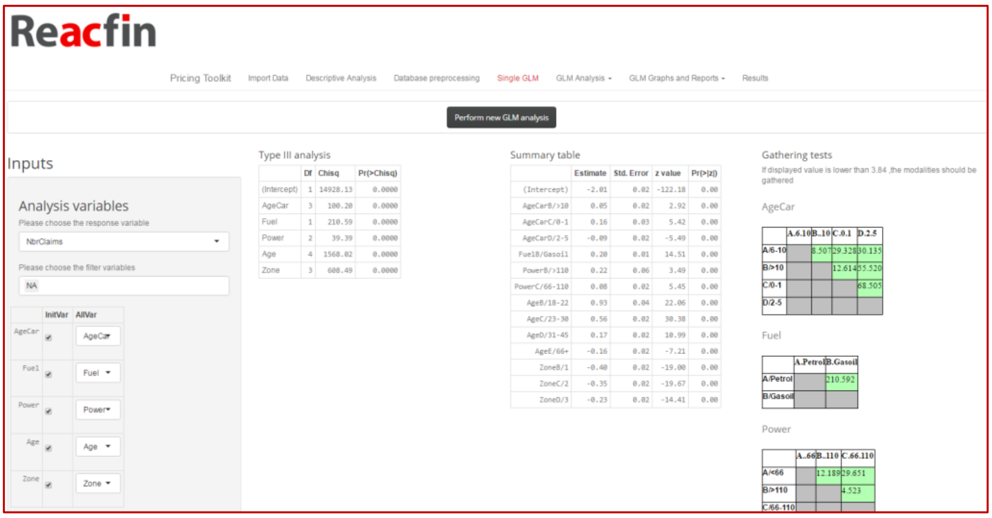


Figure 10. Single GLM interface

7 GLM ANALYSIS PHASE

The GLM analysis interface allows the user to perform the stepwise regression algorithm on the original data. In a nutshell, the algorithm will select the explanatory variables with the most predictive power. The three main approaches for the stepwise regression algorithm are the “forward selection”, “backward elimination” and “bidirectional elimination”. The tool employs the “backward elimination” method, meaning that it starts with all the candidate variables and eliminates the ones with low predictive power.

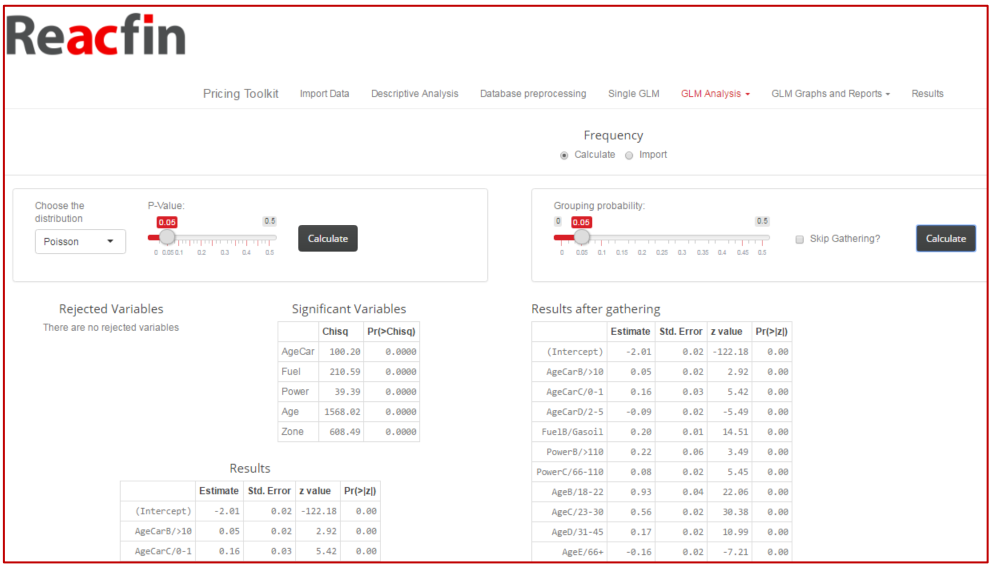


Figure 11. GLM analysis interface - frequency

The purpose of this interface is two folded, providing the possibility to perform the GLM analysis for both the frequency and the average cost. Before running the algorithm the user has the choice either to calculate the result on the current data or to import the results of a previous study.

