Université Jean Monnet

Machine Learning & Data Mining

Data Mining For Big Data: Study of the Groupama datasets

Valentin Benozillo
Josselin Marnat
Mathieu Viola
Rémi Viola

January 22, 2018

Contents

1 Introduction

This project aims at studying a dataset given by the University, in collaboration with Groupama, a French insurance group. As we will see in the next section, the dataset is composed of relational tables, and a set of surveys answers from customers. The goal is to find the actors – words, sequence of words – of (un)satisfaction according to the grades the customers gave in the surveys ; analysis the complaints in function of the customer category, profile, area ; and study the evolution of the customers satisfaction.

Section 2 presents the given dataset, section 3 explains how we loaded and the issues we had to face, section 4,5 and 6 presents the analysis of the customers satisfaction surveys according to various variables, the reclamations reasons, and the comments analysis in satisfaction surveys.

2 Description of the dataset

This dataset the following tables, split in two subsets:

1. Relational Database (6 tables):

    1.1. `BASE_Donnees_Clients` : informations about the customers (ID, age, living area, ...) ;

    1.2. `BASE_Structure_Commerciale` : informations about the company's employees (agency, region, ...) ;

    1.3. `BASE_Demandes_clients_hors_reclamations` : various requests from the customers ;

    1.4. `BASE_Actions_rattachees`_demandes : actions that has be done for a request ;

    1.5. `BASE_Reclamations_clients` : customers complaints ;

    1.6. `BASE_Avantages_clients` : customers advantages ;

2. Satisfaction Surveys (16 tables) :

    - `SATISFACTION_*` : tables containing satisfactions surveys done after a complaint, or randomly sent to customers. The main analysis will be done on these surveys.

All the tables that concerns customers are linked together with a customer ID (`ID_GRC`) Of course, this data is strictly confidential, and has been anonymized beforehand.

3 Loading the dataset

The dataset is a set of tables formatted in CSV (comma separated values), easily readable and loadable in most programming languages. It has the following form:

- the first row describes the names of the variables (IDs, dates, questions, ...);
- then, each row represents the data (either clients, an answer to a survey, a request, ...).

The CSV format works well when the semi-colons within a cell are escaped (*e.g.* ';' → '\;'), but it wasn't the case, which is a real problem if we don't fix this issue. Since the number of columns is increased by how much there is unescaped semi-colons in a row, we decided to remove all these row using this property. It's a loss of data, but treating these rows would have been to much work, and it would be the work of the data pre-processor to escape the semi-colons efficiently.
Once these rows removed, we obtained a set of dataframes.

4 Analysis of the different satisfaction surveys

In this section, we will firstly try to see where are the more and the less satisfied customers of the society. After that, to complete our study, we will also analyze the level of satisfaction of the customers according to others criteria. Finally, we will analyze the evolution of the level of satisfaction of customers.

4.1 How do we have to handle the dataset?

We started by merging all the satisfaction surveys to get an overview of the satisfaction of all the customers. In a second step, we computed the average of the satisfaction for each customers who have completed several surveys. For this customers, we named the type of survey 'Average' to see that it is a mix of several survey. In this new base, we could compute the mean and the standard deviation of the level of satisfaction of the customers. In this case, the mean is 7.94/10 and the standard deviation is 2.33. With this 2 values, we decided to set thresholds to define 3 categories of customers. The most satisfied customers are those who have a level above the mean. The less satisfied customers are those who have a level under the mean minus the standard deviation. Between these 2 group, there is what we call the neutral part. After that we plotted several diagrams to answer the question. To see the code and reproduce our results, please edit the script `q1.R` by changing the first line and put the path where the survey are.

4.2 Satisfaction according to the 'Typologie'

To see the level of satisfaction according to the criteria 'Typologie', we plot the figure 5. This diagram is the aggregation of the number of dissatisfied customers (the left part) with the number of neutral one (the central part) and the number of satisfied one (the right part), according to their typology.

In this figure, the first thing which we can see it is that the majority of the customers come from the agricultural world. This is due to the history of the company. That is why, if we just compute the number of dissatisfied customers, we will see that is for 'rural dynamique' but it is due to the number of customers in this category. To have a better overview of the satisfaction level, we also plot the pies 11 according to each category of 'Typologie' to see the one where the percentage of dissatisfied customers is the most important.

In this set of figures, we can see that the proportion of satisfied and dissatisfied customers are more or less the same with just little differences. The most satisfied group is the 'Hors Territoire' one and by decreasing order, we find the 'Hyper Centre', the 'Peri Urbain', the 'Grande Périphérie Aisée', the 'Rural Dynamique' and the 'Rural Age' group which is the most dissatisfied one.

4.3 Satisfaction according to other criteria

To complete our study of the satisfaction of the customers, we have plotted the same diagrams for the other interesting criteria given in `BASE_Donnees_Clients.csv`. All these figure are available in the appendix A.

4.3.1 Nature Personne

As expected, for this criterion, the number of person is higher than the number of PM which represent associations, companies and others. It is very hard to see the difference of proportion in this situation. But, with the pies 6, we can see that the customers coming from companies are globally less satisfied than simple person.

4.3.2 Segmentation Distributive

For this criterion 2, we have to remember that :

- N = Nouveau,
- S1 = A laisser venir,
- S2 = A fidéliser,
- S3 = A redécouvrir et multi-équiper,
- S4 = A développer et fidéliser.

The problem is that this field is not always complete in the file `BASE_Donnees_Clients`. Sometimes there is nothing, sometimes just a dot, and sometimes null and for all these situations, it seems there is no correlation with other criteria. The rest of the comparison is done without these values.

The most satisfied group is all new customers. After that, it is the sets S4, then S3, S2 and finally S1. The company have to work on these 2 last groups to improve its image.

4.3.3 Tranche age

The diagrams 3 and 10 for this criterion show that the number of young customers is very low but this is the group the most satisfied in proportion. The group of active is the biggest one and the most dissatisfied in average. It is worth noticing that the null set corresponds to the associations and companies set in this case.

4.3.4 Type of surveys

To be complete, when we have merged all the surveys we have kept the name of each ones. So we can analyze which one have the best marks and the worst. As a reminder, a big part of the dataset corresponds to the average of multiple satisfaction marks. It corresponds to the first bar of 4. After that, the most important parts of the dataset correspond to the field 'degats vehicule hors collision' and 'autres evenements ou dommages'.

All the pies of figure 8 correspond to customers having filled only one satisfaction survey. As we have say, the most importants are 'degats vehicule hors collision' and 'autres evenements ou dommages'. The second one correspond to one where the level of satisfaction is the highest with 'bris de glace(auto)'. The first one is not the worst. To find the worst, we have to look at 'demande' and mainly 'evenement entre deux vehicules' where there is more or less the same number of satisfied customers and dissatisfied ones.

Concerning the average computed diagram 9, we can see that the level of satisfaction is not bad. The proportion of dissatisfied customers is the second smallest after 'autres evenements ou dommages'. There is just a big neutral part in this diagram.

4.4 Evolution of the level of satisfaction

The third question of the company was to know the evolution of the level of satisfaction of its customers. To answer this question, we only kept the customers who filled several survey. After this selection, we built a database which contains the initial level of satisfaction and the difference between this mark and the next. If a customer filled more than 2 surveys, we only computed the difference between 2 consecutive surveys (With 3 surveys, the difference between the first and the second and between the second and the third). According to different criteria, we plotted several diagrams to analyze the problem. To see the code and reproduce our results, please edit the script `q3_1.R` by changing the first line and put the path where the survey are.

4.4.1 Global evolution

In the figure 12, we can see that most of the time the mark evolves of no more than one or two points in the decrease or in the increase. It often remains stable.

We also plotted several diagrams to see the evolution according to the previous mark. Thanks to the figures 17 and 18, we can say that it is after a 8, a 9 or a 10 that the evolution is the most frequent. It is also because it is the most frequent given marks.

4.4.2 According to other criteria

We also have computed the evolution of the level of satisfaction according to the 'Typologie', the 'Marche CSP', the 'Tranche d'age', the 'Segmentation distributive' and the 'Nature'. Because of a lack of time, we do not have to look farther in this direction. The different figure are available in the appendix B

5 Analysis of the custommers reclamations and terminations

The goal of this section is to better understood who are the client who do a reclamation or terminate their contract. To reproduce our result please edit the script `q2_1.R` by changing the first line and put the path where the table are.

5.1 Reclamations

5.1.1 Type

Firstly we simply compute the proportion of each possible type of reclamation31 (sinistre, gestion contrat, cotisation, resiliation, souscription, contrat, commerciale, encaissement). As we can see on the pie chart the most hot topic is "sinistre" that's not very remarkable, but the second one is gestion contrat" and the number of reclamation with this type is more than two times the number of reclamation with "cotisation". If we suppose that the real meaning of "gestion contrat" is : the client claim information about his contract, then maybe the custommers are not enough aware about their contract, so it's will be interesting to analysis witch are the real topic of their reclamation.

5.1.2 Typologie

Thanks to the "client" table we are able to know what is the "typologie" of a client for each reclamation. In a first hand we can simply count the number of reclamation for a given "typologie"32. But it exist a bias, by looking this chart we can deduce that custommers "rural dynamique" has more "reclamation" than the others, but they also are the most representative categories in the "client" table. So if we look in term of proportion, by dividing our result by the number of "client" with this "typologie", we got a second chart33, and as we can see the "typologie" as no impact on the proportion of reclamation.

5.1.3 Marche PSO

We do the same thing as before but with the characteristic "MARCHE_PSO" (agricole, acps, particulier, collectivites)3435. But in this case we can see the class "agricole" have more reclamation (in proportion) than the other class. We compute their "TYPE" of reclamation37.

5.1.4 Departement

Thanks to the "COD_INSEE" in the table "client" we are able to know their department. So we compute the proportion of reclamation per department3839.

5.2 Termination

We do the same study for the "resiliation" table.

- Typologie41,42
- MARCHE_PSO43,44
- Departement45

6 Comments analysis in satisfaction surveys

The goal of this section was to understand the reasons for customer satisfaction and dissatisfaction. To reproduce our results please edit the script `q2_keywords.R` by changing the first and put the path where the table are.

At first, we gathered all the comments from the satisfaction according to the satisfaction rating: one set for the good ratings – the highest 30%) and the bad ratings. Then, we removes the French stop-words: prepositions, articles, pronouns ; as well as numbers. And retrieved the roots of words by stemming, in order to remove conjugations, feminine or plural forms. This allowed us to group the words together more efficiently. Finally, we computer the n-grams for $n \in [1..5]$. For example, the sentence ‘I like your car’ has the following 2-grams: ‘I like’, ‘like your’, and ‘your car’ ; and we extracted the 20 most frequent terms on this n-grams dataset. In order to keep only the meaning-full n-grams, we computed the intersection between the satisfied and unsatisfied n-grams, and removed them from the tables. Thus, the n-grams are exclusively attributed to on set of comments.

We obtained what is shown in tables 1 to 5. You can see for each set of two tables: 1) on the left: the 20 most frequent n-grams according to satisfied customers and 2) on the right: according to the unsatisfied clients.

