VICAN: Relationship between socio-economic factors, geographic inequalities and breast cancer sequelae

Audrey Bergès, Swann Chelly, Inès Multrier, Ludovic Pailloux

***Abstract*—This paper analyzes possible links between socio-economic factors that distinguish French departments, and breast cancer sequelae. This study has been carried out on the VICAN French dataset gathering information about 642 women who suffer from breast cancer. Thanks to unsupervised learning techniques, we tried to find correlations between socio-economic and geographic data and breast cancer sequelae. We found two methods of clustering to perform this analysis: one based on the *t-SNE* projection and one based on the *UMAP* projection. We had difficulties picking one clustering over the other. Finally, we concluded that even though the small size of the dataset hindered our study, some of our results were encouraging: living in the North seems to favor the decrease of sequelae and living in Paris, its suburbs or Marseille seems to favor the apparition of sequelae related to aesthetics or psychology.**

Introduction

VICAN is a French dataset that gathers answers from women to the VICAN survey, all of these women having been diagnosed with breast cancer. The survey has been carried out in order to better understand the sequelae of cancer over the years.

In 2012, the French National Institute of Cancer (FNIC) has carried out VICAN 2, a survey that questioned women 2 years after their cancer diagnostic. Since one survey is not enough to understand the outcomes of such a disease, the FNIC has led another survey that questioned the same women 5 years after the detection. A total of 2009 patients responded to both VICAN 2 and VICAN 5 [1].

The following paper analyzes the subset of a data-set gathering breast cancer cases (642 women) and tries to find relationships between sequelae and socio-economic factors.

The idea of this study is to be able to predict the kind of sequelae a women is more likely to endure, on the mere basis of her department of residence. That way, we could prevent the apparition of frequent sequelae and alleviate them by setting up a specific medical support process and monitoring for each department.

In section I we analyse the dataset we were given for this study. In section II, we present the clusterings we performed and describe them using socio-economic factors and data on the patients from the VICAN dataset. In section III we perform an analysis of the evolution of sequelae for each of these clusterings in order to search for geographic or socio-economic differences.

I. VICAN Dataset & Departments Dataset

A. Description of the Datasets

In order to find links between breast cancer sequelae and socio-economic factors, we have used three datasets.

1) **”vican_sequelles”**: This is a subset of the VICAN survey gathering patients suffering from breast cancer and that have both answered to VICAN 2 and VICAN 5. This dataset contains both quantitative (age, weight, salary, family salary...)

and qualitative (presence of arm sequel, ear sequel ...) information. A total of 642 patients are identified in this dataset and they live in various French departments.

2) **”données_geo”**: This dataset gathers socio-economic information of 2011 and 2016 about French departments. It is composed of information related to economic factors (mean salary, unemployement rate...) and social factors (number of births/deaths, poverty rate...).

3) **”dep_distance”**: This dataset gathers the geographic data of each department, meaning their average latitude and their average longitude. This dataset enabled us to draw the map of France in order to better visualize our clusterings.

For the last three weeks, we have looked into these datasets in order to better understand if any relationship could be drawn between breast cancer sequelae and socio-economic factors. The guideline was first to analyse the *vican sequelles* data-set, second to create a clustering of departments, based on socio-economic data, and finally to find out any relation between such clustering and the diversity of *vican sequelles* data set.

B. VICAN Description

The VICAN dataset we worked on gathers information from 642 women on several criteria. Considering personal information, most of the women (67%) were between 18 and 50 years old when their cancer was detected and most of them did not smoke (76%). Looking at socio-economic information, 97% of the women didn’t change their department of residence from VICAN 2 to VICAN 5. 282 women (44 %) had an estate credit at the time of diagnostic and 142 (22%) had a consuming credit.

Regarding employment Figure (1) analyzes the distribution of women according to their professional social categories at the time of cancer diagnostic and their professional situation. Most of them were employed (69%) or retired (17.2%) at the moment of the diagnostic and were employees (38.4%) or in an intermediate position (26.6%). Regarding salary, the mean salary per month is 1092.45 € and the mean salary per family is 2597.43 €. The normalization of the salary in Figure (2) shows that without considering the outliers, there is no family where one parent earns much more than the other. Most of the women have been treated with surgery (99%) which was combined with radiotherapy (81.7%), hormone therapy (71,6%) and chemotherapy (57%). The sequelae are numerous but we can’t say that there is an important evolution of the overall sequelae distribution from VICAN 2 to VICAN 5. Functional, arm and pain sequelae are the most common (respectively 31%, 24% and 23% in VICAN 2 and 33%, 26% and 24% in VICAN 5).

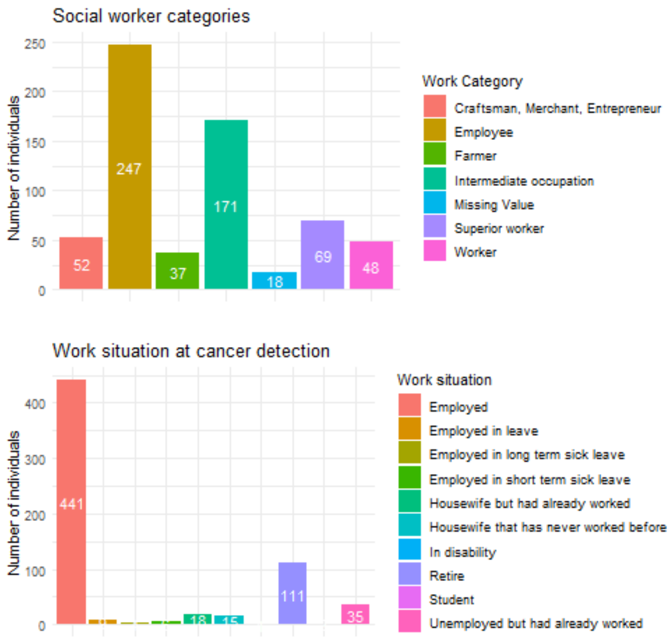


Fig. 1. Top - Social Worker categories. Bottom - Work situation at cancer detection

Overall, from this preliminary analysis, we deduced that the women from VICAN 2 and VICAN 5 had regular socio-economic characteristics: their distribution approaches that of the average population in France. This means that these properties will not influence our study.

We will now cluster the French departments based on the methods we have learned in class, in order to study the distribution of sequelae across these clusters and infer potential geographical inequalities.

II. Department Clustering

There are 101 departments in France. Almost all departments are being represented by at least one individual but there are not enough samples in this dataset to analyze socio-economic impact per department. Therefore, we decided to use the table *donnees_geo* that gathers socio-economic data per department in order to clusterize them.