The GLM analysis is applied for the frequency and severity analysis as described in Section 2.

7.1 Frequency analysis

As a starting point for the frequency analysis the user could choose between the existing data or a previous result. In the second case, two files must be loaded containing the results of a previous GLM analysis:

- Data file containing the predicted number of claims named by default ResultGLM_DB_Freq_Poisson.csv
- The table with the modality and relativities named by default ResultGLM_Relativities_Freq_Poisson.csv

We underline that the correct regional settings must be applied in order to accommodate for the correct data separator in the csv file.

Before running the algorithm the following parameters can be adjusted:

- The distribution model
- The statistical selection threshold (p-value, by default set to 5%)
- The gathering p-value in case we allow the algorithm to group modalities

As a result of the calculation, three tables are provided:

- A list of the rejected variables with their corresponding p-values
- A list of the significant variables and their corresponding p-values
- A table containing the estimate, the standard error, the z-value and the p-value of all the modalities of the significant variables.

As a general guideline, in the last table, a high p-value indicates that this modality should be gathered with the reference class. It is noticeable that the reference classes are not displayed in this table since their estimates are equal to zero.

In case the process is restarted, and new data is provided as input, the results of the previous analysis are still displayed on the interface. However, after performing the variable selection algorithm the new results are displayed.

Following the selection algorithm the user has the option to enable or not the automatic gathering process by ticking the corresponding checkbox ("Skip Gathering?"). If gathering is enabled, the following two-step process is performed.

In step one of the process the gatherings with the reference class are tested for the non-ordinal variables. First, the modalities are rank ordered according to the p-value (highest first). Second, the modalities with p-value above the user defined threshold are gathered with the reference class.

In step two of the process the gatherings with adjacent modalities are tested for the ordinal variables. Additionally, for the non-ordinal variables, the testing of the

Reacfin

gatherings between all the modalities is performed (with the exception of the reference class).

It is obvious that the modality order as defined by the user during the Importing phase (see Section 3) is of the outmost importance since it will impact the gathering process as described above. For instance, given the age of the policyholder as an explanatory variable, the tool will test if “20-21” should be gathered with “18-19” and ”22-23”, but not with “>65”.

It should be underlined that this step is the most computational intensive since the testing of the gathering is perform heuristically (i.e. all possible gatherings are tested and each time a new GLM analysis is performed).

The results of this step consist of three distinct files:

- A data file containing the all initial data and an additional column with the predicted number of claims
- A table with all the modalities and their respective regression coefficients and relativities. This table contains all the variables from the initial file. If a variable was not selected or if it is not significant, its relativity is one.
- An excel file with two sheets containing the result of the GLM analysis and the grouping algorithm.

7.2 Severity analysis

Figure 12 depicts the interface corresponding to the severity analysis. It is noticeable that this interface bears similarities with the frequencies analysis with the exception of two additional parameters: “use scaling” and “atypical threshold”.

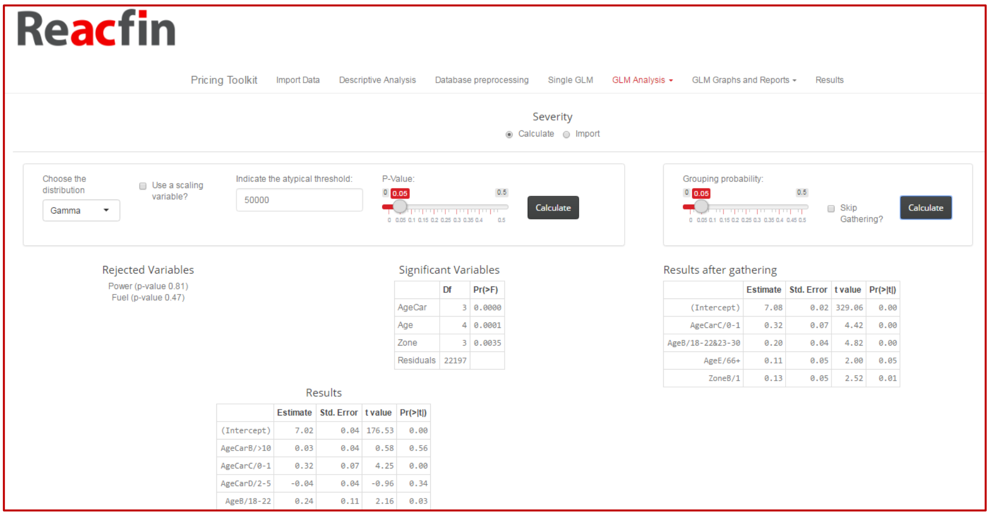


Figure 12. GLM analysis interface - severity

If the response variable is not the cost but the severity rate, the user should employ the scaling variable (for instance to express the cost as a percentage of the sum insured).