7 Conclusion

A Levels of satisfaction

A.1 Global

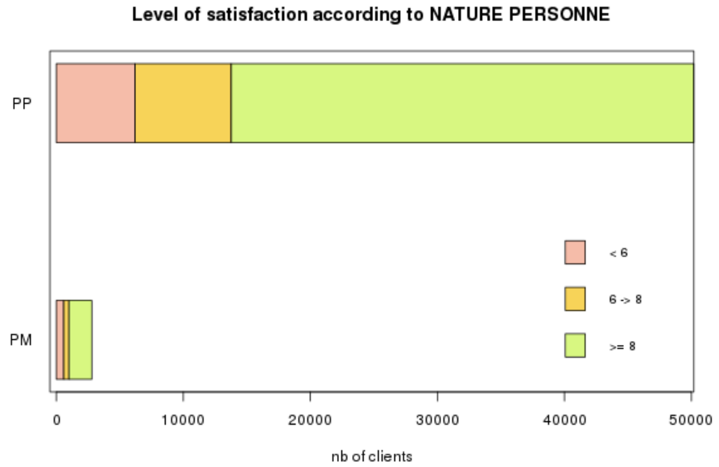


Figure 1: Level of satisfaction according to NATURE PERSONNE

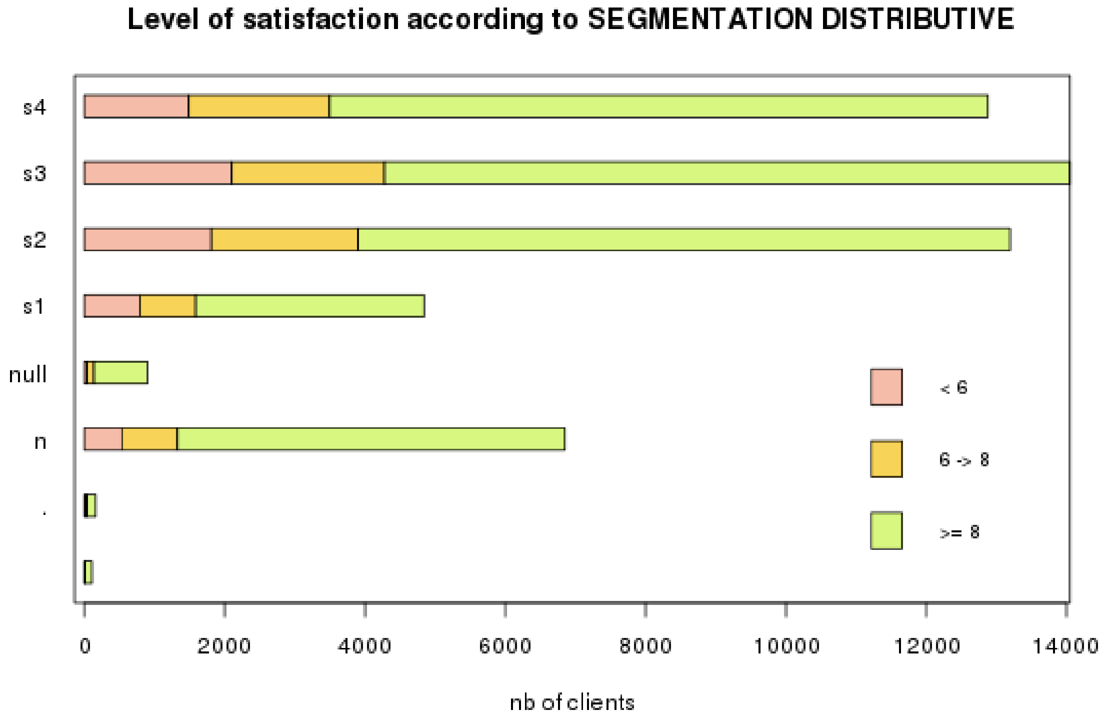


Figure 2: Level of satisfaction according to SEGMENTATION DISTRIBUTIVE

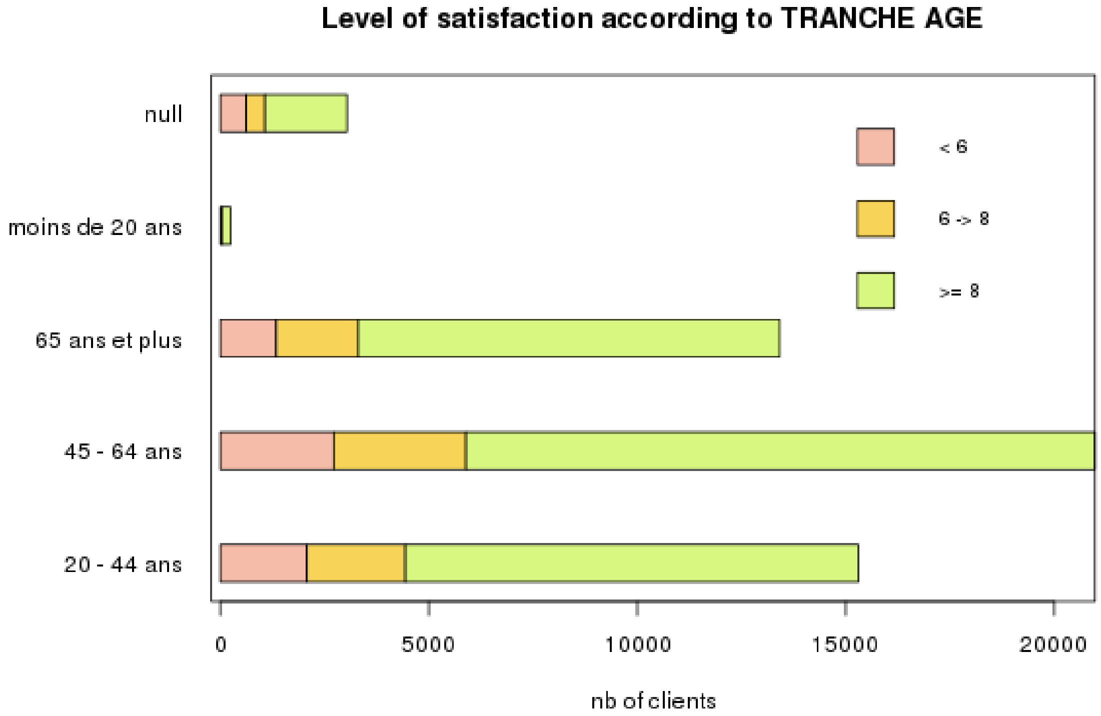


Figure 3: Level of satisfaction according to TRANCHE AGE

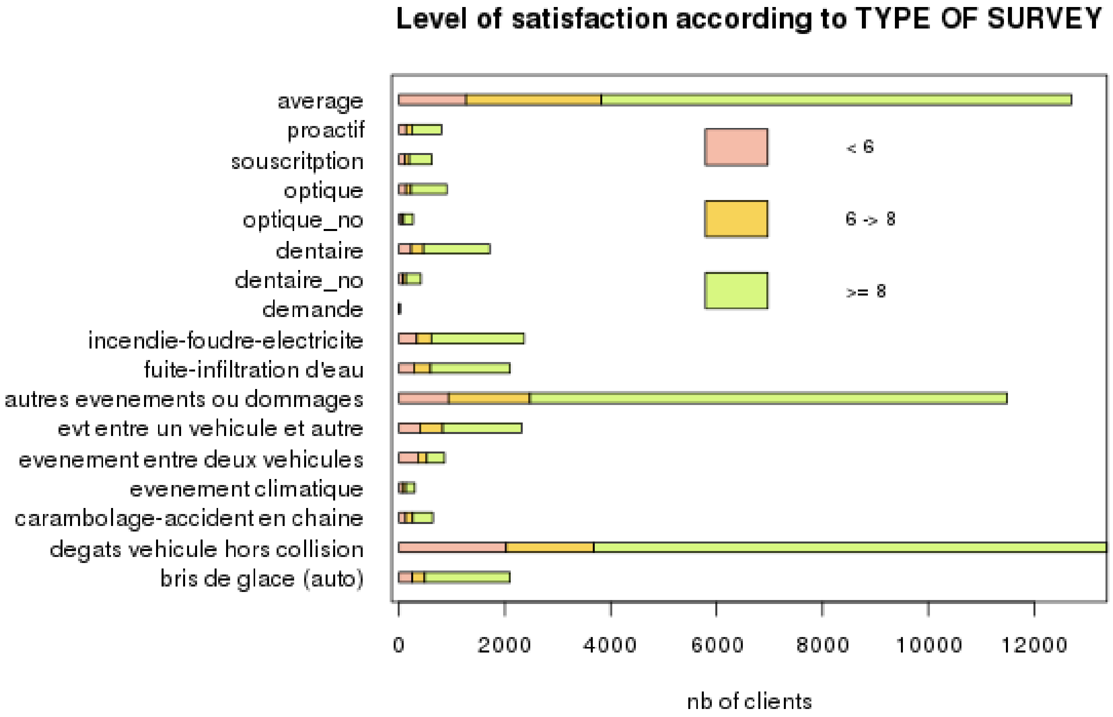


Figure 4: Level of satisfaction according to TYPE OF SURVEY

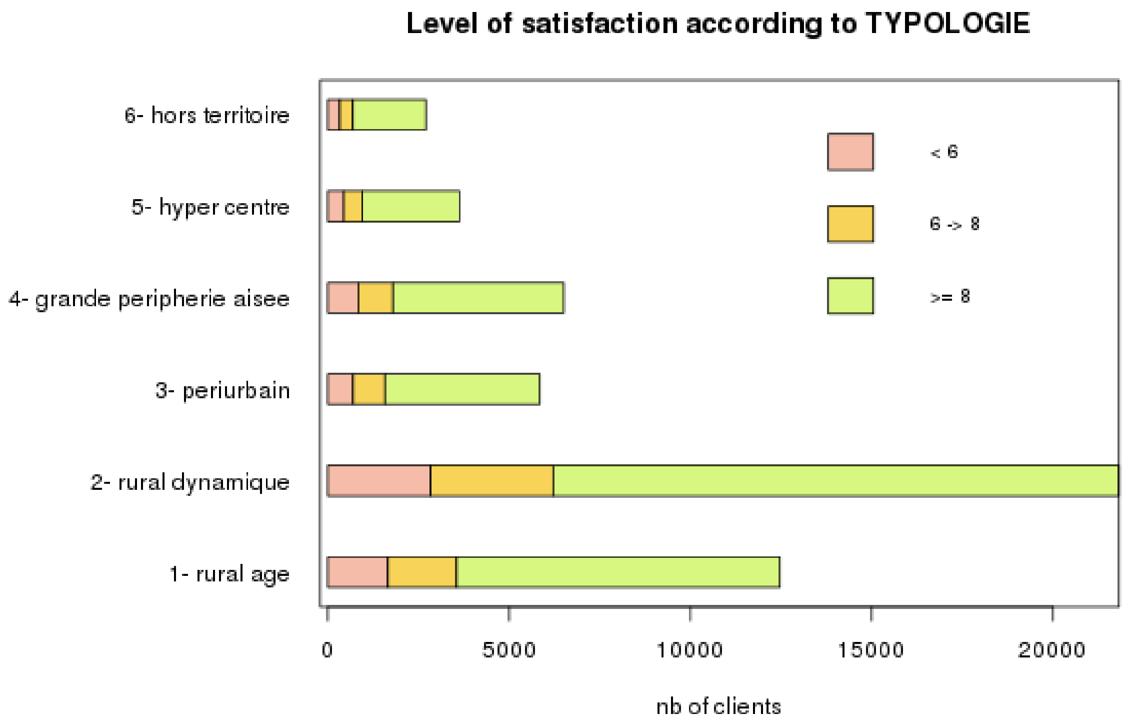


Figure 5: Level of satisfaction according to TYPOLOGIE

A.2 Nature Personne

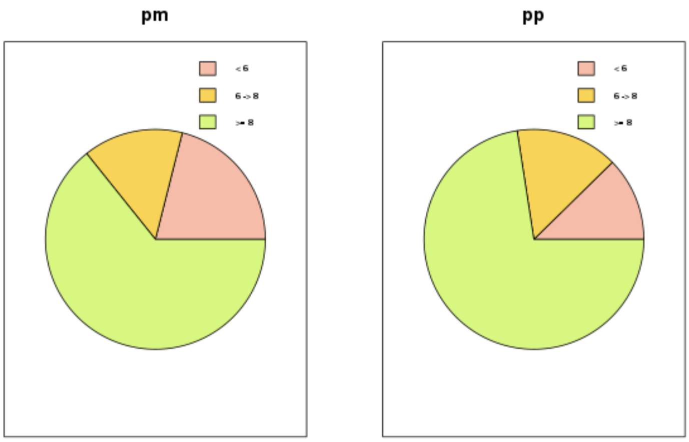


Figure 6: Level of satisfaction according to NATURE PERSONNE

A.3 Segmentation Distributive

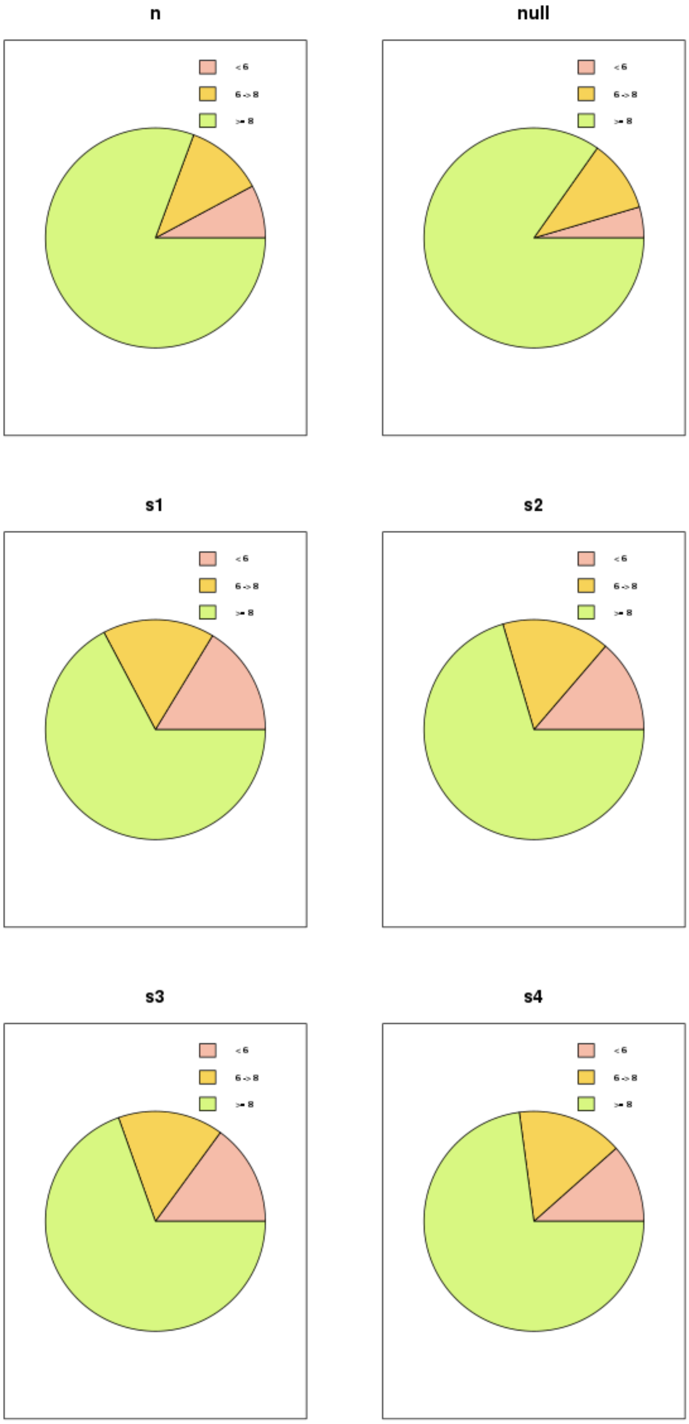


Figure 7: Level of satisfaction according to SEGMENTATION DISTRIBUTIVE

A.4 Type Survey

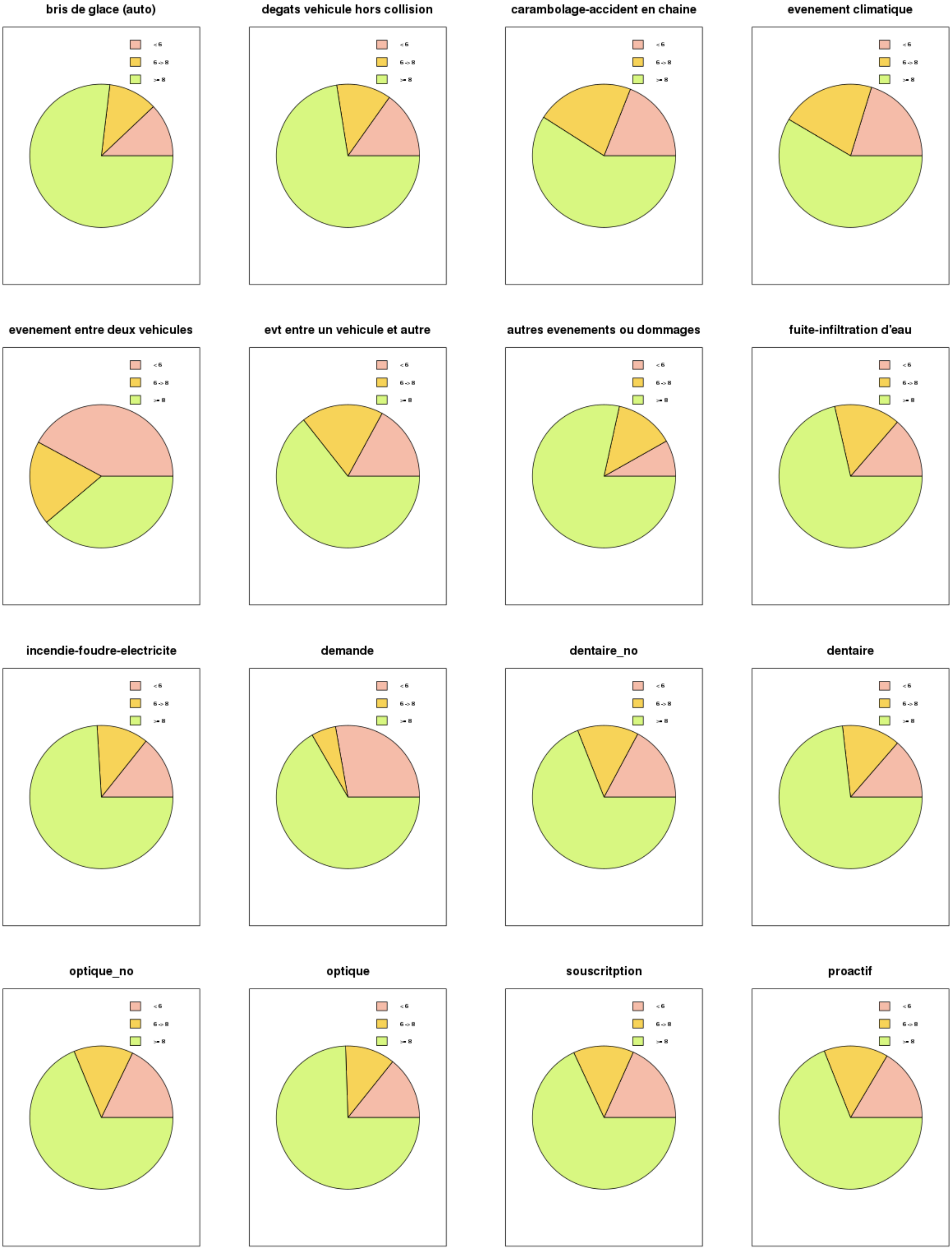


Figure 8: Level of satisfaction according to TYPE OF SURVEY

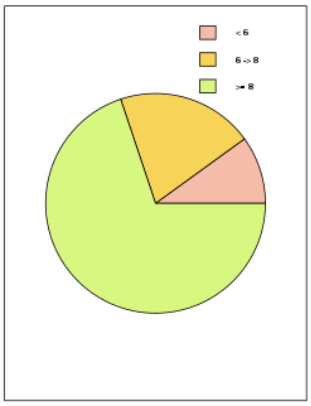
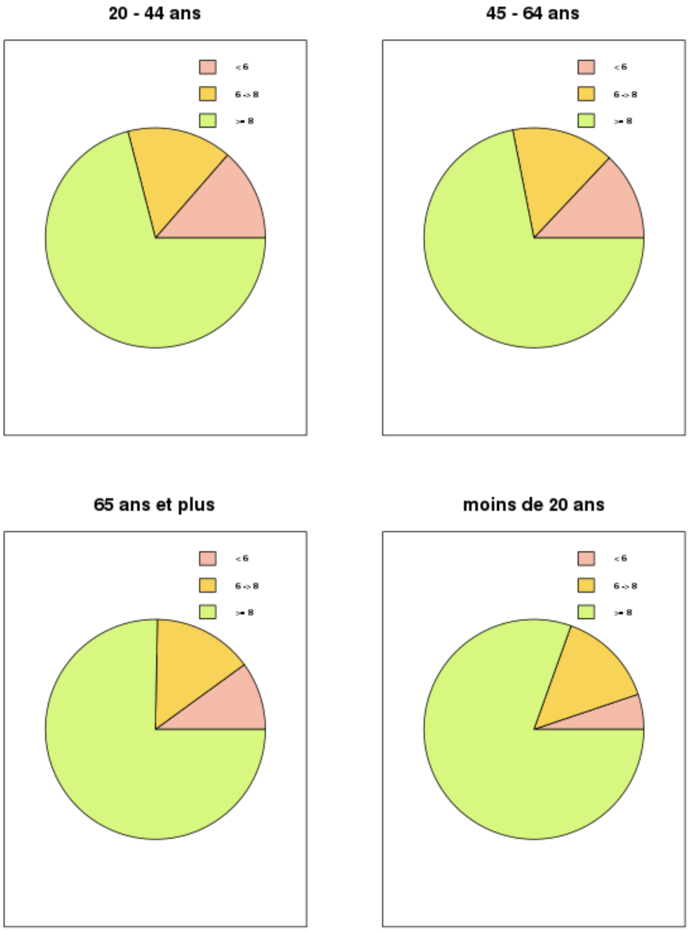