Before executing these different methods of clustering, we preprocessed the data. The four last departments of

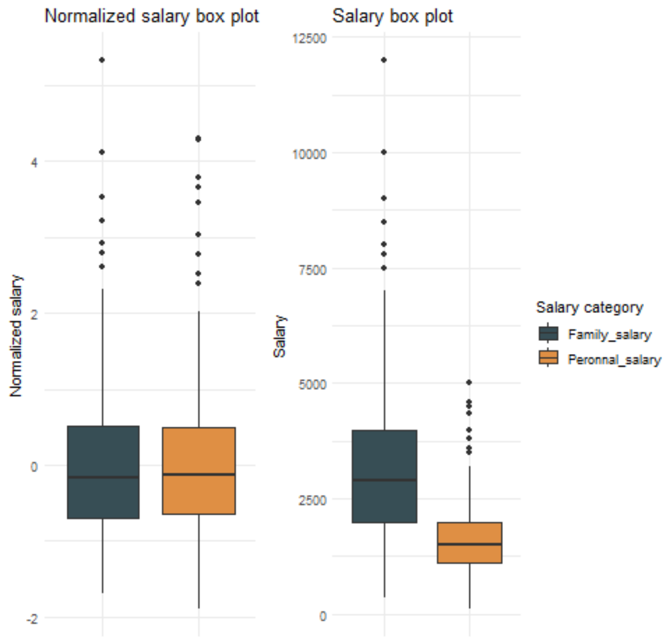


Fig. 2. Left - Salary Boxplot. Right - Normalized salary boxplot

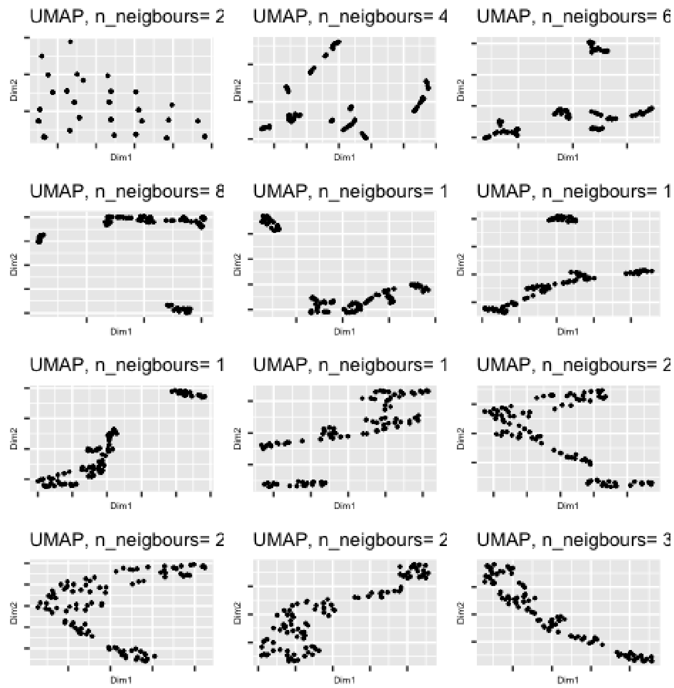


Fig. 3. Plot of the different 2D-projections using the *UMAP* method, with different number of neighbours

the database *donnees_geo* had missing values. In addition, there was no patient at all coming from these departments in the database *vican_sequelles*. Therefore, we decided not to take into account these departments and deleted them from the database. These four departments were overseas ones such as Guadeloupe, Martinique, French Guyana and Reunion.

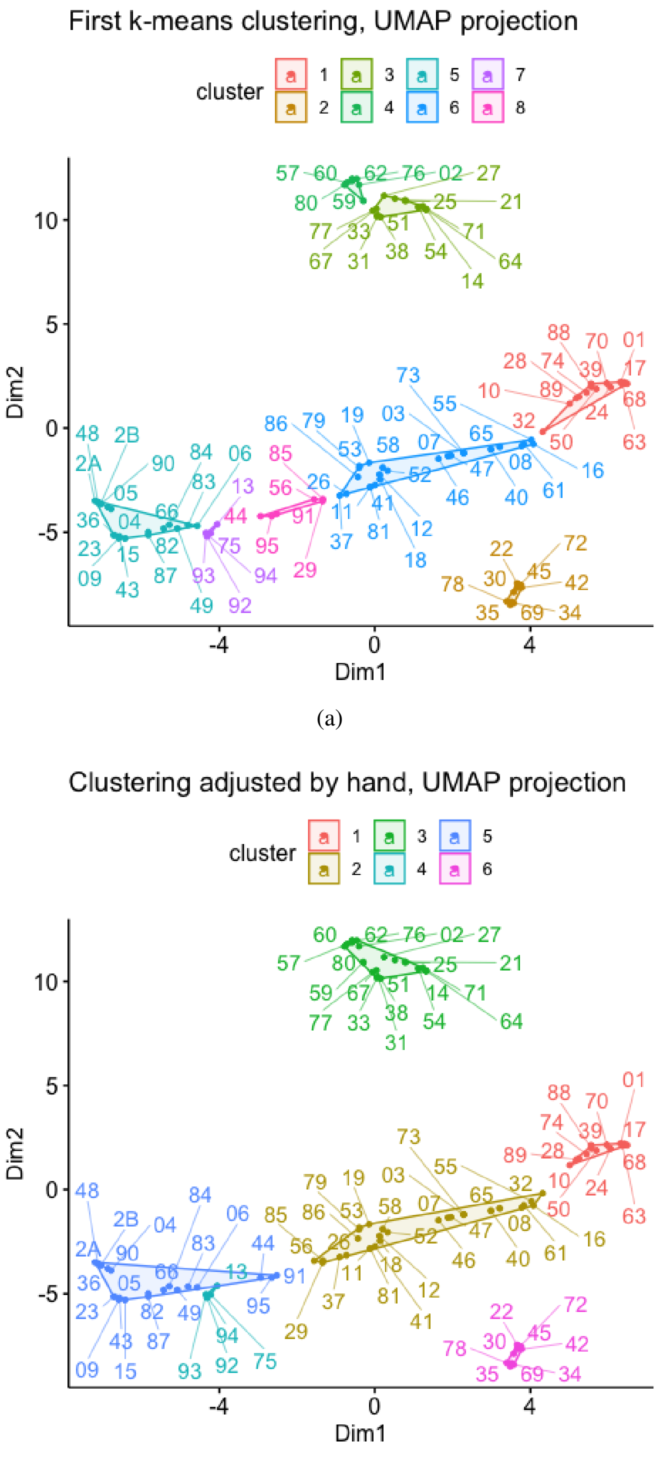
We tried several methods in order to perform this clustering. On the one hand, we failed to perform a good clustering when using *PCA* and *ISOMAP* but on the other hand we succeeded in getting a good clustering when using *t-SNE* and *UMAP*. At first glance, a good clustering should create a low number of clusters which gather similar cities (Paris and Marseille or Bordeaux and Toulouse) in terms of socio-conomic factors.

A. *UMAP clustering*

By performing a clustering using the *UMAP* method, we can display the projection of departments according to this clustering. We tried to find the best projection changing the number of neighbours in the *UMAP* algorithm. We chose to test the number of neighbours from 2 to 30 as it should approximate the number of clusters we then want. All these projections are represented in Figure (3). The result of these plots shows that taking a number of neighbors of 6 yields a pretty good clustering. Indeed, the number of clustering is not too high and the different clusters are well separated. We will keep this value of the number of neighbours for further study.