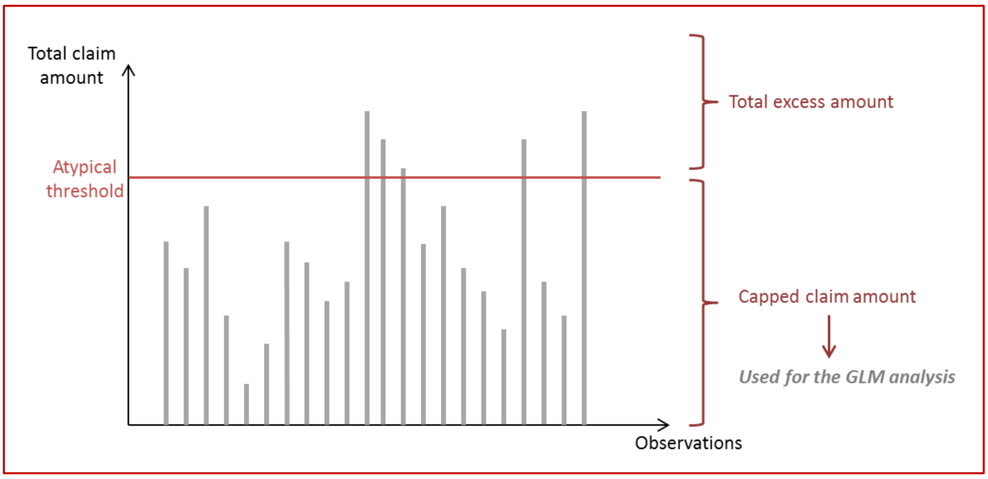


Figure 13. Atypical threshold's application

The atypical threshold is employed in order to apply a cap to the claim values. In the corresponding cell, the user can indicate the atypical threshold (set by default to a very large number). In practice, if the total claim amount exceeds the threshold, the capped claim amount will be equal to the threshold and the excess is then defined as the difference between the total claim amount and the capped claim amount. If the user has chosen to use the scaling variable, the atypical threshold introduced should be expressed in function of this. It should thus be an "atypical threshold rate". At the end of the GLM analysis, the predicted claim amount is then multiplied by the factor:

$$1 + \frac{Total\ excess\ amount}{Total\ observed\ cost}$$

Since, as mentioned above, atypical values are not removed from the database, no atypical threshold parameter was introduced in the frequency analysis. In other words, the number of claims is not impacted by the atypical treatment.

8 GLM GRAPHS AND REPORTS

The purpose of the GLM Graphs and Reports is to present the final results of the analysis performed. The results are presented both in terms of frequency analysis and severity analysis. As presented in Figure 14 the user can select via two dropdown menus the parameters of the graphical results displayed. One of the variables available in the dataset should be selected as the first parameter. The type of graphs is defined by the second parameter that can be either prediction or relativity. By prediction we imply the average of the predicted claim frequency or the average of the predicted cost

Figure 14. GLM Graphs and Reports

Figure 14 depicts as example the result for the "Age Car" variable representing its risk exposure histogram. The solid line is dependent on the type of graph selected by the user. In this example, it is noticeable that people having a new car (aged between 0-1 years) have a claim frequency higher than others. Indeed, all other parameters being equal, these persons have a claim frequency around 15% higher than people with a car aged between 6-10 years. On the contrary, people with a car age between 2-5 years are much less riskier.