Figure 9: Level of satisfaction according to TYPE OF SURVEY - Computed Average

A.5 Tranche Age


20 - 44 ans

45 - 64 ans

65 ans et plus

moins de 20 ans

Figure 10: Level of satisfaction according to TRANCHE AGE

A.6 Typologie

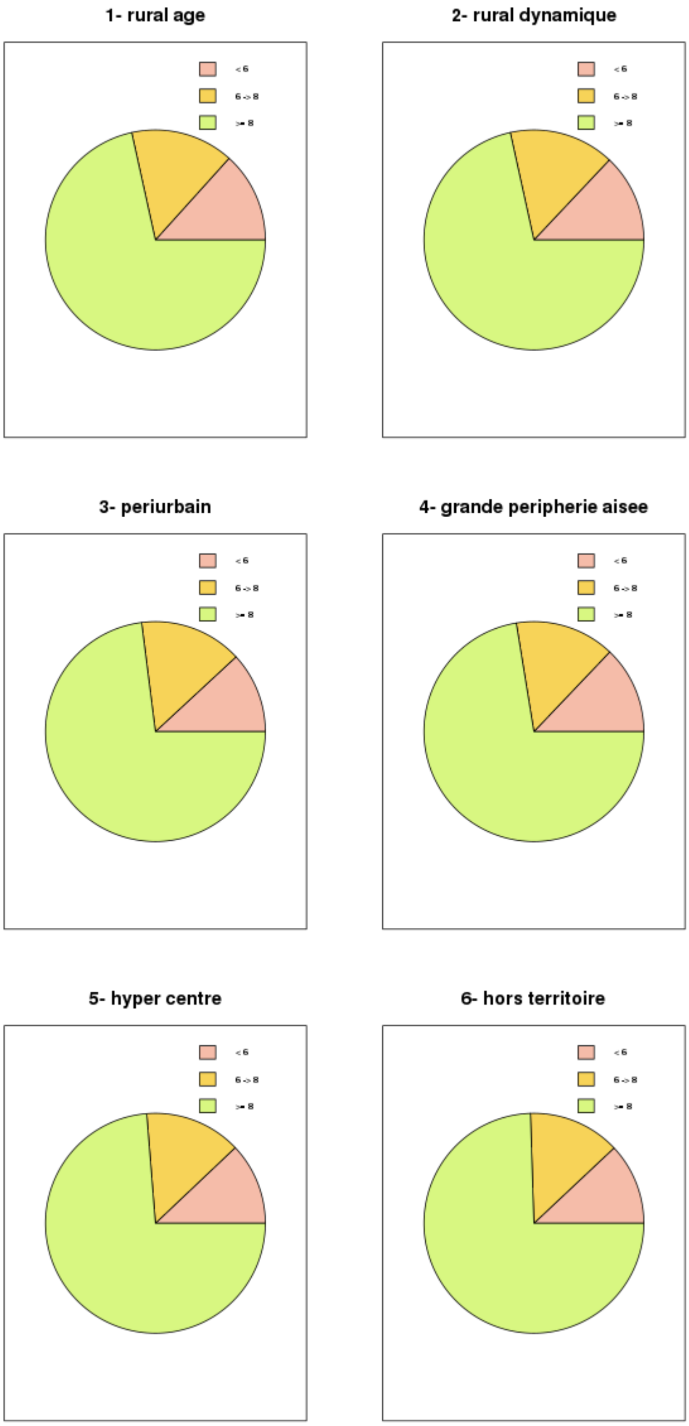


Figure 11: Level of satisfaction according to TYPOLOGIE

B Evolution of the satisfaction

B.1 According to the previous mark

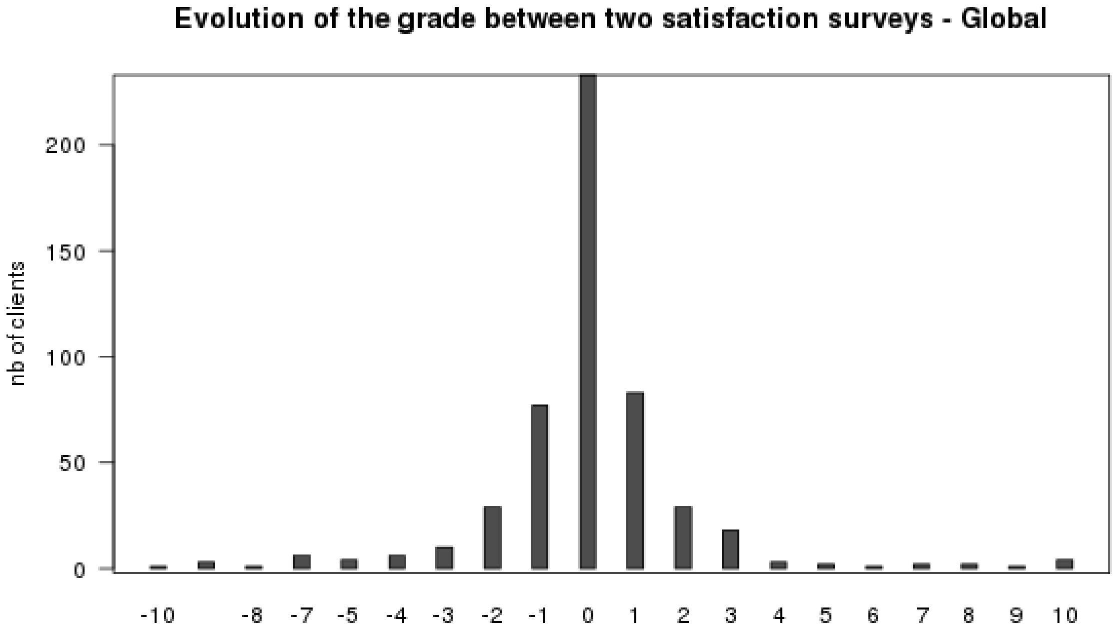


Figure 12: Evolution of the grade between two satisfaction surveys - Global

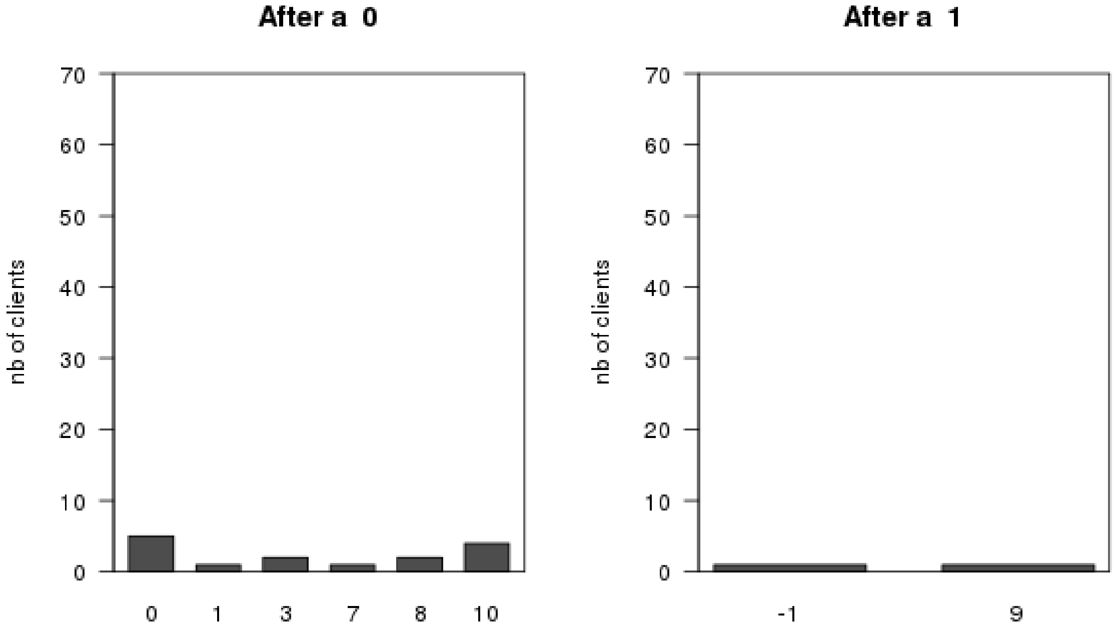


Figure 13: Evolution of the grade after a 1

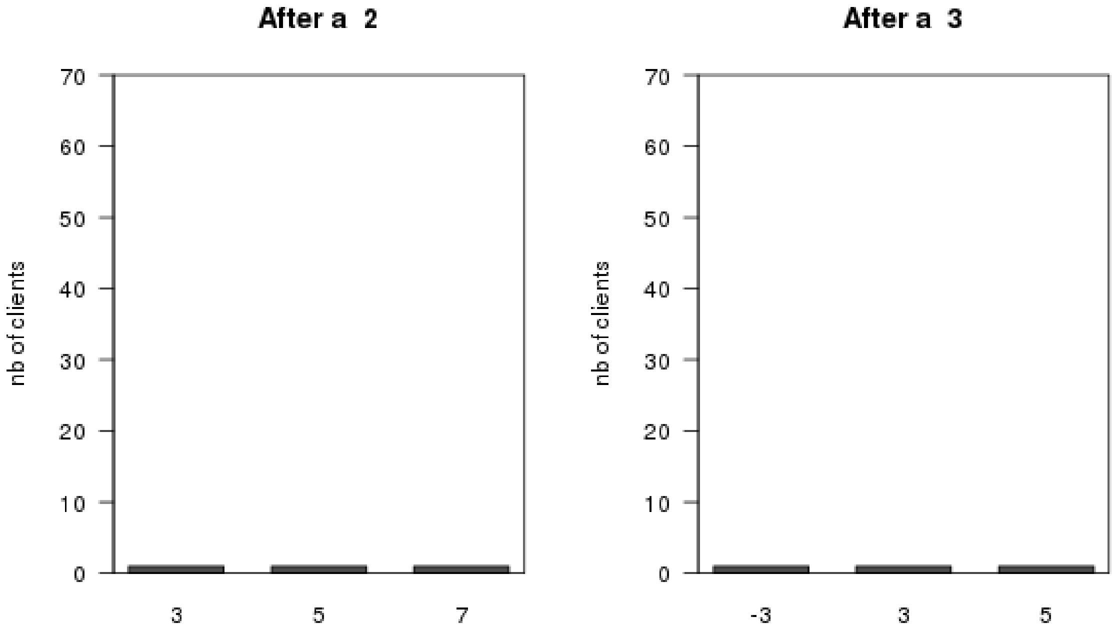


Figure 14: Evolution of the grade after a 3