Then, we computed a *k-means* clustering in order to get a first base of clustering that could be adjusted. We performed this *k-means* on the 2D-projection we had done with the *UMAP* algorithm. We did it with different values for the number of clusters. After a few tests, we found that 8 clusters was a good trade-off since it really divides clusters well and we can merge certain clusters and certain points by hand in order to get the best possible clustering. We finally got 6 clusters: we decided to keep the small cluster of 5 departments including the 75 (Paris) as it was computed in the *k-means* (displayed in Figure (4.a)). Indeed, we deemed relevant to put the department of Paris a little bit isolated from the majority of French departments as it has the specialty to only correspond to a whole city, and has an important weight in terms of population. Moreover, *UMAP* combined with *k-means* succeeded into creating this cluster alone in comparison with *t-SNE*. These two clusterings are displayed in Figure (4).

In order to better visualize what the clustering we found looked like, we plotted the map of France thanks to the data of *dep_distance*. In this scatter plot, y corresponds to the average latitude of each department and x the average longitude. Then, we coloured each point with the clustering we built. We obtained the Figure (5). Therefore, we can see that the clustering is quite relevant. For example, the green cluster is approximately the ancient industrial region and the blue one corresponds



(a)

(b)

Fig. 4. a) *K-means*clustering with 8 clusters on the 2D projection of the *UMAP* function - b) *K-means*clustering adjusted by hand on the 2D projection of the *UMAP* function - The label of each point corresponds to its department code

to the *Ile-de-France* and the department of *Bouches-du-Rhône* (Marseille). We detail the naming procedure below.

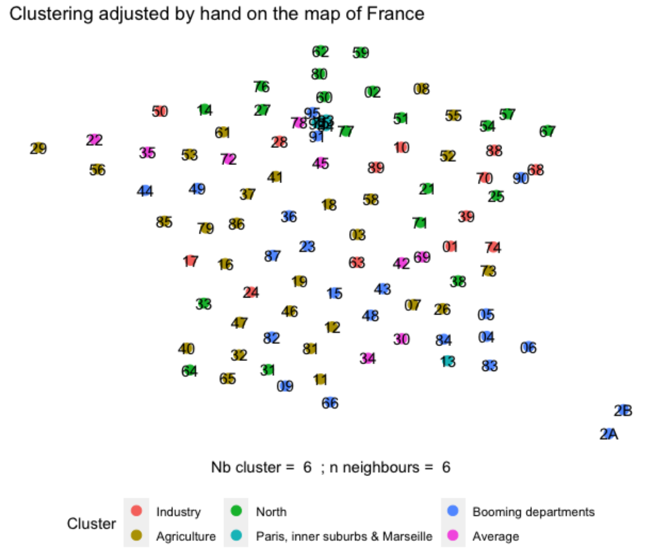

Clustering adjusted by hand on the map of France
Nb cluster = 6 ; n neighbours = 6
Cluster
Industry
Agriculture
North
Paris, inner suburbs & Marseille
Booming departments
Average

Fig. 5. Clustering of Figure (4) b) displayed on the map of France

B. *t-SNE clustering*

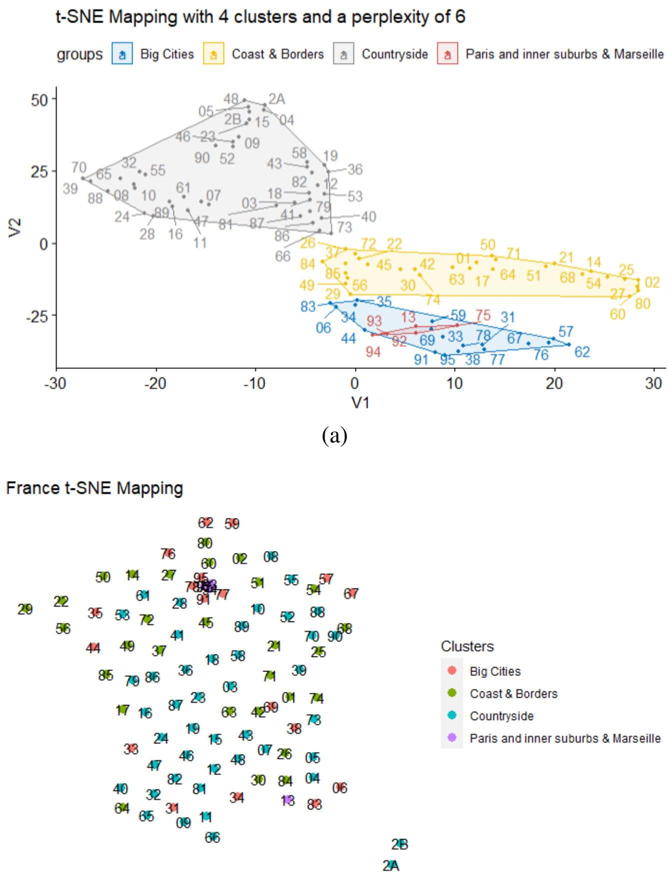
Considering the *t-SNE* clustering we chose to select a clustering with 4 clusters and a perplexity of 6. A smaller perplexity leads to less clusters which was not the goal here and a higher one divided cities that were supposed to be together.[1]. This clustering is represented in Figure (6). On Figure (6.a) we represent the 3 clusters that have been found with *k-means*. The blue one is being divided in two clusters. In fact, we decided to manually differentiate Paris, its inner suburbs & Marseille from the other french big cities. The idea was to create a new cluster that we think as different from the others and also to have a cluster that will be in both *UMAP* & *t-SNE* clusterings.

This clustering is interesting since it divided France in 4 really different clusters we can name according to their geographic distribution. In fact, looking at the display of France map according to the *t-SNE* clustering represented in Figure (6.b), we can really see that this clustering can be interpreted as a geographic clustering. Coast & Borders cluster seems to take into account touristic regions meanwhile Countryside's cluster highlight farming and industrial region. In the next section, we will discuss more about this clustering according to socio-economic factors.

C. *Other clustering methods*

The *ISOMAP* function enabled us to do a 2D-projection of all the points representing each department.