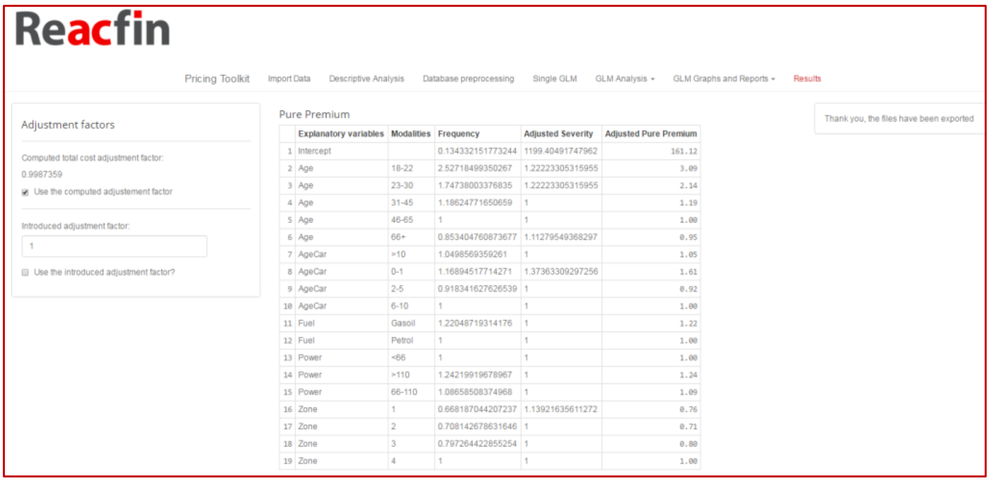
All the graphical results are available for download (see the pop-up camera icon on the upper right corner of the graph).

9 RESULTS

Figure 15 depicts the Result interface where the pure premium values obtained as the product between the frequency and the severity regression process are presented (after applying the adjustment factors). For the variable eliminated during the backward search the frequency and adjusted severity (consequently adjusted pure premium as well) are equal to one. Explain why there are ones in the table.

It is noticeable that the results are adjusted for Severity and Pure Premium. The selection of the adjustment factors values is pseudo automated, meaning that the tool provides a default value that can be further tuned or overridden (as noticeable on the left side of Figure 15).

The default value of the cost adjustment factor is calculated as ratio of the total observed cost and the total predicted cost and is automatically provided via the interface. The user might choose to use this value or not by ticking the corresponding checkbox. Indeed, while the Poisson regression ensures that the predicted number of claims is equal to the observed number of claims, this is not



	Explanatory variables	Modalities	Frequency	Adjusted Severity	Adjusted Pure Premium
1	Intercept		0.134332151773244	1199.40491747962	161.12
2	Age	18-22	2.52718499350267	1.22223305315955	3.09
3	Age	23-30	1.74738003376835	1.22223305315955	2.14
4	Age	31-45	1.18624771650659	1	1.19
5	Age	46-65	1	1	1.00
6	Age	66+	0.853404760873677	1.11279549368297	0.95
7	AgeCar	>10	1.0498569359261	1	1.05
8	AgeCar	0-1	1.16894517714271	1.37363309297256	1.61
9	AgeCar	2-5	0.918341627626539	1	0.92
10	AgeCar	6-10	1	1	1.00
11	Fuel	Gasoil	1.22048719314176	1	1.22
12	Fuel	Petrol	1	1	1.00
13	Power	<66	1	1	1.00
14	Power	>110	1.24219919678967	1	1.24
15	Power	66-110	1.08658508374968	1	1.09
16	Zone	1	0.668187044207237	1.13921635611272	0.76
17	Zone	2	0.708142678631646	1	0.71
18	Zone	3	0.797264422855254	1	0.80
19	Zone	4	1	1	1.00

Figure 15. Results

the case for the other regressions. Applying this factor ensures that the total predicted cost is equal to the total observed cost.

If the user choses to employ a different adjustment factor, such as to accommodate for the IBRN correction, its value has to be introduced via the interface and the corresponding checkbox ticked.

If the checkboxes for both the computed and user defined adjustment factors are ticked then the final factor is calculated as the multiple of the two.

10 INPUT AND OUTPUT FILES

10.1 Input

# Column	Variable	Format
1	Policy number	Integer
2	Year	Integer
3	Risk Exposure	Double
4	Offset	Double
5	Scaling	Double
6	Number of observed claims	Integer
7	Total claim amount	Positive double
8->11	Premiums	Positive double
12->...	Explanatory variables	String, number,...

Table 1. Input file structure.

It is imperative that the input data file, used as a basis for the GLM analysis, should follow the template described in this section in order to ensure the proper functioning of the tool.

The required format for the input file is provided in Table 1 in terms of column number, variable name and variable format, respectively.

The **Policy number** column must contain only unique values. In most cases, the **Offset** will be equal to the **Risk Exposure** and it will be used in the frequency analysis. The **Scaling** can be used, for instance, as the sum insured in case of motor damage insurance or fire insurance. It is noticeable that the input file allows for a number of four different premiums to be introduced (columns 8 to 11). Currently, only one premium value is necessary for the pricing tool.