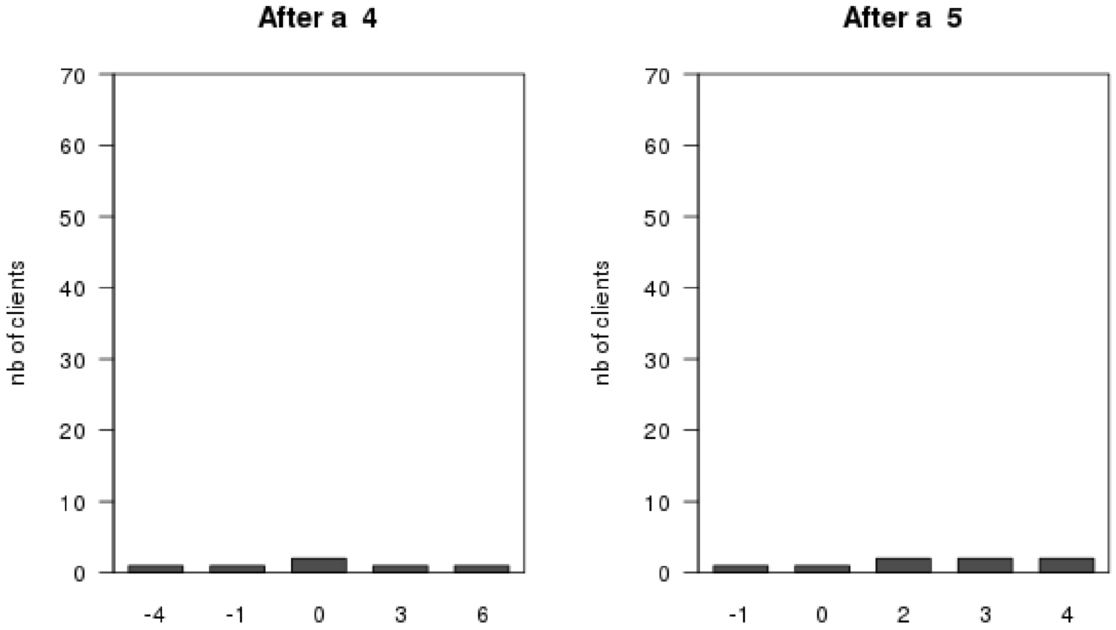


Figure 15: Evolution of the grade after a 5

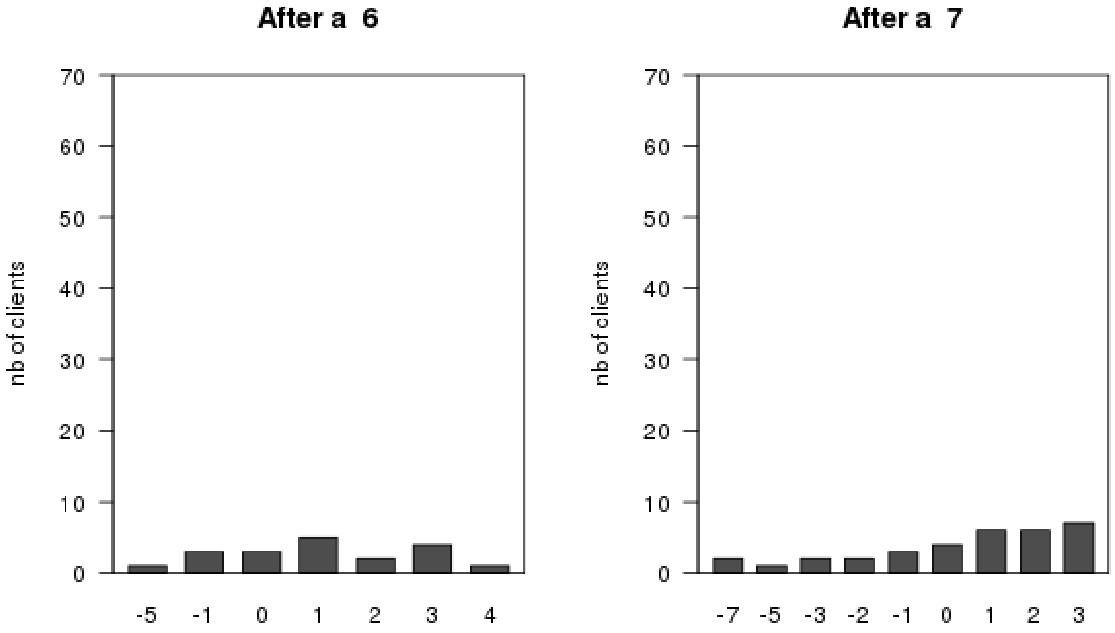


Figure 16: Evolution of the grade after a 7

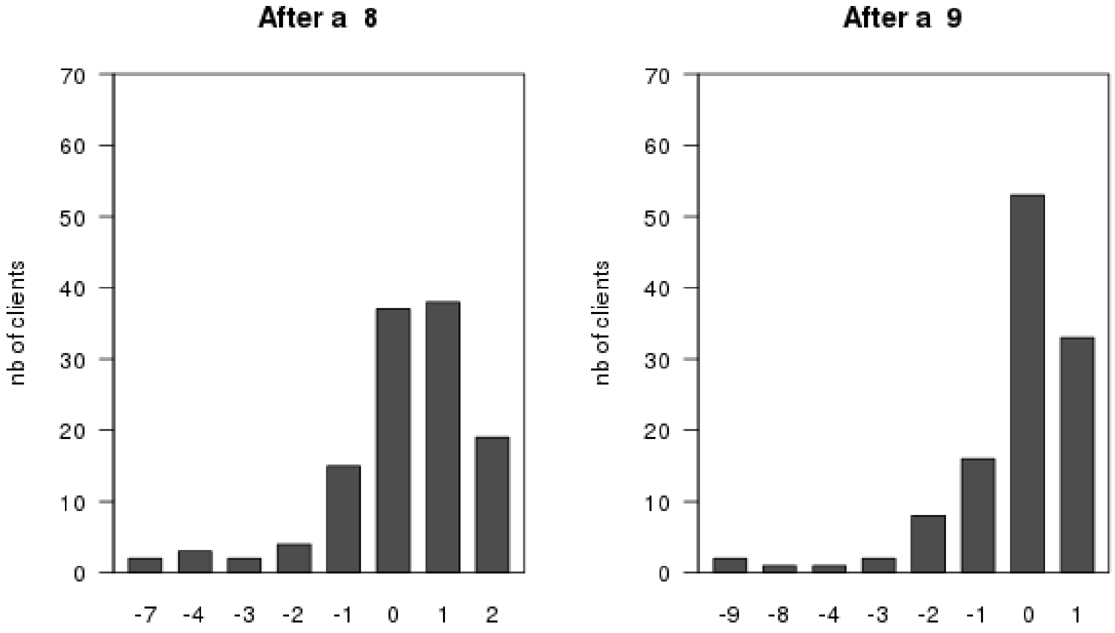


Figure 17: Evolution of the grade after a 9

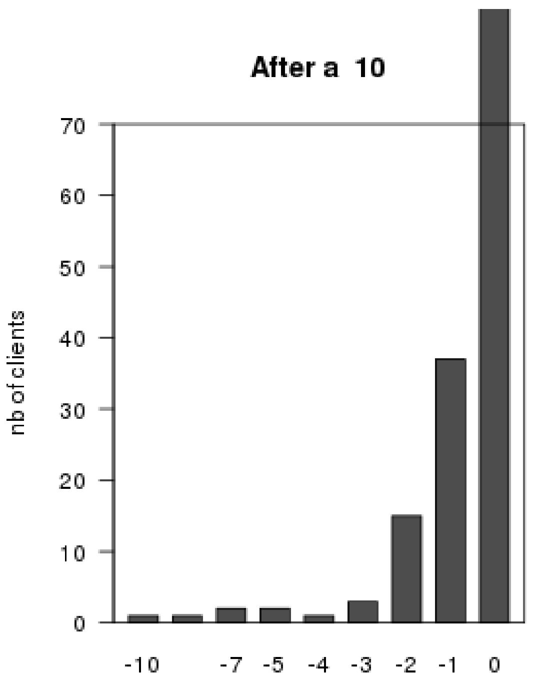


Figure 18: Evolution of the grade after a 10

B.2 According to Marche CSP

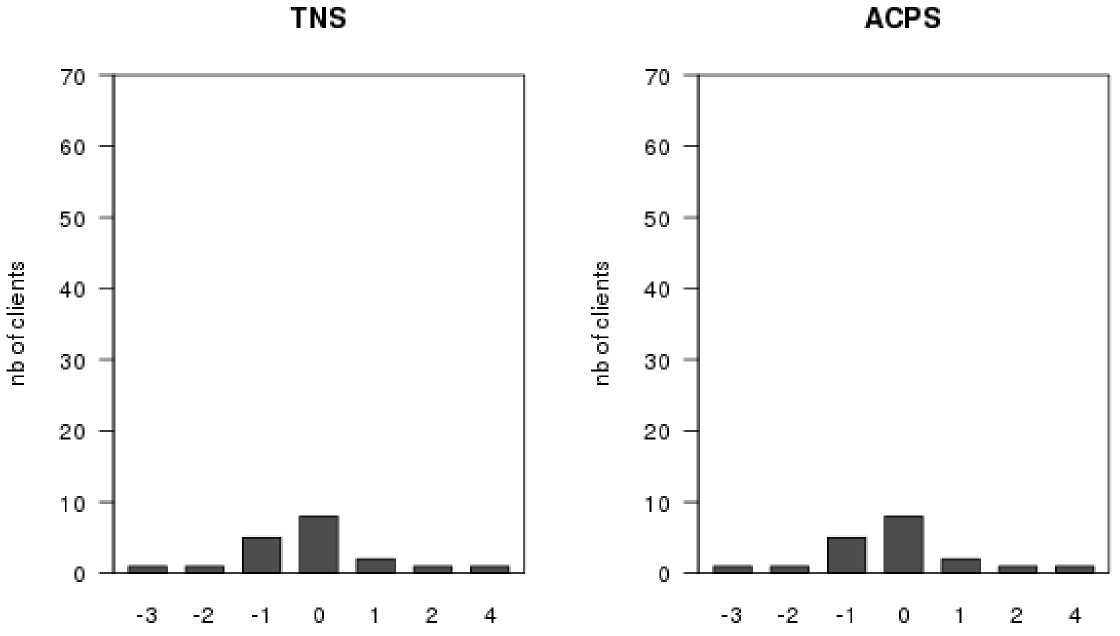


Figure 19: Evolution of the grade for marche CSP ACPS

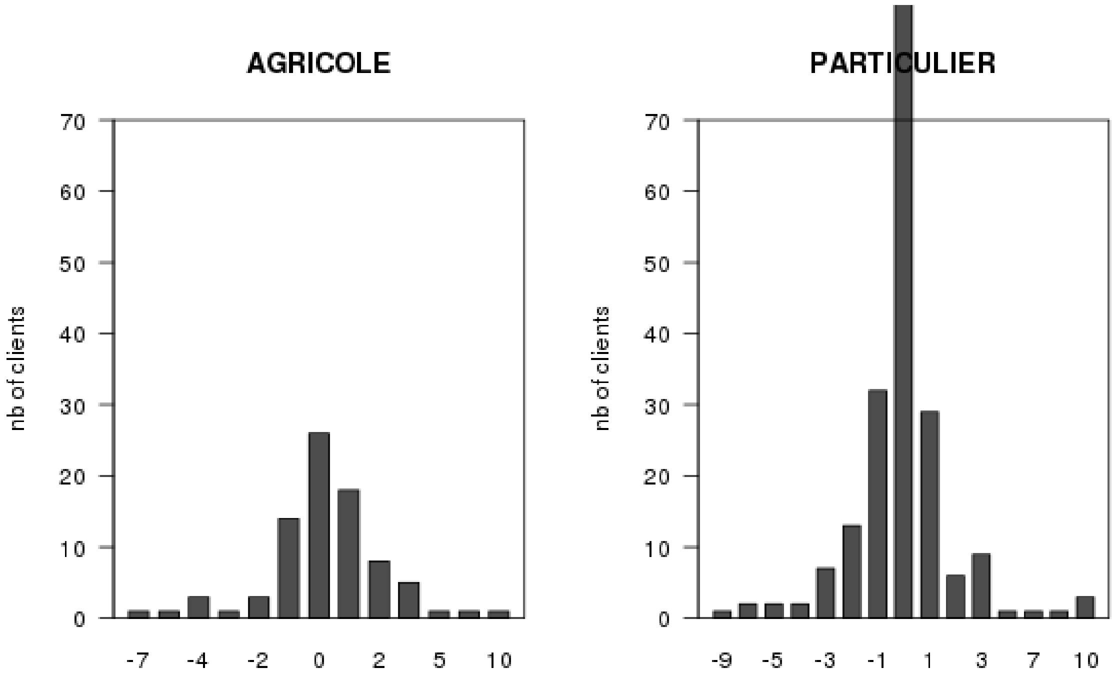


Figure 20: Evolution of the grade for marche CSP PARTICULIER

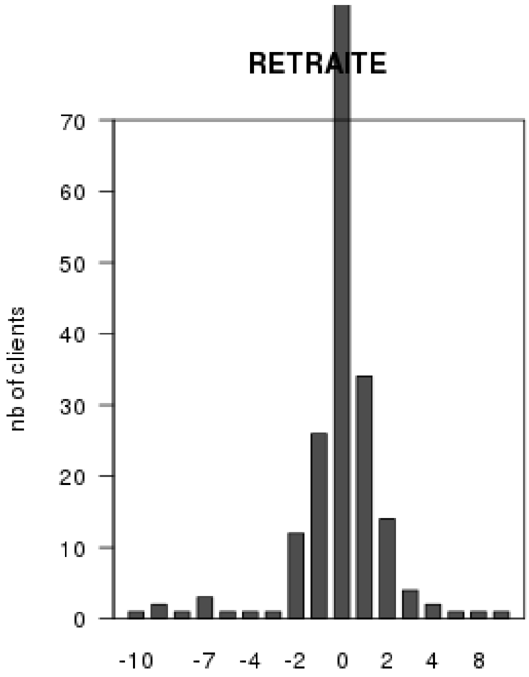


Figure 21: Evolution of the grade for marche CSP RETRAITE

B.3 According to Nature