[1] The randomness of *t-SNE* was challenging and in order to find a clustering that could be reproduced, it is important to set the seed in your code



(a)

(b)

Fig. 6. a) Plot of the clustering according to *t-SNE* coordinates. b) Optimal department clustering

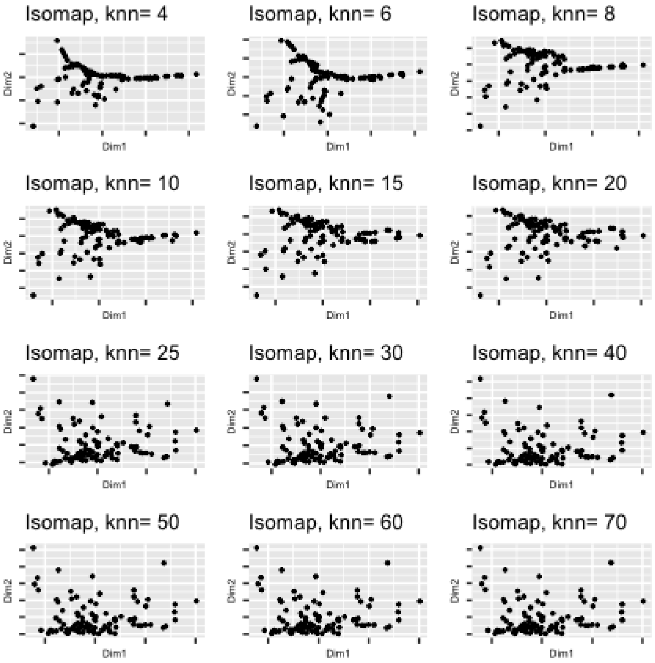


Fig. 7. Plot of the different 2D-projections using the *ISOMAP* method, with different number of neighbours

First, we wanted to find the number of neighbours that gave the best possible projection using the *ISOMAP* function. Therefore, we computed several projections changing its parameter (see Figure (7)).

Unfortunately, none of these projections gave a satisfying clustering. Indeed, none of them displayed distinguishable clusters, separated from each other. As a result, we did not choose to continue with this method.

Finally, performing a *PCA* before and even after removing any size effect did not succeed in creating a good department clustering.

D. Clustering selection

From this point, we therefore had selected two different clusterings: one produced by *UMAP* method and one produced by *t-SNE* method. In order to compare both department clusterings we first computed several indexes to have a more accurate assessment of each clustering and then performed a socio-economic analysis of both.

1) Analysis of indexes: To do so, we downloaded the package `fpc` in `R`. In this package, we used the function *cluster.stats* that computes a series of indexes to assess a clustering. Among these indexes, we chose to use *pearsongamma*, *sindex* and *dunn*. Indeed, most of other indexes were not relevant in our case as they take into account the scale of distances of the clustering which were not the same at all between the two clusterings we selected.

1) ***pearsongamma*** is better known as the Gamma index. Its mathematical formula is: $gamma = \frac{s_+ - s_-}{s_+ + s_-}$ where s_+ represents the number of times where the distance between one point and another point from the same cluster is *smaller* than with a point from another cluster and s_- represents the number of times where the distance between one point and another point from the same cluster is *bigger* than with a point from another cluster. Therefore, the better the clustering is, the closest to 1 *pearsongamma* gets.
2) ***sindex*** is the separation index which means it represents the difference between the maximum and the minimum distance between points and the barycenter of their cluster. Therefore, a low separation index indicates that clusters have approximately the same size and are quite homogeneous.
3) ***dunn*** index divides the minimum distance between two clusters by the maximum distance separating two points in the same cluster. Therefore, the dunn index should be the highest possible for a clustering.

	persongamma	sindex	dunn
UMAP	0.2636	48573	0.0099
t-SNE	0.3828	116238	0.0283

TABLE I
INDEXES RESULTS FOR *t-SNE* CLSUTERING AND *UMAP* CLUSTERING

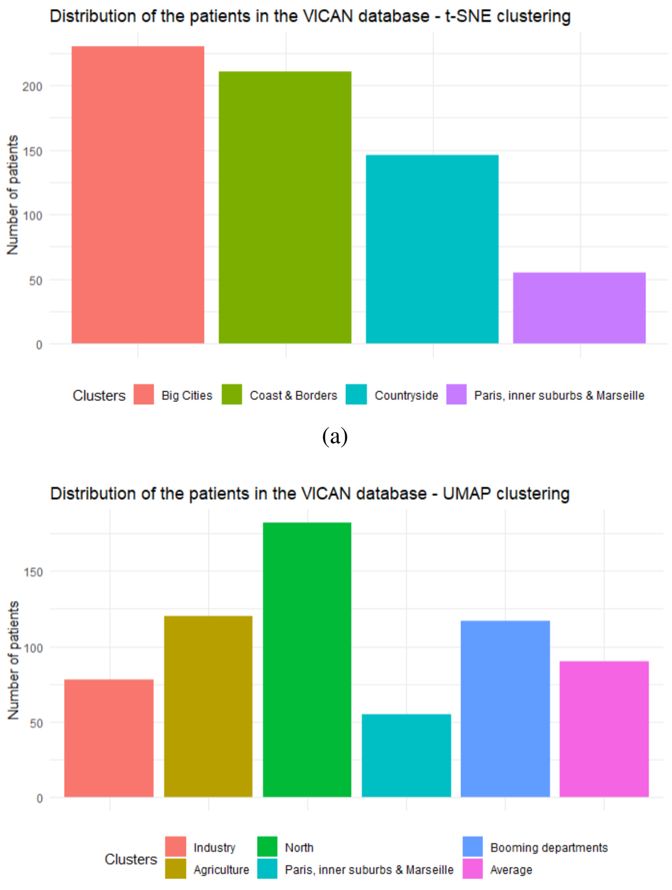
We computed these indexes on both clusterings. The results are gathered in Table (I). The second clustering performs better in separating the clusters regarding the *pearsongamma* index and the *dunn* index. However, when regarding the *sindex*, clusters seem more homogeneous in the *UMAP* clustering, we would thus select the clustering based on the *t-SNE* projection through this analysis.

However, we then performed a socio-economic analysis of both clusterings in order to see which one described best the population differences that were relevant for our further analyses of sequelae. Still, the analysis of indexes was a good first indicator of the relevance of these clusterings.

2) Socio-economic analysis: First, we observed the distribution of the patients in the database VICAN in order to see whether or not the analysis we would do further could be relevant with such clusterings. In Figure (8), we can see that for the clustering based on the *t-SNE* projection, there is an important disparity between the number of patients in the cluster 4 and the other clusters. For the clustering based on the *UMAP* projection, the distribution of patients is a little bit more balanced.

Then, we analysed the distribution of professional categories in clusters for both clusterings. The results are plotted in Figure (9).

First of all the **assumptions made in the previous section concerning the names given to the *t-SNE* clusters are being highlighted** by this plot. Both *Coast & Borders* and *Countryside* clusters have a higher proportion of farmers and industry. However the commerce sector is less developed (”shop” indicator), confirming our intuition that the tertiary sector is less developed in these clusters than in the other two clusters. Not only is this tertiary sector more developed in *Big Cites* and *Paris, inner suburbs and Marseille* but the access to healthcare is much better there (see Figure 10). Therefore the *t-SNE* performs well in splitting France between rural departments, semi-rural departments and urban departments.

