الجمهوريــة الجزائريــة الديمقراطيــة الشعبيــة
République Algérienne Démocratique et Populaire

وزارة التعليـم العالــي و البحــث العلمــي الجمهوريــة
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المديريـة العامـة للبحـث العلمـي و التطويـر التكنولوجـي
Direction Générale de la Recherche Scientifique et du Développement Technologique

Rapport final de projet
التقرير العام لمشروع البحث

PNR البرنامج الوطني للبحث في:	**Organisme pilote** الهيئة المشرفة
ENSSEA	**CREAD**

Domiciliation du projet : مؤسسة توطين المشروع:

Intitulé du projet : عنوان المشروع :

Détection de Clusters de Cancers de la thyroïde par Balayage Statistique

Intitulé du domaine : domaine 1 : les secteurs d'activité	
Intitulé de l'axe : Axe 5 : Economie et politiques sociales	
Intitulé du thème : Identification et étude des clusters (groupements) de malades de cancer de la thyroïde et détermination et analyse des facteurs ayant provoqué ou favorisé l'émergence de ces clusters	**الموضوع**

Chef de projet رئيس المشروع

Nom et prénom اللقب و الاسم	**Grade** الرتبة	**Etablissement de rattachement** المؤسسة المستخدمة
Moussi Oumelkheir	MC A	ENSSEA

Equipe de recherche أعضاء المشروع

Nom et prénom اللقب و الاسم	**Grade** الرتبة	**Etablissement de rattachement** المؤسسة المستخدمة	**Observation** الملاحظة
Boudrissa Naima	MC A	ENSSEA	
Bouakline Siham	MA A	Université de Bougie	Doctorante
Semrouni Mourad	Professeur	EHPS Mustapha	
Hasbellaoui Fella	MA A	EHPS Mustapha	Doctorante
HadjArab Samir	MA A	EHPS Mustapha	

Titre du projet PNR :

Nom et prénom du chef du projet: Moussi Oumelkheir Grade :Maître de Conférences A

Nom et prénom des membres du projet :

Semrouni Mourad Professeur

Boudrissa Naima Maître de Conférences A

Hadjarab Samir Maître de Conférences B

Bouakline Siham Maître Assistante A

Hasbellaoui Fella Maître Assistante A

Originalité de la recherche

Les cancers en Algérie sont devenus émergents et constituent une préoccupation majeure. Le cancer de la thyroïde augmente de manière spectaculaire depuis une quinzaine d'années, c'est le troisième cancer chez la femme d'après le registre des tumeurs d'Alger. Entre 2007 et 2010 il a été multiplié par 6 et ce seulement au service endocrinologie du centre Pierre et Marie Curie d'Alger. Ces observations nous amènent à se poser plusieurs questions :

●Certaines wilayas ont-elles un nombre de cas de cancers excessif?

●Les cas de cancer de la thyroïde sont-ils anormalement concentrés?

● La distribution spatiale de ces cas, est-elle aléatoire ?

Répondre à ces questions revient à décrire l'hétérogénéité spatiale.

Un cluster est une organisation spatiale définie comme un agrégat, un regroupement de cas proches les uns des autres ; la proximité étant définie au sens d'une distance géographique.

Les méthodes de détection de clusters identifient les regroupements de cas incohérents sous l'hypothèse nulle d'absence de clusters et évaluent leur niveau de significativité .Quant aux méthodes de détection globale d'agrégation de cas, elles étudient la corrélation spatiale et détectent la tendance des cas à la clustérisation.

Dans ce travail nous allons utiliser des méthodes de détection d'agrégats de cas, dont les statistiques sont souvent basées sur les distances afin d'analyser l'existence des clusters de cancers de la thyroïde en Algérie. Ce sont le test du coefficient de corrélation de Moran, le

test de Tango et la méthode de balayage de Kulldorff. Le test d'ajustement a été utilisé pour vérifier l'hypothèse des risques constants de l'incidence du cancer de la thyroïde.