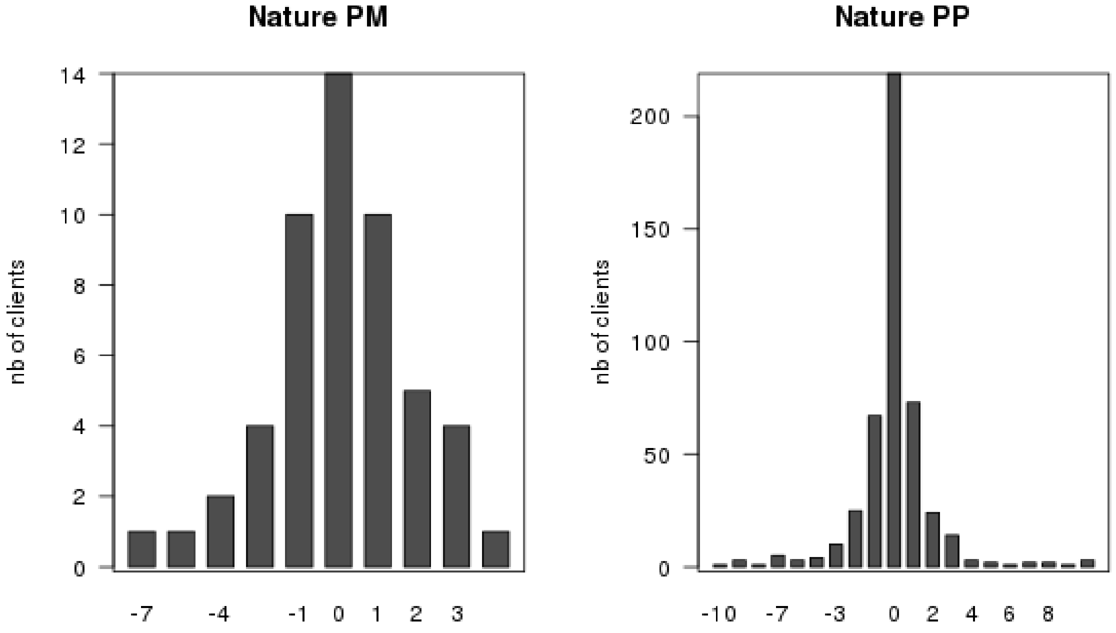


Figure 22: Evolution of the grade for nature PP

B.4 According to Segmentation Distributive

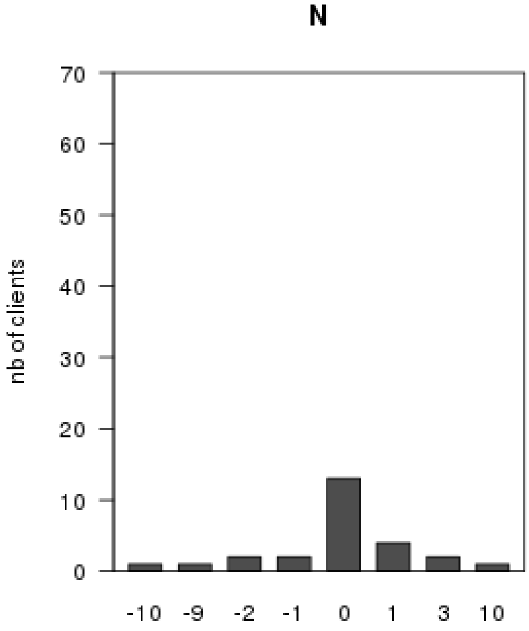


Figure 23: Evolution of the grade for segmentation N

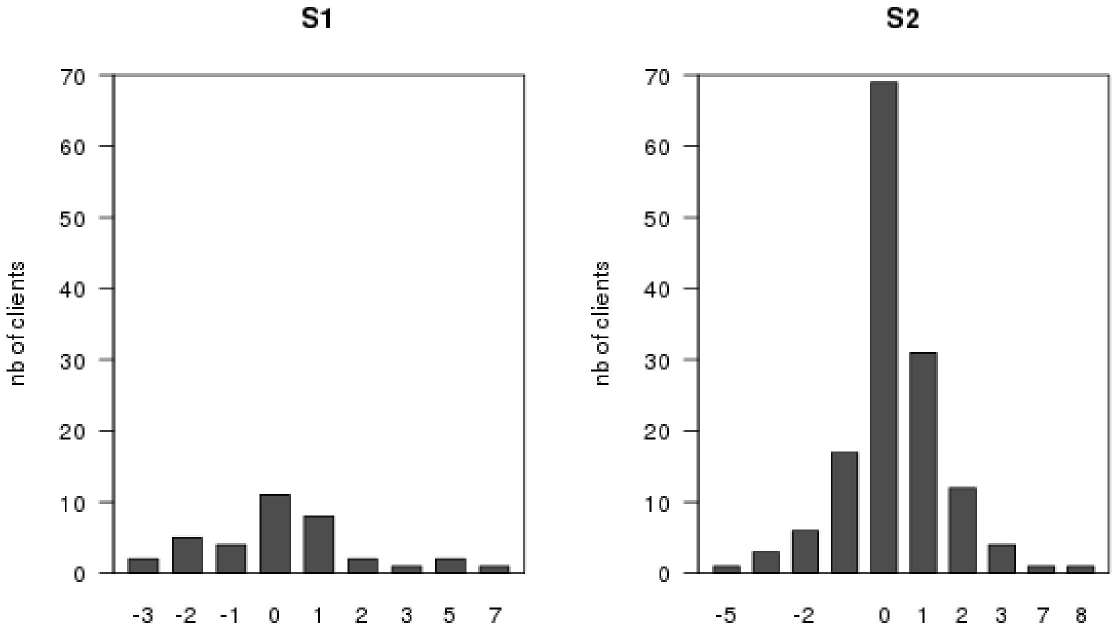


Figure 24: Evolution of the grade for segmentation S2

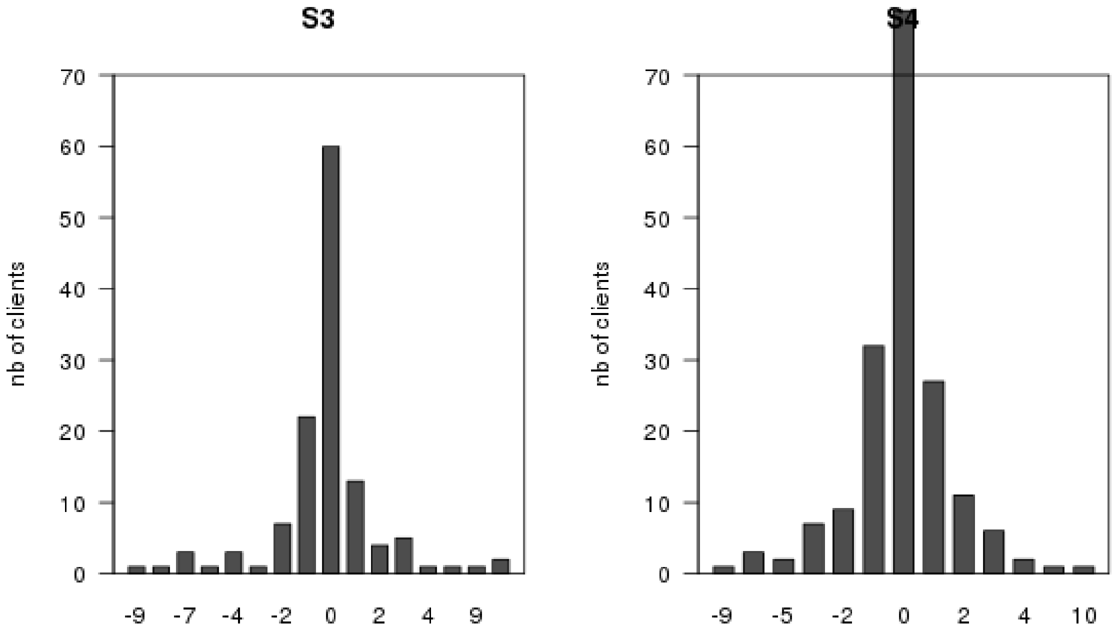


Figure 25: Evolution of the grade for segmentation S4

B.5 According to Tranche d'age

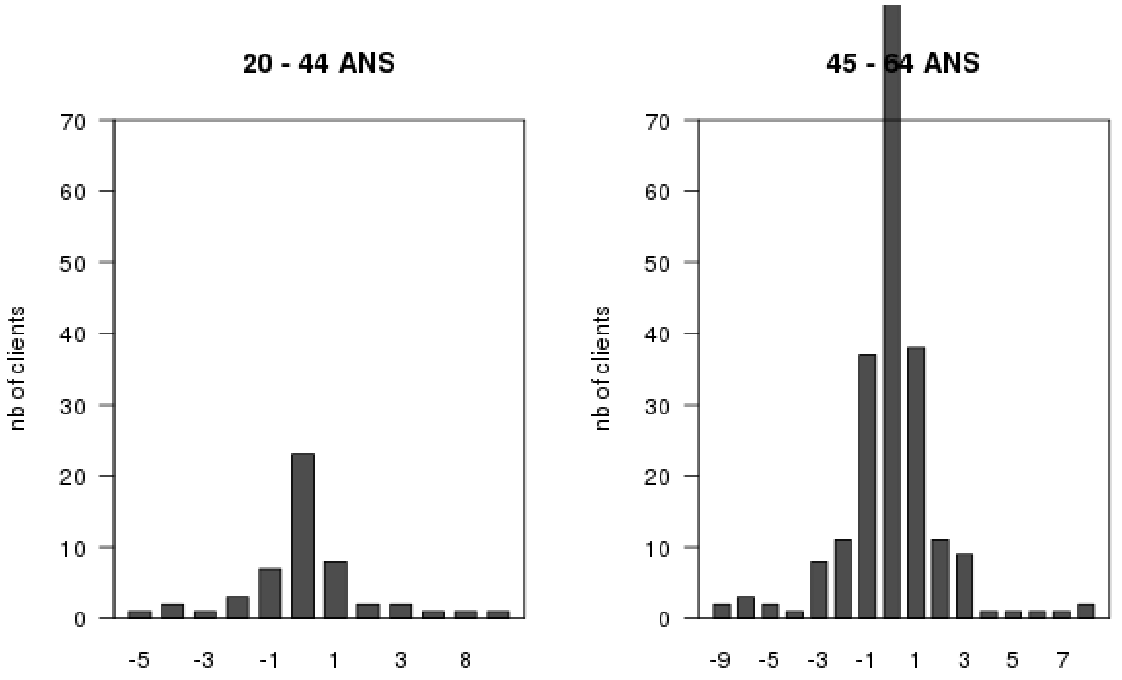


Figure 26: Evolution of the grade for tranche d age 45 - 64 ANS

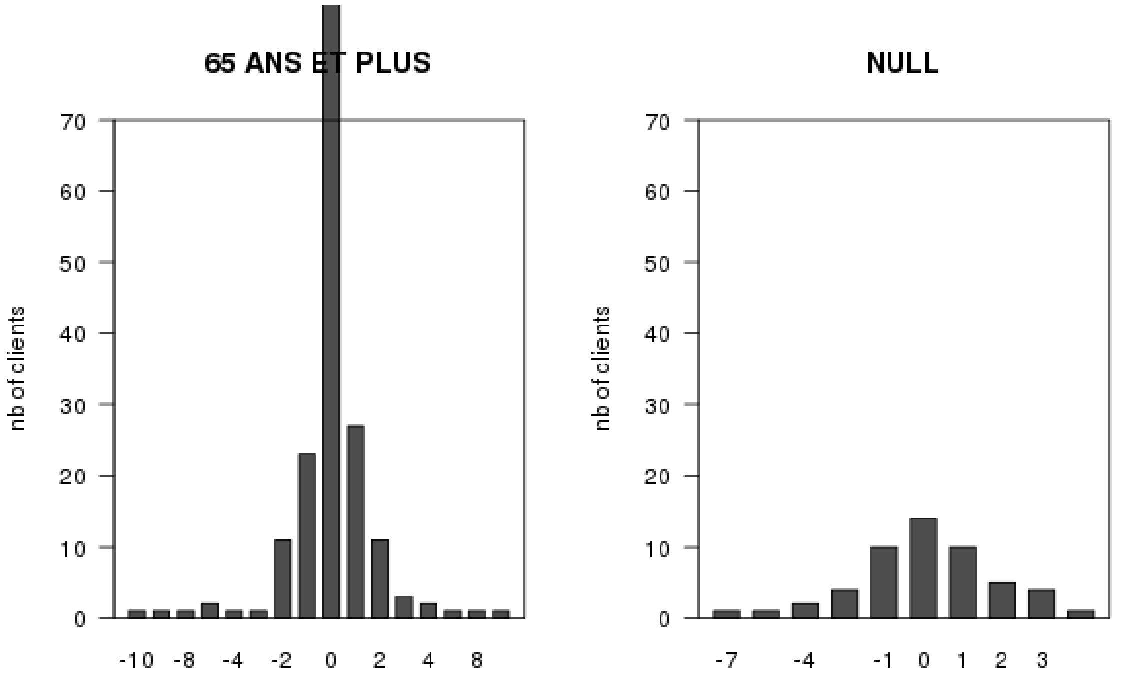


Figure 27: Evolution of the grade for tranche d age NULL

B.6 According to Typologie

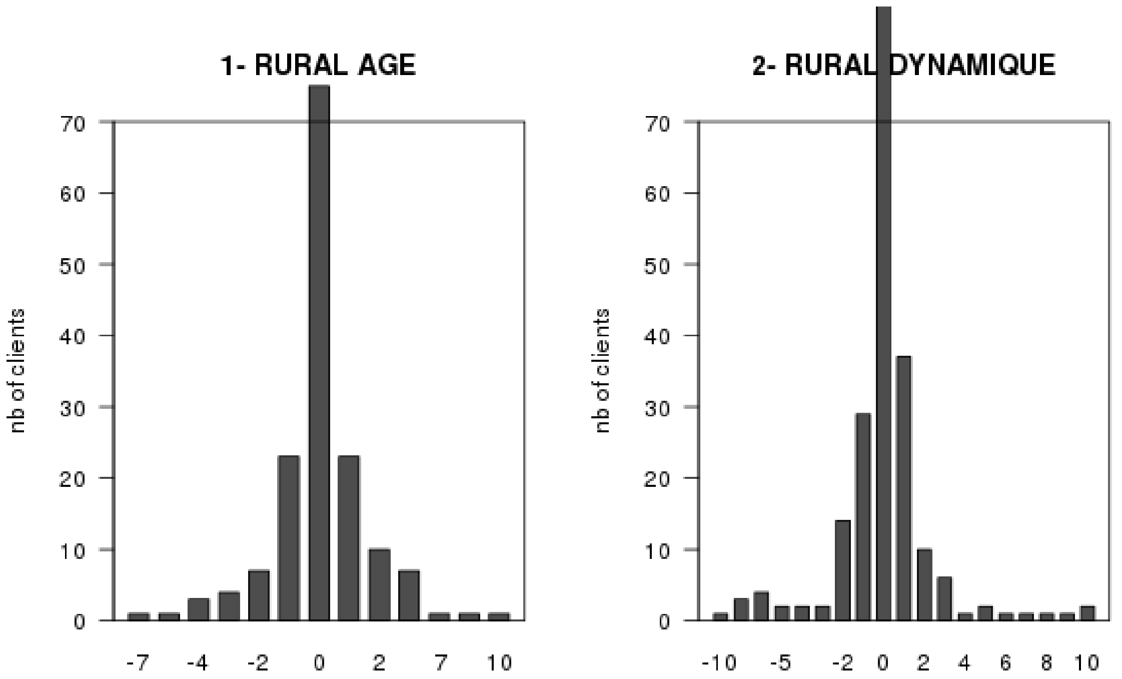