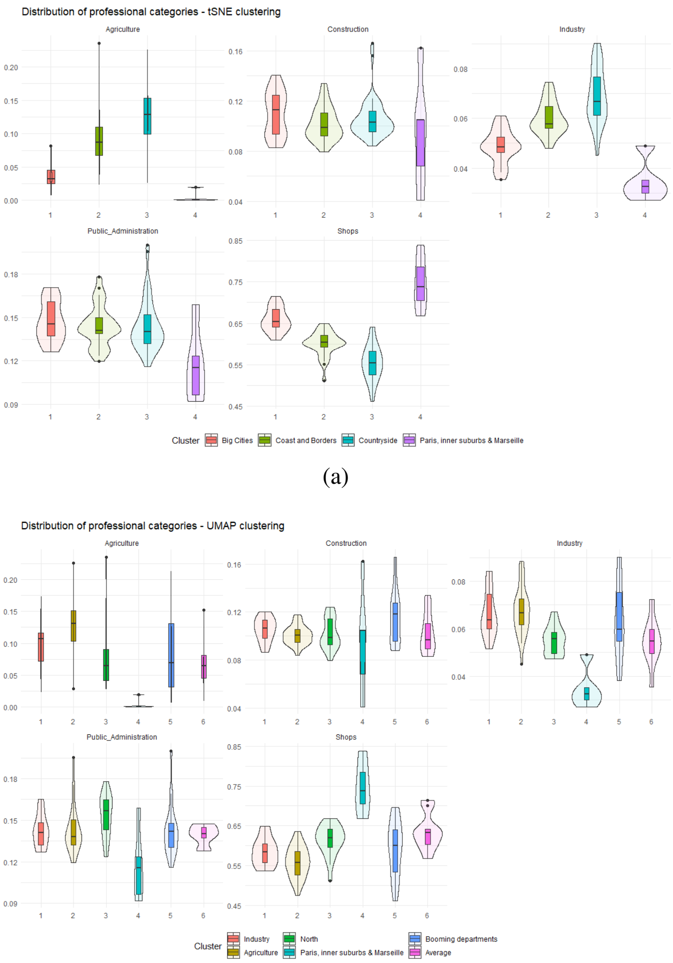


(a)

(b)

Fig. 8. a) Distribution of patients in the VICAN database among the clusters of the clustering based on the *t-SNE* projection - b) Distribution of patients in the VICAN database among the clusters of the clustering based on the *t-SNE* projection

The *UMAP* clustering is much more challenging since at first glance it is not a geographical clustering. On a closer look to the distribution of the patients in this clustering, it is very interesting that the only cluster that can been named geographically is the bigger one (230 women, 35.8% of the dataset).
With the help of Figure (9.b) we named the other clusters based on their economic activity.
We linked the first cluster to its industrial activity and labelled it *Industry*. With the same reasoning we labelled the second cluster *Agriculture*.
Considering both its industrial and its construction sectors and also its geographical position, the fifth cluster is very interesting. In fact when looking at the map, it gathers departments containing fast changing areas such as Nantes or Saclay. We interpreted the high dispersion of professional categories distribution among this cluster as the fact that the departments are changing really fast. Some areas in these departments are composed



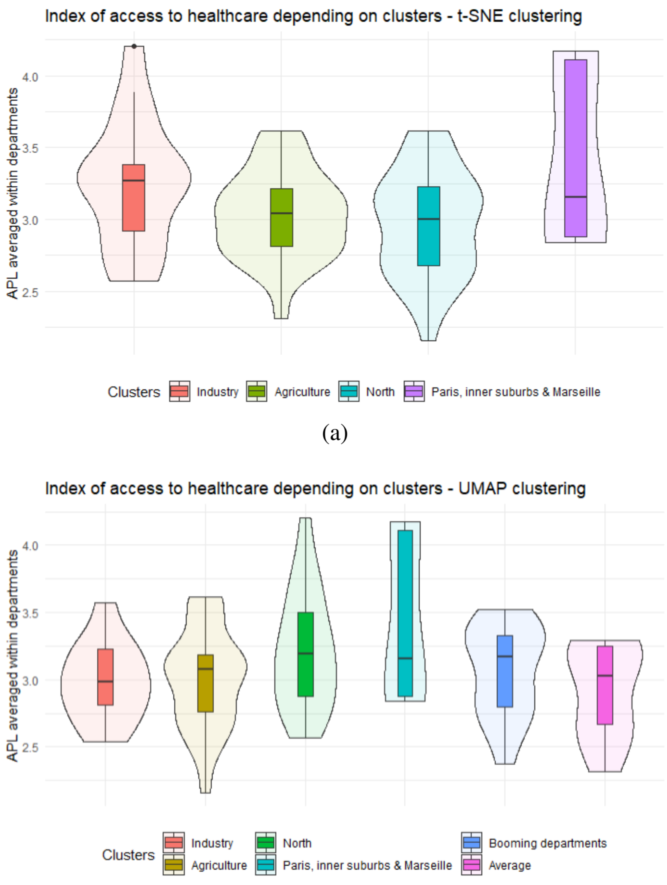
(a)

(b)

Fig. 9. a) Distribution of professional categories in clusters for the clustering based on the *t-SNE* projection - b) Distribution of professional categories in clusters for the clustering based on the *UMAP* projection - In each plot, y represents the rate of establishments from the category studied among all professional establishments in the cluster

with an high number of farms and industries and others are not. This is why we called this cluster *Booming departments*. In parallel with constructions, public administration is booming too (maybe because new cities are being created or expanded). It is interesting to look at the violin plot on Figure (10.b) since inside the *Booming Departments*, there are two bellies which could be relevant when considering our last analysis. Finally the last cluster has averaged value in every professional category. Thus we labelled it *Average*.

The *UMAP* clustering is then a new way to see France: not only with geographical criteria but trough a socio-economic analysis. We first thought that having too much clusters would be an issue for an in-depth understanding of the phenomenon but since the clusters are really different, it could finally be an asset to our analysis.



(a)

(b)

Fig. 10. a) Averaged APL per cluster for the clustering based on the *t-SNE* projection - b) Averaged APL per cluster for the clustering based on the *UMAP* projection - APL means *Accessibilité Potentielle Localisée* in French - In each cluster, we take into account the averaged APL index computed in each department