Résultats *:*

Notre base de données est constituée par l'ensemble des dossiers de patients pris en charge pour cancers thyroïdiens différenciés diagnostiqués sur une période de 05 ans (2007-2011) au CPMC et hospitalisés pour cancers de la thyroïde.

Ayant construit un échantillon de 527 malades et 60 variables, des investigations statistiques ont été entreprises et un état des lieux sur les clusters de cancer de la thyroïde a été réalisé. Les résultats ont fait l'objet d'une communication internationale intitulée : **Détection et identification des clusters de cancer de la thyroïde en Algérie (Journées de statistique organisées à Toulouse par la Société Française de Statistique en Mai 2013).**

L'ensemble des méthodes statistique utilisées rejettent l'hypothèse d'un risque constant et suggère une tendance à une concentration de cas de cancer de la thyroïde. La recherche de clusters spatiaux montre l'existence d'un cluster significatif regroupant les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6. Il contient 389 cas.

La recherche de clusters spatiaux- temporels montre un cluster significatif entre 2007 et 2011 regroupant les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6, 19, 14, 18, 17, il contient 438 cas.

Nous faisons remarquer que la constitution de l'échantillon se poursuit dans le but d'atteindre les 1000 malades. Par ailleurs le développement des outils statistiques dans le cadre du balayage statistique et l'analyse bayésienne afin d'identifier et étudier l'émergence de clusters se fait dans le cadre d'une thèse de doctorat d'Etat en cours de réalisation.

Introduction et intérêt de la recherche :

De nombreuses méthodes ont été développées pour tester une tendance à l'agrégation de cas d'une pathologie .Leur objectif est de mieux comprendre la distribution géographique des maladies et d'en étudier l'hétérogénéité spatiale.

Une première approche consiste à analyser globalement la distribution spatiale et temporelle d'une maladie. Une deuxième approche s'intéresse à l'estimation du risque d'une maladie par rapport à un point source.

Un cluster ou agrégat peut être défini comme une concentration de cas "anormalement élevée", dans un groupe de personnes, une zone géographique ou une période de temps.

Les tests proposés dans le but de savoir si les événements sont agrégés dans l'espace peuvent être classés selon leur objectif. On distingue :

- les tests globaux pour évaluer la tendance globale à la clusterisation ou à l'agrégation de l'incidence d'une maladie dans une région d'étude. Les méthodes de « clusterisation globale » étudient la corrélation spatiale et détectent la tendance des cas à l'agrégation.

- les tests de détection pour identifier la localisation des clusters potentiels et tester si ces derniers sont significatifs. . Les méthodes de détection de cluster identifient les regroupements de cas incohérents avec l'hypothèse nulle "pas de clusterisation" et évaluent leur niveau de significativité.

-les tests focalisés ou de concentration sont utilisés quand une information permet à l'avance de spécifier une coordonnée géographique autour de laquelle la recherche d'une concentration de cas va se focaliser.

Les analyses de clusters peuvent être classées selon le type de données qu'elles permettent d'étudier. Les deux catégories de données sont définies par leur niveau de résolution : elles sont soit agrégées ou de comptage (par exemple, le nombre de cas et la population par commune de la zone géographique étudiée, c'est le cas dans notre projet) ; soit ponctuelles ou individuelles (par exemple, les coordonnées spatiales des cas et de la population à risque ou des témoins).

Pour effectuer les tests globaux, il est nécessaire de décrire la proximité entre les unités spatiales, cette dernière est donnée par la matrice de proximité noté W. W est une matrice carrée et elle résume la relation entre chaque couple d'unités spatiales de la zone étudiée. Cette proximité peut être la distance entre deux unités spatiales ou par dichotomie en donnant la valeur 1 si le couple a des frontières en commun et zéro sinon (Waller & Gotway (2004)). Cela permet d'affecter un poids à chaque couple.