Figure 28: Evolution of the grade for typologie 2- RURAL DYNAMIQUE

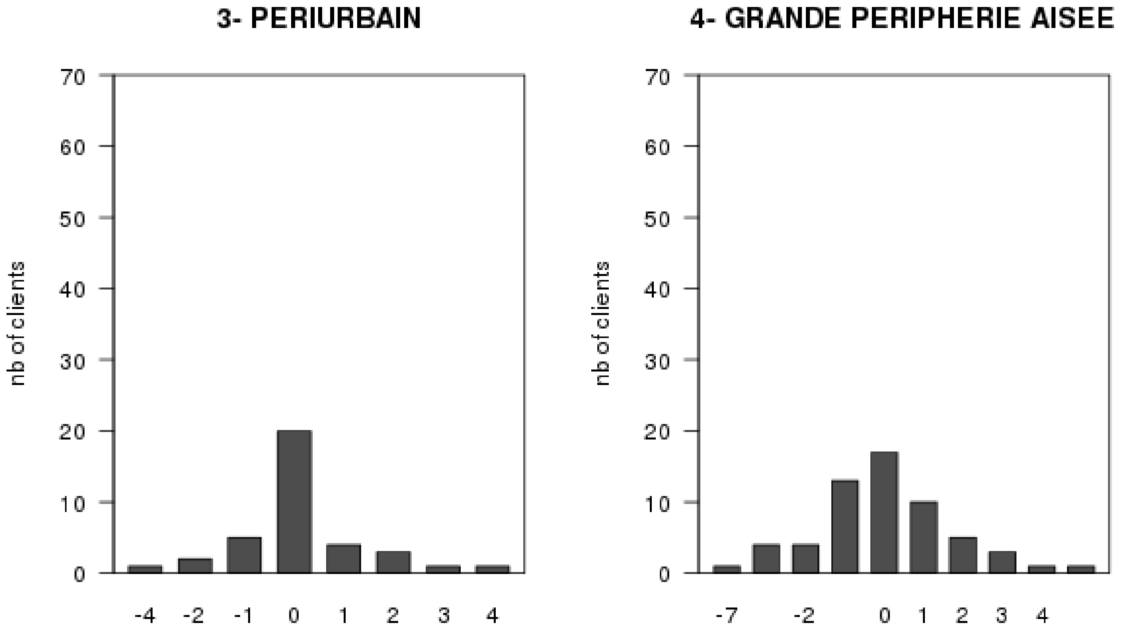


Figure 29: Evolution of the grade for typologie 4- GRANDE PERIPHERIE AISEE

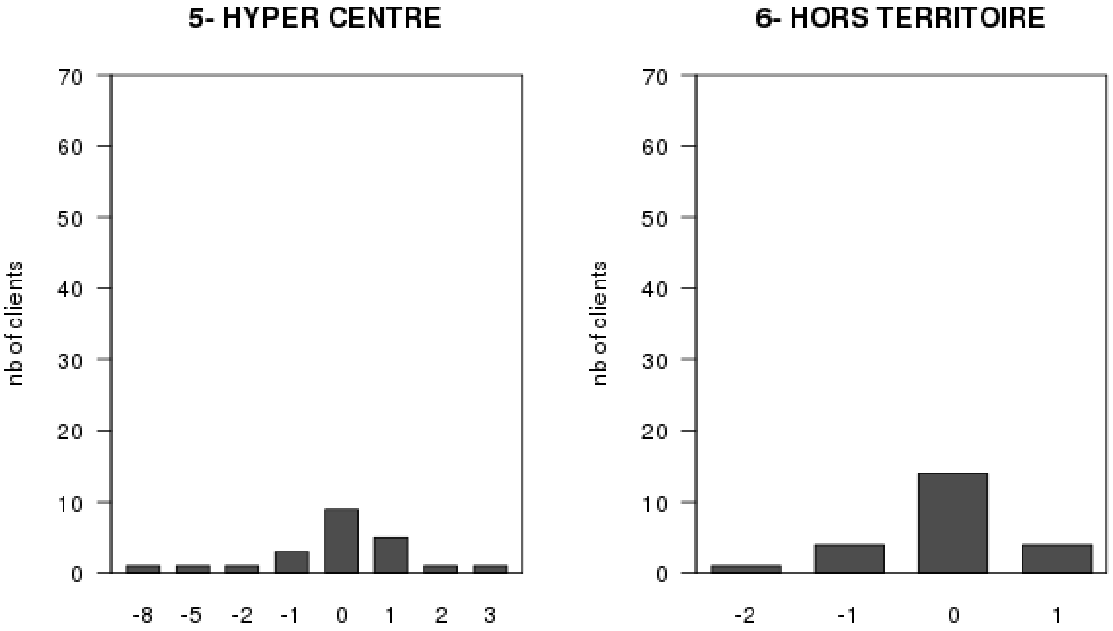


Figure 30: Evolution of the grade for typologie 6- HORS TERRITOIRE

C Reclamation & termination

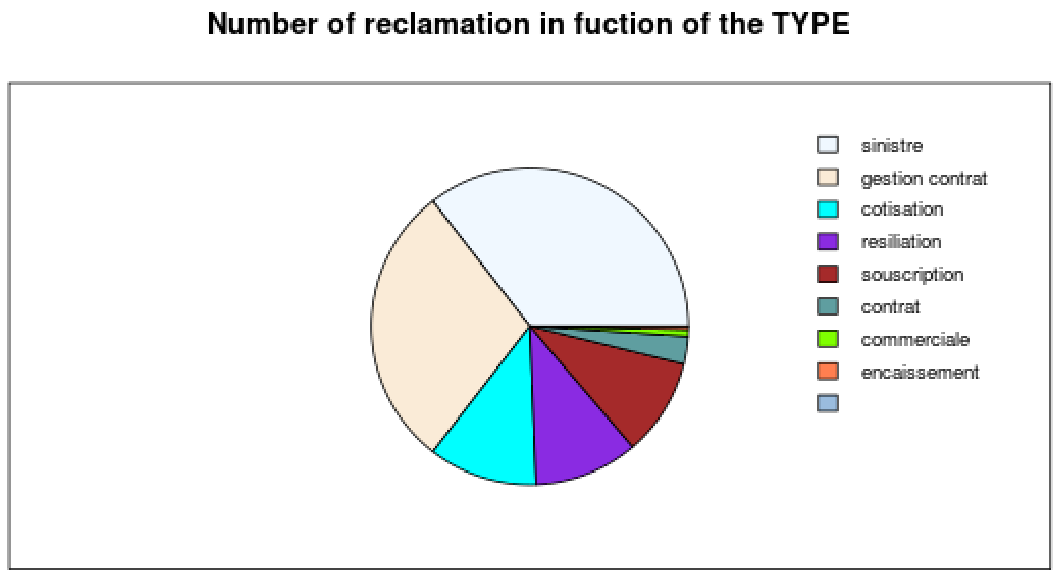


Figure 31: Number of reclamation according to their TYPE

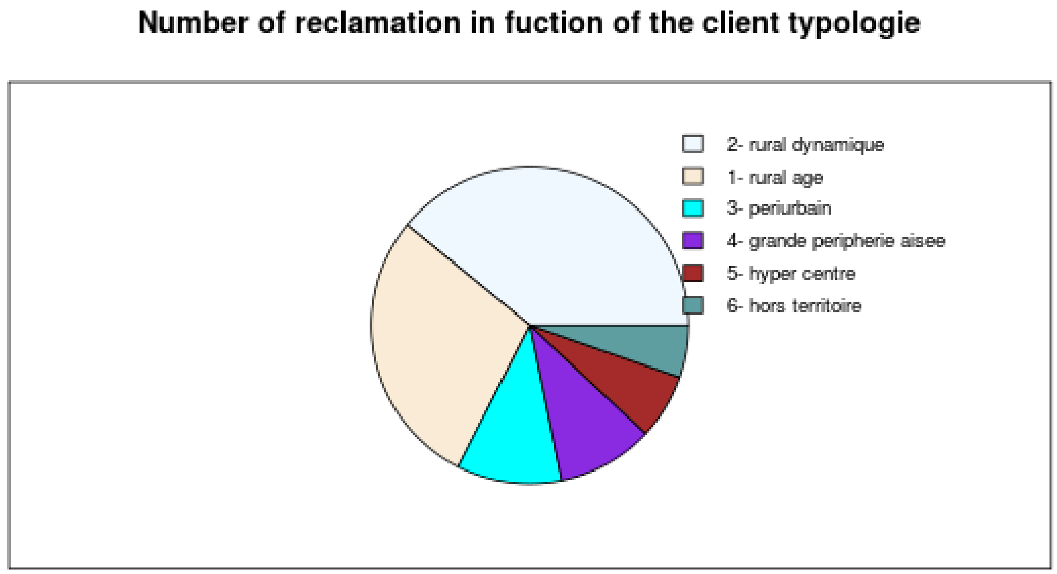


Figure 32: Number of reclamation according to the client TYPOLOGIE

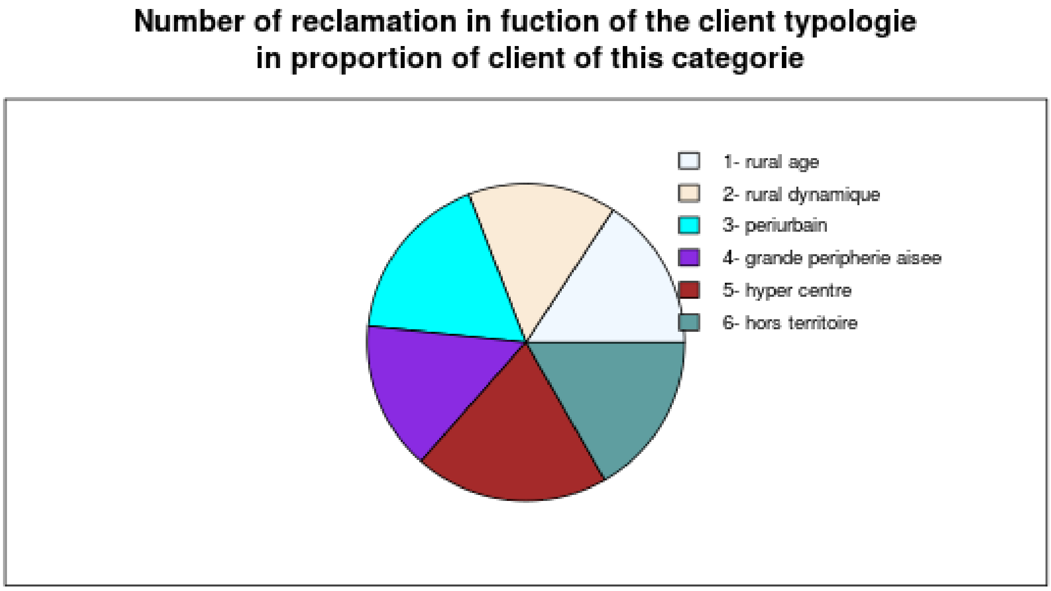


Figure 33: Number of reclamation according to the client TYPOLOGIE in proportion of the client of this categorie