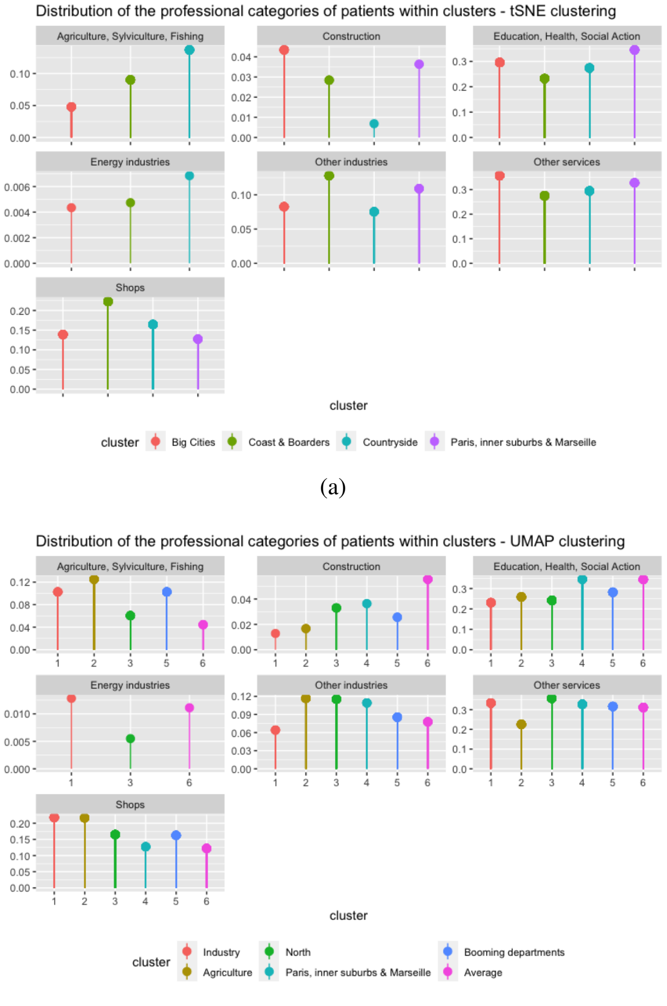
Once we had compared the socio-economic features of all clusters in both clusterings, we wondered if the patients recorded in the VICAN database had approximately the same socio-economic distribution as in our clusterings. Therefore, we analysed the business sector in which each patient from each cluster worked. The results are displayed in Figure (11). Once again, we see that our sample is representative of the overall population: for example, if we look at the distribution of the professional category ”Agriculture” among patients in the *UMAP* clustering (i.e. the proportion of patients that work in the agricultural sector for each cluster), the cluster with the highest proportion of patients working in agriculture is the one where the agricultural field was dominant in number of companies. The logic is the same for the cluster with second most patients in the agricultural sector, etc. (Figure (9.b)).

Regarding the field of Construction, however, the patients from the *Average* cluster have a bigger proportion than they should have if we only looked at the the distribution of professional establishments in this cluster. The same phenomenon occurs in the field of Industry with the cluster of *Paris, inner suburbs and Marseille*: Patients from the VICAN database in this cluster have a bigger weight that is comparable to other clusters whereas the distribution of professional establishments showed that there were really few industries in this cluster compared to others. Finally, Shops in the cluster *Paris, inner suburbs and Marseille* are also really less represented among the patients in the VICAN database compared with the ratio it should have considering the rate of establishments. To sum up, it seems that the VICAN dataset patients from Paris, Marseille and its suburbs (in the *UMAP* clustering) are not ”separated” by socio-economic factors as well as they could be.

For the clustering based on the *t-SNE* projection, the same goes for the field of Agriculture: the distribution of patients in this sector follows the same trend as the distribution of professional establishments. But patients in the cluster *Countryside* working in the Construction sector are really less represented than they should be. Moreover, the analysis we made on the cluster *Paris, inner suburbs and Marseille* in the *UMAP* clustering can be applied in this clustering too as this cluster is exactly the same in both clusterings. As a result, as we saw previously, patients from this cluster are less represented in Shops and too represented in Industry. All those differences we spotted first come from the fact that some clusters may have a sample of patients that is too small: this could lead to an important bias. Another reason may be the fact that the professional categories in the distribution of establishments and in the patients' data are not exactly the same. Therefore, some categories may overlap for instance.

3) Mathematical analysis of clustering: We have described how the clusters look like in terms of activities or geographical criteria, but we cannot compare our clusterings to a reference. Thus, we decided to go further in the understanding of the mechanisms hidden behind the terms *t-SNE* and *UMAP* which are used before using a k-means algorithm.

Regarding *t-SNE*, the probability of observing distances between any two points in the initial space decays exponentially with the standard deviation. For a small deviation, the probability of observing distances between any two points is null for distant points and grows very fast for the nearest neighbors. For a high standard deviation, points become equidistant. Thus, we should use an averaged standard deviation linked to perplexity values between 5 and 50 basically. Although local structures are preserved with this method, the global structure is not. That way, after a k-means algorithm, we cannot say if two clusters are close or different based on the projection. [2]



(a)

(b)

Fig. 11. a) Distribution of the professional categories of patients in each cluster of the clustering based on the *t-SNE* projection - b) Distribution of the professional categories of patients in each cluster of the clustering based on the *UMAP* projection - In each plot, y corresponds to the percentage of workers in such cluster working in such business sector

Considering *UMAP*, the use of the number of nearest neighbors instead of perplexity makes it easier to use. This method allowed us to have a better understanding of the projection because both global and local structures are preserved with the projection [3]. In fact, *UMAP* uses a different cost function (Cross-Entropy instead of KL-divergence) [4]

4) Conclusion: When we first computed both clusterings, our goal was to compare them in order to decide which one was the best and thus which of them we would keep for the sequelae analysis. After the

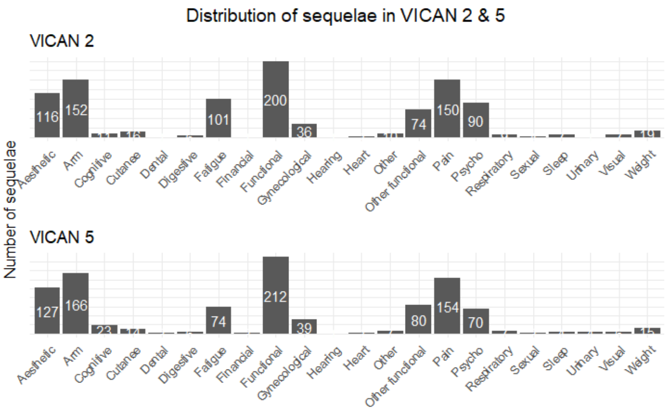


Fig. 12. Distribution of the number of sequelae according to each VICAN survey

previous analysis, we chose to keep them both since they describe different phenomona. Moreover, since they have a different size, they absorb the statistic bias in different ways which is an asset for interpretation.

III. Analysis of sequelae within clusters

Once both department clusterings had been analysed, we studied patients' sequelae within each cluster in order to see if there was a correlation between the socio-economic clusters in which patients live and their sequelae 2 years and 5 years after diagnostic.

In order to perform such an analysis we first chose to look into the distribution of sequelae in both VICAN surveys (Figure (12)). Some sequelae are much more common than others and therefore in order to do an easier analysis, we split sequelae in two groups. The first one is composed with *Aesthetic*, *Arm*, *Fatigue*, *Functional*, *Gynecological*, *Other Functional*, *Pain*, *Psycho*. The other is composed with the other type of sequelae. In VICAN 5 there was also the *lymphodedeme* sequel that is not displayed in this bar plot since it was not taken into account in VICAN 2. Overall we observe an increase of the number of sequelae over time (Figure (12)): this highlights the need to better help and monitor the patient in the years after the diagnostic.