10.2 Output

The results provided by the tool are downloadable and consist of two distinct csv files.

The first one, named PurePremium_Relativities, contains exactly the same table as the one displayed in the Results interface (see Figure 15). This output table will include all the variables available in the initial data file with their modalities. The Predicted Frequency, Severity and Pure Premiums relativities are provided. The parameters of the frequency are the same as the ones produced after the frequency analysis. The intercept for the severity will be different from the one produced after the severity analysis if adjustment factors are applied (as presented in Section 9).

The second one, named PurePremium_DB, contains the data available in the initial file with three additional columns, namely: the predicted number of claims, the predicted average cost and the pure premium.

It is noticeable that these output files could be used as input for complementary tools such as the Dispersion tool.

11 BIBLIOGRAPHY

[1] M. Denuit, X. Maréchal, S. Pitrebois and J.-F. Walhin, Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems, New York: Wiley, 2007.

[2] Wikipedia, "Stepwise regression," [Online]. Available: https://en.wikipedia.org/wiki/Stepwise_regression.

[3] M. Denuit and A. Charpentier, Mathématiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement., Paris: Collection Economie et Statistique Avancées, Economica, 2005.

Reacfin

Reacfin

Excellence: our outstanding feature
To deliver more than is expected from us, we attract the best people and develop their skills to the most cutting-hedging techniques supported by a robust and rigorous knowledge management framework.

Innovation: our founding ambition
Leveraging on our profound academic roots, we are dedicated on creating inventive solutions by combining our extensive professional experience with the latest scientific research.

Integrity: our commitment
We put work ethics, client's best interest and confidentiality as the foundation of our work. We are fully independent and dedicated at telling the truth.

Solution-driven: our focus
We produce for our clients tangible long-term sustainable value. We help our clients not only to reach the top, we help them reaching the stable top.

Reliability: our characteristic
We never compromise on the quality of our work, the respect of deadlines & budgets and our other commitments. We don't produce reports, we deliver results!

Risk & Portfolio Management

ALM, Portfolio Management & Quantitative Finance

- Implementation and calibration of stochastic models for valuation, trading and risk Management purposes
- Times series analysis & modelling
- Pricing of financial instruments & development of ALM models
- Design/review/implementation of systematic trading & hedging strategies
- Business intelligence in ALM or Portfolio Management
- Tools development (Valuation, Pricing, hedging, portfolio replication, etc.)
- Design of Capital Management solutions

Qualitative Risk Management, Restructuring & Operations

- Organization & Governance
- Businesses restructuring & change management
- Implementation and industrialization of processes
- Internal & regulatory reporting (KRI's & KPI's dashboards)
- Model Review frameworks
- Model Documentation

Insurance specialities

Life, Health and Pension

- DFA* Models
- Capital Requirement assessment
- Business valuation support
- Product development (pricing, profitability,..) & Reserving
- Model validation

Non-Life

- Reserving: triangle methods, individual claims modelling
- Pricing: frequency and severity modelling, large claims analysis, credibility methods, commercial constraints
- DFA models: cash-flows projection, calibration of models
- Reinsurance: modelling covers, optimal reinsurance programs

Xavier Maréchal
CEO
M +32 497 48 98 48
xavier.marechal@reacfin.com

Aurélie Miller
Director
M +32 486 31 60 99
aurelie.miller@reacfin.com

François Ducuroir
Managing Partner
M +32 472 72 32 05
francois.ducuroir@reacfin.com

Samuel Mahy
Director
M +32 498 04 23 90
samuel.mahy@reacfin.com

Maciej Sterzynski
Managing Partner
M +32 485 97 09 16
maciej.sterzynski@reacfin.com

Fabien Verdicq
Director
M +32 498 23 76 30
fabien.verdicq@reacfin.com

Reacfin ***s.a. is a consulting firm, spin-off*** *of the University of Louvain.*

We develop ***innovative solutions*** *and robust practical* ***tools*** *to manage our customers'* ***risks, products, capital & portfolios.***

Reacfin s.a./n.v.
Place de l'Université 25
B-1348 Louvain-la-Neuve
www.reacfin.com
+ 32 (0)10 84 07 50

Linking Academic Excellence with Market Best Practice