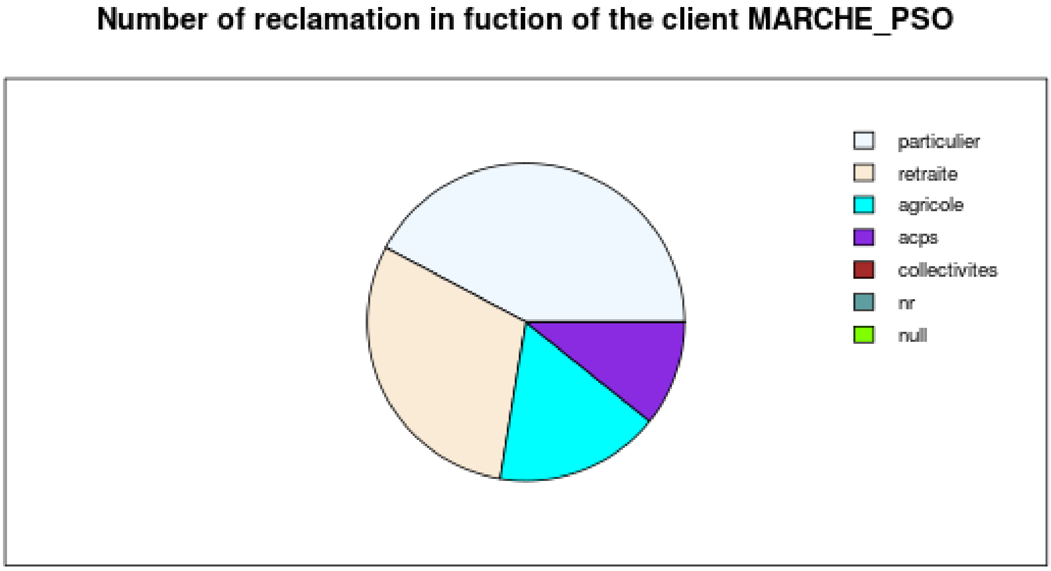


Figure 34: Number of reclamation according to the client MARCHE_PSO

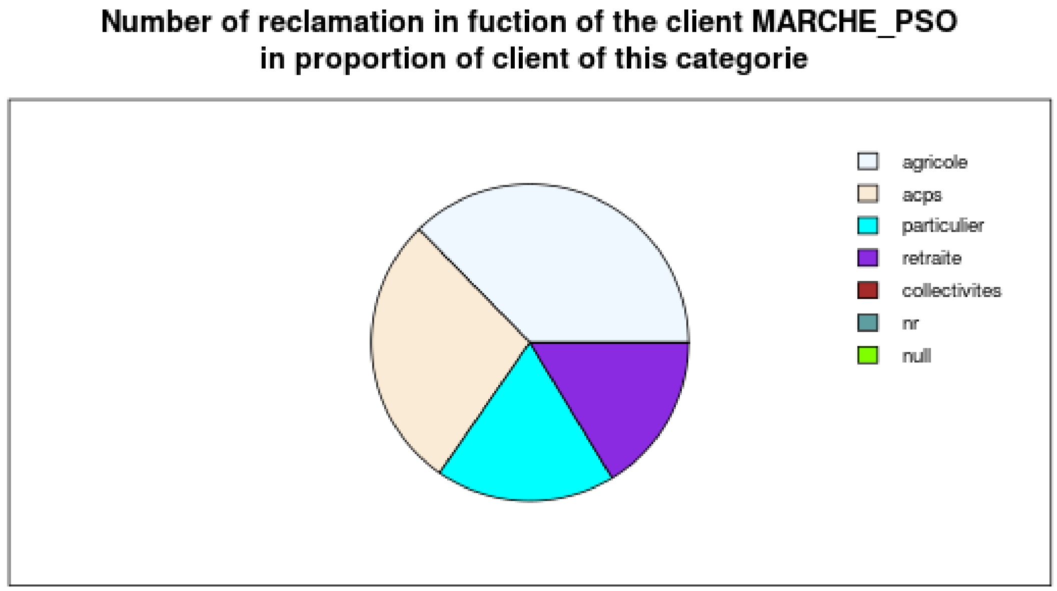


Figure 35: Number of reclamation according to the client MARCHE_PSO in proportion of the client of this category

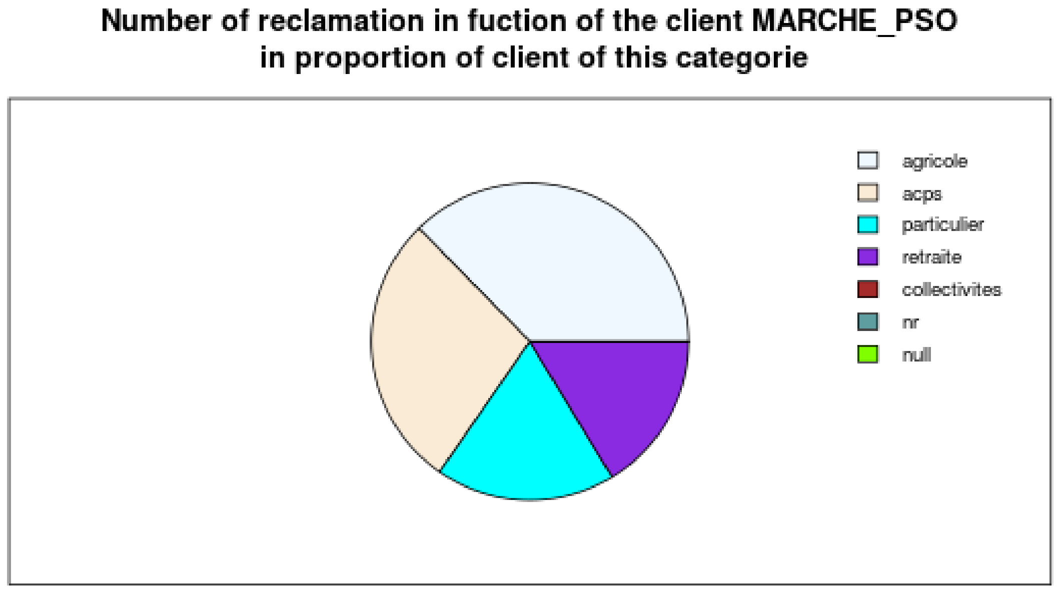


Figure 36: Number of reclamation according to the client MARCHE_PSO in proportion of the client of this category

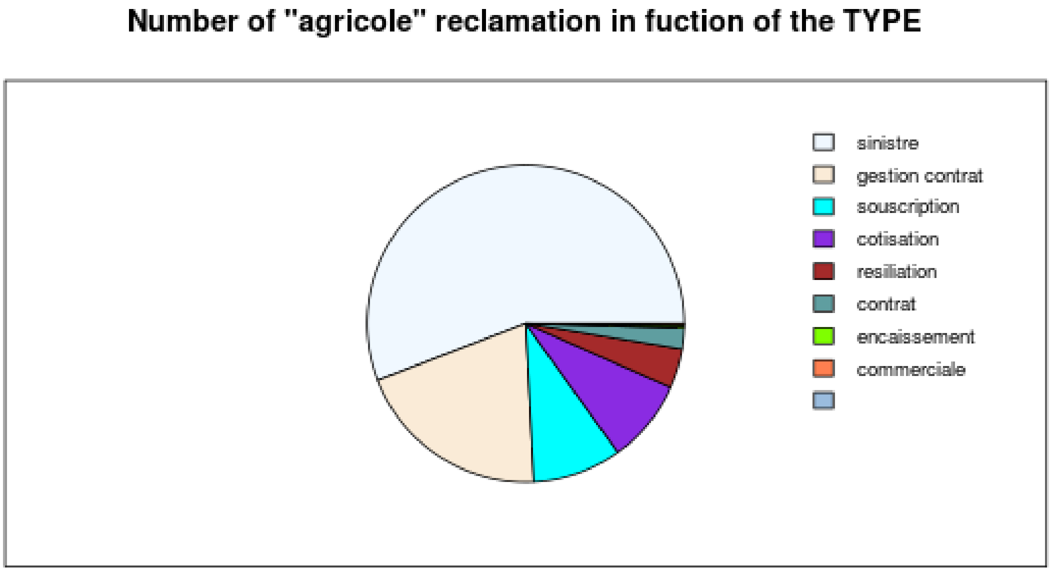


Figure 37: Number of "agricole" reclamation according to the TYPE of reclamation

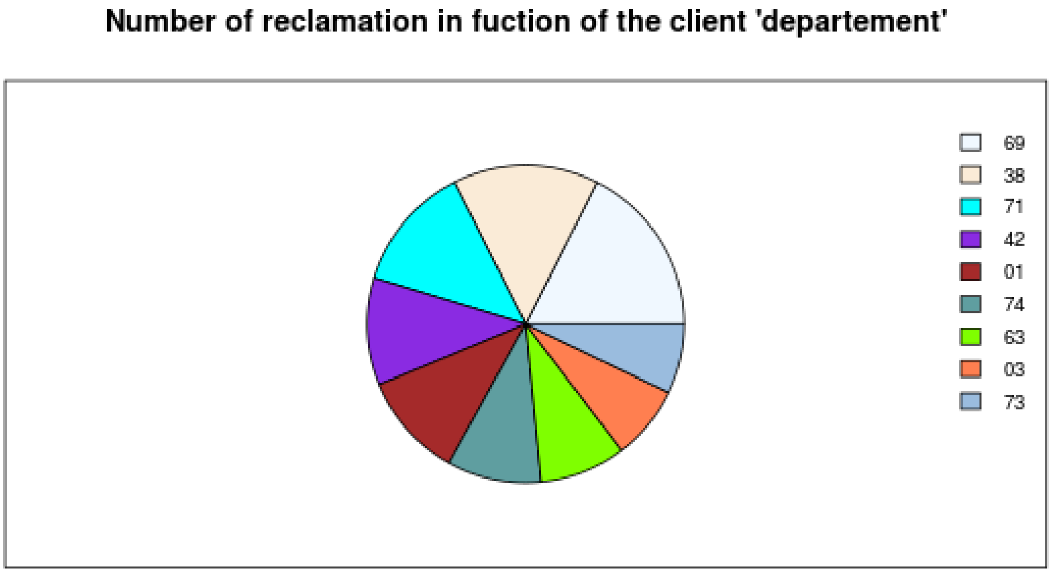


Figure 38: Number of reclamation according to the client department

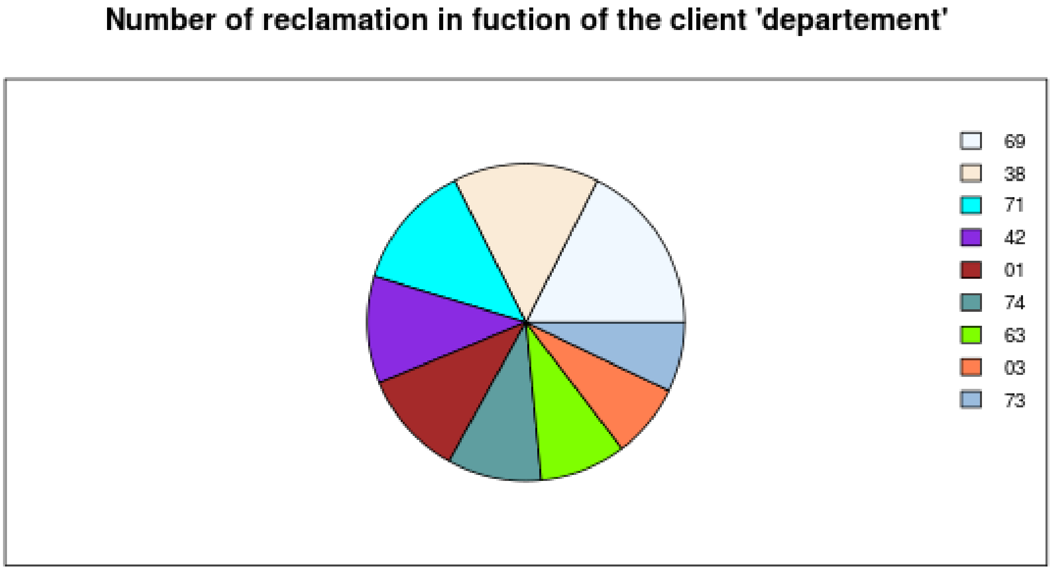


Figure 39: Number of reclamation according to the client department

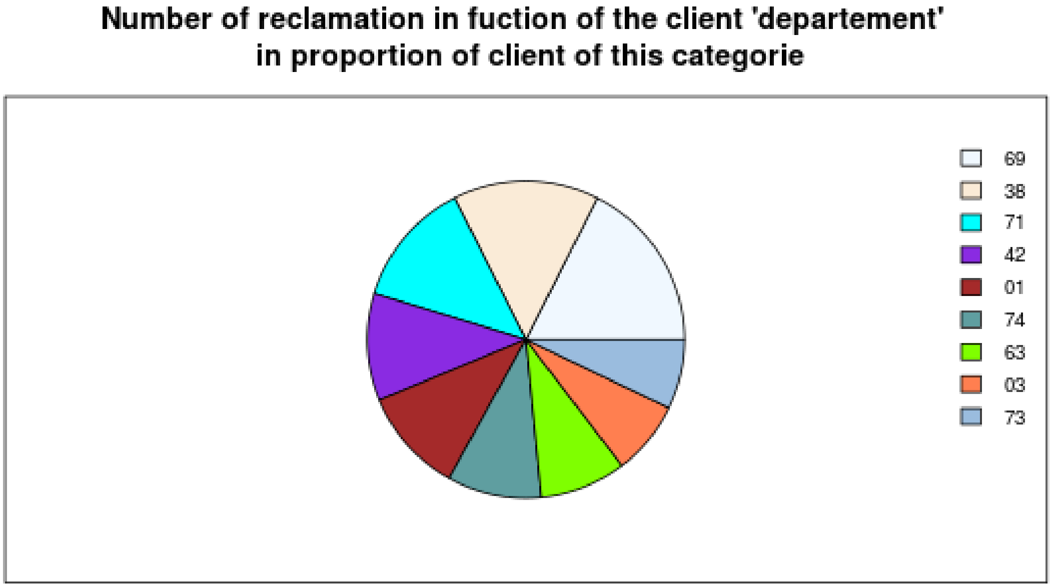


Figure 40: Number of reclamation according to the client department in proportion of the client of this category

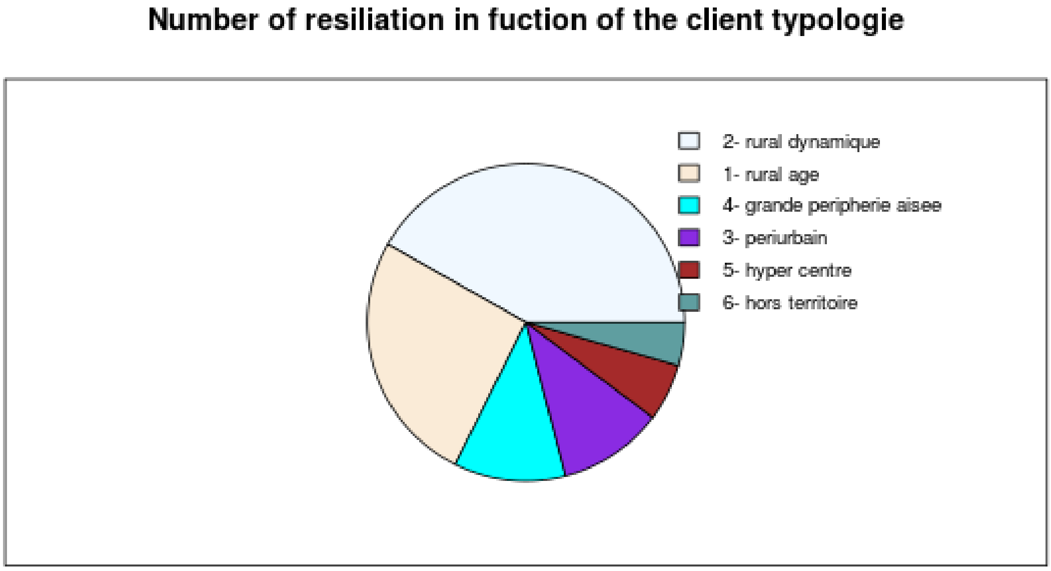


Figure 41: Number of resiliation according to the client TYPOLOGIE

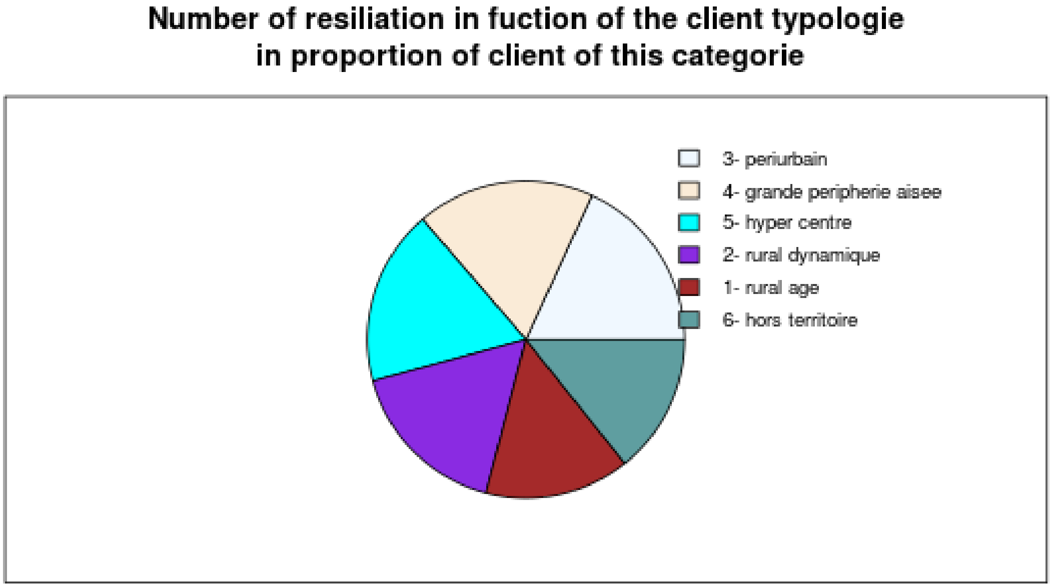


Figure 42: Number of resiliation according to the client TYPOLOGIE in proportion of the client of this category