A. Clustering based on the t-SNE projection

First, we considered the clustering based on the *t-SNE* projection. We observed the rate of patients within each cluster who declared having sequelae, categorizing by the type of sequelae in order to see whether there was a specific sequel that patients may have in specific clusters. We plotted these graphs for 2 years after diagnostic and 5 years after diagnostic (Figure (13 a. & b.)). For instance, among women in *Countryside* cluster, there is 25% of chances that they develop arm sequelae after 2 years. We then observed that 2 years after diagnostic, patients in the cluster *Paris, inner suburbs & Marseille* are more likely to suffer from aesthetic and psychological sequelae. 5 years after diagnostic, we can underline that these patients are generally more likely to have all forms of sequelae, and this time, more specifically aesthetic and functional ones. This observation is being reinforced by the Figure (13.c). In fact just by looking at the previous Figure we can't say that the increasing of the percentage of sequelae is a real increasing since some women's sequelae disappearance can compensate other women's sequelae appearance. Therefore this figure gives us a real tool to analyze the variation of overall sequelae among clusters.

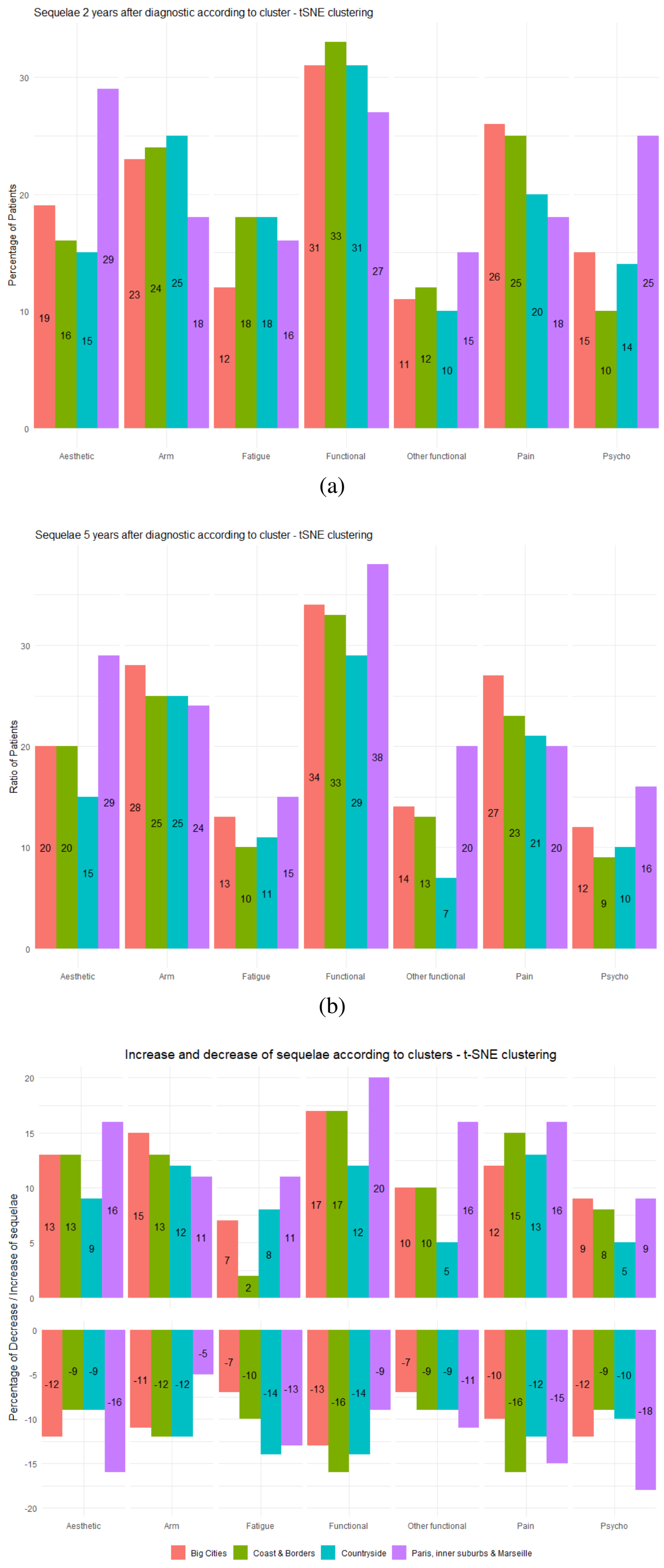
The interesting point here is that for the *t-SNE* clustering, we extracted artificially one cluster from the other. When looking at the sequelae distribution in (Figure (13)), there is a real difference between *Paris, inner suburbs and Marseille* cluster and the other clusters. However, asserting this is not completely accurate since the number of women in this cluster is very low and a variation of 5% corresponds to a small variation in absolute numbers in comparison with 3 other clusters. Looking at the two first graphs of Figure (13), one may infer that there is no real evolution in each cluster between *VICAN 2* and *VICAN 5*. However, the third graph shows that there is actually one. Indeed it displays only the patients who changed their answer in the questionnaire with an increase when they declared a new sequel and a decrease when their sequel ends. That way, the lack of differences between *VICAN 2* and *VICAN 5* shows that our clustering fails to model the evolution of sequelae throughout time. We actually see that almost each new sequel declared between the two studies is compensated by another patient who declared that the same sequel came to an end. In any cluster, and for almost every kind of sequelae, there is no real evolution in the total number of positive answers whereas there is a big evolution regarding individuals.

We admit that the *t-SNE* clustering fails to draw a relation between socio-economic factors, geographic inequalities and breast cancer sequelae, as it is unable to account for any kind of evolution of most sequelae.

B. Clustering based on the UMAP projection

We then performed the same analysis of sequelae on the clustering based on the *UMAP* projection. The results (Figure (14)) are quite different. Still, once again, the cluster containing Paris, Marseille and its suburbs gathers more aesthetic and psychological sequelae than all the other clusters as no cluster stands out in terms of this category of sequelae. Also, living in *Booming departments* seems to favor the apparition of sequelae concerning arms, pain and functional ones but especially 2 years after diagnostic. Since the clusters of the *UMAP* method are more balanced in terms of number of people, it may be more interesting to compare different percentages between clusters. Looking at the *Booming departments* we find a high proportion of patients touched by arm, functional and pain sequelae. However, 5 years after the diagnostic, the *Averaged* cluster is dominant in terms of functional sequelae: it has between 12% and 100% more patients with functional sequelae than the other clusters; this can also be seen in the other functional sequelae. We could infer that the patients have been better monitored in the *Booming departments*, but to confirm or infirm this intuition we would need a larger set of patients. Some patients might also have changed departments and still experience the same sequelae, although only 3% patients have moved between *VICAN 2* and *VICAN 5*. Sequelae tend to grow in the industrial cluster between *VICAN 2* and *VICAN 5*. This is all the more interesting as this cluster's APL (accessibility to healthcare) is the lowest of all clusters.

This trend establishing a link between a cluster's evolution from VICAN 2 to VICAN 5 and its APL measurement is not always confirmed: Paris, its suburbs and Marseille have a high APL but growing sequelae between *VICAN 2* and *VICAN 5*. For the *Average* cluster as well, sequelae tend to grow except for psychological and pain sequelae but for some sequelae, the increase is very important such as for functional sequelae or the ones concerning arms. It might also be explained by a big part of people in this cluster living in a place where the APL index is quite low compared to other places. In the *North* cluster, sequelae tend to decrease in most cases, except the sequelae regarding pain. This also might be implied by the fact that the APL index is slightly above the average of clusters as we can observe in (Figure (10)). Therefore, the analysis of increases and decreases of sequelae in the clustering based on the *UMAP* clustering seems to say that patients in the *North* cluster may have less sequelae than in other clusters, and especially the industrial cluster.