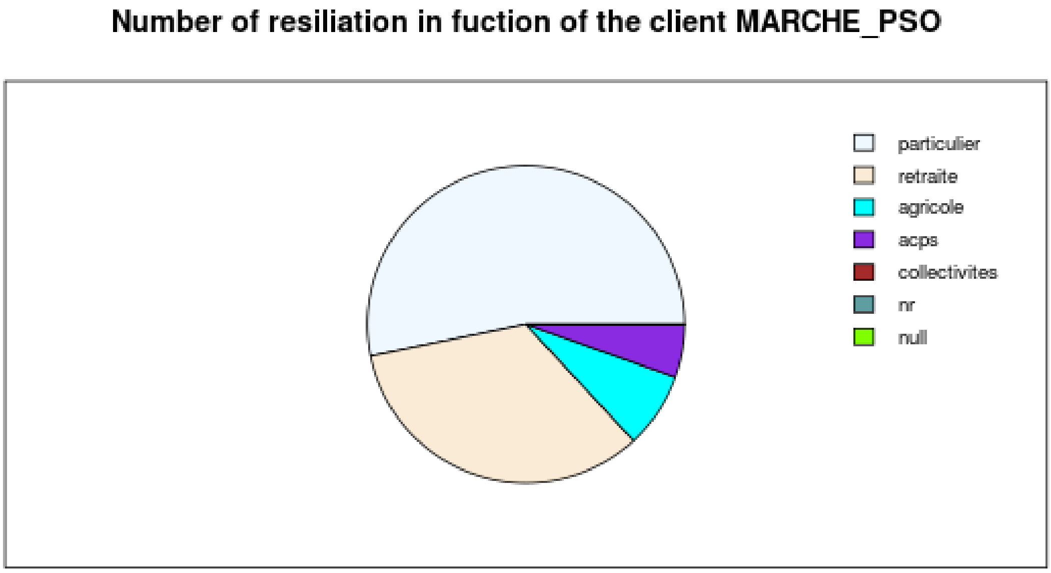


Figure 43: Number of resiliation according to the client MARCHE_PSO

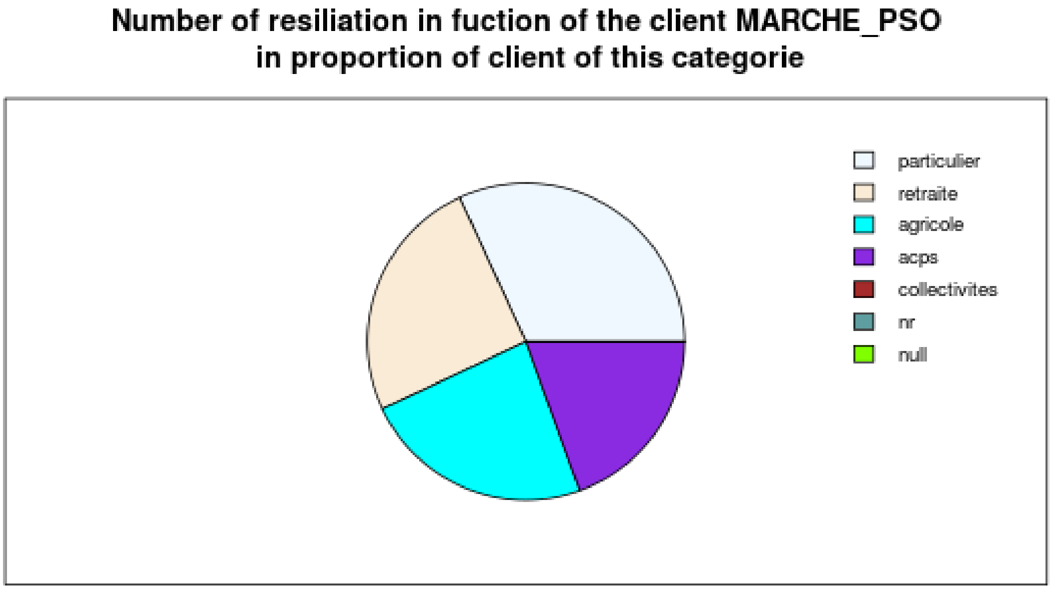


Figure 44: Number of resiliation according to the client MARCHE_PSO in proportion of the client of this category

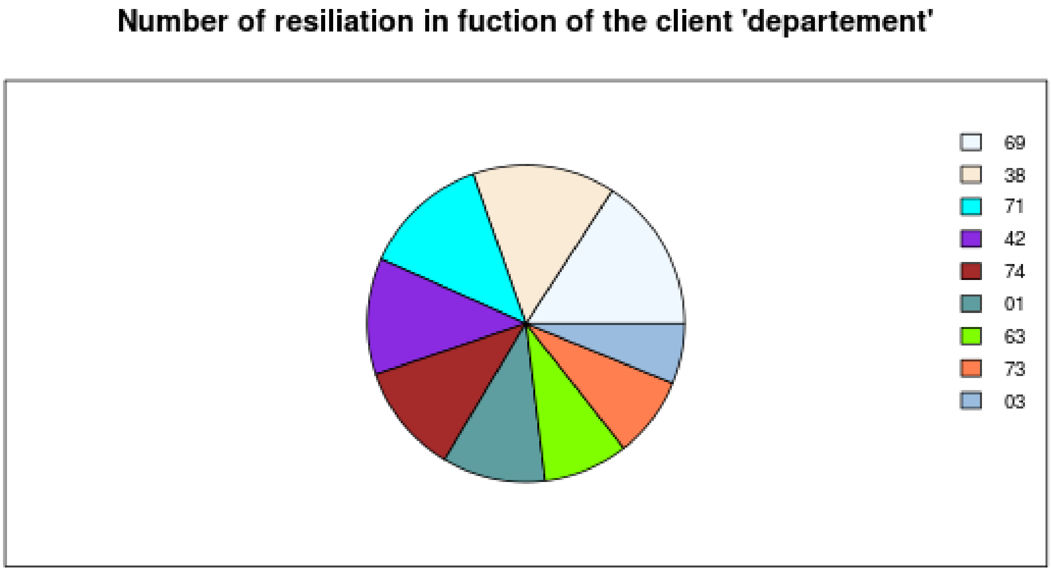


Figure 45: Number of resiliation according to the client department

D Term frequency on comments

	ngrams	nappear
1	rapid	8564
2	accueil	5712
3	efficac	3426
4	écout	2650
5	satisf	2465
6	tre	2136
7	satisfait	2056
8	réactiv	1643
9	compétent	1559
10	expliqu	1468
11	renseign	1348
12	clair	1290
13	qualit	1276
14	agréabl	1242
15	attent	1140
16	question	1091
17	personnel	1074
18	not	968
19	parf	832
20	prestat	786

	ngrams	nappear
1	alor	720
2	mois	692
3	non	585
4	apres	569
5	pai	563
6	expert	494
7	plusieur	483
8	envoi	456
9	quand	432
10	moin	403
11	tous	399
12	comm	391
13	dit	376
14	cotis	369
15	beaucoup	358
16	mutuel	358
17	euros	355
18	deux	338
19	nouvel	336
20	courri	332

Table 1: 1-grams frequencies for satisfied and unsatisfied

	ngrams	nappear
1	bon accueil	2203
2	rapid efficac	1126
3	bon conseil	826
4	tres rapid	761
5	bien reçu	706
6	tres satisf	626
7	répons rapid	569
8	tres satisfait	535
9	tre bon	526
10	bon contact	484
11	bien conseil	445
12	tres agréabl	443
13	tres professionnel	431
14	bien renseign	403
15	bon servic	380
16	trait rapid	373
17	servic rapid	361
18	satisf servic	358
19	efficac rapid	351
20	accueil bon	343

	ngrams	nappear
1	trop long	143
2	trop cher	130
3	contrat assur	125
4	moin cher	113
5	attend toujour	112
6	gest commercial	111
7	assur voitur	110
8	cel fait	108
9	plus cher	107
10	résili contrat	107
11	nouveau contrat	101
12	plusieur fois	100
13	assur auto	97
14	tous contrat	93
15	depuis plus	92
16	tres déçu	91
17	contrat chez	90
18	assur habit	88
19	beaucoup trop	87
20	suit sinistr	85

Table 2: 2-grams frequencies for satisfied and unsatisfied

	ngrams	nappear
1	tres bon accueil	961
2	bon accueil bon	300
3	pris charg rapid	265
4	bon pris charg	240
5	tres bon contact	213
6	tres bon conseil	189
7	tre bon accueil	152
8	tres bien accueil	136
9	bon accueil téléphon	134
10	tres bien conseil	129
11	tres bien renseign	129
12	tres bon servic	126
13	rapid pris charg	124
14	bon accueil agenc	120
15	accueil bon conseil	108
16	rapid trait dossi	104
17	pris compt demand	100
18	bon accueil tres	96
19	s bien pass	95
20	bon accueil expliqu	92

	ngrams	nappear
1	client depuis an	29
2	del trop long	28
3	tous contrat chez	27
4	contrat chez groupam	26
5	tout assur chez	23
6	assur tous risqu	22
7	sinistr non respons	22
8	toujour rien reçu	22
9	beaucoup trop long	21
10	appel plusieur fois	20
11	jour plus tard	20
12	non pris compt	20
13	aller voir ailleur	19
14	assur tout risqu	19
15	chang par bris	19
16	cel fait plus	18
17	résili tous contrat	18
18	mois plus tard	17
19	an assur chez	16
20	aucun gest commercial	16

Table 3: 3-grams frequencies for satisfied and unsatisfied

	ngrams	nappear
1	bon accueil bon conseil	97
2	tres bon accueil tres	78
3	tres bon accueil bon	77
4	tres bon accueil téléphon	74
5	tout s bien pass	72
6	bon accueil tres bon	62
7	tres bon accueil agenc	60
8	bon rapport qualit prix	56
9	tres bon pris charg	48
10	bon accueil expliqu clair	46
11	tres bon accueil expliqu	44
12	tout s tres bien	41
13	s tres bien pass	40
14	pris charg rapid efficac	37
15	pris charg tres rapid	35
16	bon accueil bon expliqu	33
17	tres bien reçu agenc	33
18	bon accueil bon renseign	31
19	tres bon accueil répons	31
20	bon accueil bon écout	29

	ngrams	nappear
1	tous contrat chez groupam	11
2	chez depuis plus an	8
3	bonjour mis not car	7
4	ni plus ni moin	7
5	plus an assur chez	7
6	quelqu jour plus tard	7
7	assur depuis plus an	6
8	client chez groupam depuis	6
9	suit sinistr non respons	6
10	toujour reçu cart vert	6
11	trop cher rapport concurrent	6
12	cel fait plus mois	5
13	chez groupam depuis plus	5
14	client groupam depuis an	5
15	contact fair point situat	5
16	depuis plus an chez	5
17	jour toujour rien reçu	5
18	rembours beaucoup trop long	5
19	résili tous contrat chez	5
20	agenc trop souvent ferm	4

Table 4: 4-grams frequencies for satisfied and unsatisfied

	ngrams	nappear
1	tres bon accueil tres bon	54
2	tout s tres bien pass	32
3	tres bon accueil bon conseil	25
4	bon accueil tres bon conseil	21
5	tres bon accueil expliqu clair	21
6	tres bien reçu tres bien	13
7	parc tout s bien pass	12
8	tres bon accueil bon expliqu	11
9	tres bon accueil répons rapid	11
10	bon accueil expliqu tres clair	9
11	mis not car tres bien	9
12	tres bien accueil tres bien	9
13	tres bon accueil bon écout	8
14	tres bon accueil renseign clair	8
15	tres bon accueil tres bien	8
16	bien accueil tres bien renseign	7
17	bien reçu tres bien conseil	7
18	demand pris compt tres rapid	7
19	tout s tres bien déroul	7
20	tres bon accueil téléphon rapid	7

	ngrams	nappear
1	client chez depuis plus an	3
2	client chez groupam depuis plus	3
3	goélet don silvano mor bihan	3
4	jour plus tard toujour rien	3
5	oblig aller chez partenair groupam	3
6	accueil cap larg abordag gourmet	2
7	alor client depuis plus an	2
8	appel plusieur fois avoir bon	2
9	apres avoir vain tent joindr	2
10	apres plus an chez groupam	2
11	assur auto beaucoup trop cher	2
12	assur auto trop cher rapport	2
13	assur chez depuis nombreux anné	2
14	assur chez groupam depuis an	2
15	assur depuis plus an chez	2
16	aupres conseil groupam plus proch	2
17	auto trop cher rapport concurrent	2
18	cap larg abordag gourmet entrepris	2
19	capabl fair gest commercial prendr	2
20	cel fait plus an chez	2

Table 5: 5-grams frequencies for satisfied and unsatisfied