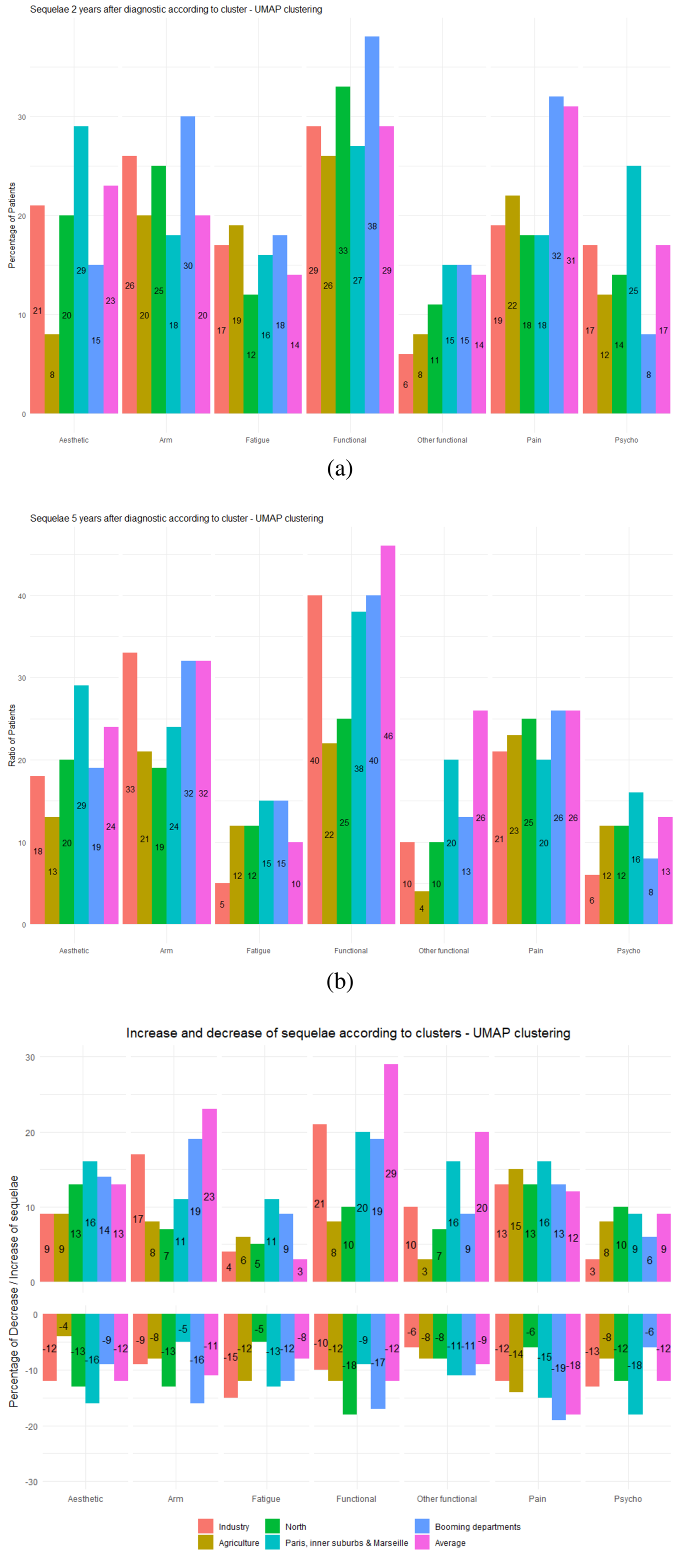


(a)

(b)

(c)

Fig. 13. a) Percentage of patients having different types of sequelae 2 years after diagnostic, within each cluster from the clustering based on the *t-SNE* projection - b) Percentage of patients having different types of sequelae 5 years after diagnostic, within each cluster from the clustering based on the *t-SNE* projection - c) Percentage of appearance and disappearance of sequelae between both survey per cluster



(a)

(b)

(c)

Fig. 14. a) Percentage of patients having different types of sequelae 2 years after diagnostic, within each cluster from the clustering based on the *UMAP* projection - b) Percentage of patients having different types of sequelae 5 years after diagnostic, within each cluster from the clustering based on the *UMAP* projection - c) Percentage of appearance and disappearance of sequelae between both survey per cluster

IV. Conclusion

Throughout our work, we understood how difficult it could be to select good hyperparameters for our clusterings. In fact we selected clusterings that matched with what we thought was the best but without rigorous justification. We also were faced with the small size of the dataset. We had to be careful not to generalize some cases where only a few number of individuals have a big impact on a criterion.

However we did find some encouraging results. The clustering obtained with *UMAP* separates well the North with a good recovery overall in comparison from other clusters. *Paris, inner suburbs Marseille* stand out among other clusters when it comes to *psycho* and *aestetic* sequelae.

V. Discussion

We have to be careful about our interpretations of sequelae and their evolution. Indeed, biases can exist. For example, a patient can know her disease better 5 years after than 2 years after. She can also better accept the effects it has had on her. This does not imply these effects have grown, but only a change of perception of these effects.

If we want our study to be thorough, we should also study objective, medical measurements of a patient's ability to move her arm, etc. as a complement to these questionnaires.

To pursue our work, it would have been interesting to perform a clustering on a table merging the socio-economic data of departments and the data concerning the sequelae of patients. Then, it would be helpful for the study to compare this clustering with our results to see if there are relevant similarities. This work could validate our results for example.

Finally, our project led us to first cluster the departments and in a second time, analyse the distribution of sequelae of the patients in the VICAN database to find correlations. But we also thought of another method that could describe the relationship between geographic inequalities and breast cancer sequelae. This method planned to compute the characteristics of an average patient in each department. Then, we would cluster all these average patients with a k-means method projected thanks to different methods of visualization (*t-SNE*, *UMAP*, *ISOMAP*...) just as we did in our study. Therefore, we could observe the average type of sequelae in each cluster. A variant of this method would be to cluster the distribution of patients' sequelae and then see if there was a correlation with geographic and socio-economic data. However, the main issue with this method would be that we may loose a lot of information and data when computing an average patient in each department. Still, it would be very interesting for further studies to implement this method and compare the results to our method.

References

[1] I. national du cancer, "La vie cinq ans après un diagnostic de cancer," 2018.

[2] G. Hinton and L. van der Maaten. (2008) Visualizing data using t-sne. [Online]. Available: https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

[3] J. H. James Melville and L. McInnes. (2020) UMAP: Uniform manifold approximation and projection for dimension reduction. [Online]. Available: https://arxiv.org/pdf/1802.03426.pdf

[4] N. Oskolkov. (2019) How exactly UMAP works? [Online]. Available: https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668