I-Détection de clusters et méthodes de balayage spatial

L'objectif des méthodes de balayage spatial est la surveillance géographique d'un territoire dans le but de détecter les zones pour lesquelles une incidence plus élevée de cas d'une maladie est observée, sans hypothèses à priori.

Les méthodes de balayage spatial cherchent à détecter l'emplacement des clusters dans la région étudiée. Elles appliquent des fenêtres (souvent des cercles) sur toute la région et dénombrent les cas et les individus à risque à l'intérieur et à l'extérieur de chaque fenêtre. Il existe différentes méthodes de balayage spatial, la méthode d'Openshaw, la méthode de Besag et Newell et la statistique de scan spatiale, et elles se distinguent entre autres par la construction de la fenêtre qu'elles utilisent.

I-1**Méthodes de balayage spatial : la statistique de scan spatiale**

La statistique de scan spatiale est la plus usitée. Son objectif est d'identifier les zones ayant une incidence anormalement élevée et qui sont les moins "cohérentes" avec l'hypothèse nulle de risque constant. Cette méthode est basée sur un test du rapport de la vraisemblance. Elle est très puissante et s'applique aussi bien sur des données agrégées que ponctuelles.

Une fenêtre, de forme prédéfinie (cercles ou ellipses), de taille variable, balaye la zone d'étude. Pour chaque fenêtre, une statistique, basée sur le rapport de vraisemblance et les nombres de cas observés et attendus, est calculée.

Les fonctions de vraisemblance s'écrivent selon le choix de la distribution théorique associée au nombre de cas.

Deux distributions peuvent être définies : la loi de Poisson (données agrégées ou lorsque le nombre de cas est négligeable face à la taille de la population) et la loi binomiale (données individuelles des cas et témoins). L'hypothèse alternative, pour chaque "position spatiale" et taille de fenêtre, est qu'il existe un risque élevé à l'intérieur de la fenêtre par rapport à l'extérieur de la fenêtre. La fenêtre qui correspond au maximum de vraisemblance est le cluster le plus probable, celui qui a le moins de chance de survenir par hasard. Une probabilité p, calculée à partir de simulations de Monte Carlo, est assignée à ce cluster.

Le logiciel SaTScan peut être utilisé pour mettre en œuvre la statistique de scan spatiale (et spatio-temporelle).

Il s'agit d'un logiciel développé par Kulldorff . SaTScan permet de détecter des clusters spatiaux ou spatio-temporels, et de voir s'ils sont statistiquement significatifs ; de tester si la maladie est distribuée aléatoirement dans l'espace, le temps ou dans l'espace et le temps.

Le nombre de cas, la population et les coordonnées géographiques de chaque unité spatiale (dans notre projet le chef-lieu de Wilaya) de la zone étudiée doivent être définis. Des covariables (sexe, classes d'âge, densité de population, score socio-économique…) peuvent être prises en compte.

D'autres méthodes de détection de clusters ont été développées mais, la méthode de balayage spatiale de Kulldorff reste l'outil le plus utilisé pour identifier des clusters potentiels.

II-Tests de détection globale de clusters

Ces méthodes s'intéressent à l'existence d'une hétérogénéité globale de la distribution spatiale d'une maladie.

L'objectif de ces méthodes est d'étudier la corrélation spatiale et de détecter la tendance des cas "à la clusterisation ". Ces méthodes ne donnent pas la localisation des clusters.

Il existe de nombreuses méthodes de détection globale de clusters. On présente le test de Moran et le test de Tango qui sont très utilisés dans les études de corrélation spatiale.

II-1 La statistique de Moran :

Une deuxième méthode évalue l'existence d'une hétérogénéité spatiale globale en termes d'autocorrélation spatiale.

La statistique de Moran est l'indice d'autocorrélation spatiale le plus utilisé. Cette statistique résume le degré de ressemblance des unités géographiques voisines par une moyenne pondérée de la ressemblance entre observations.

Le test repose sur la statistique dite Indice de Moran (Waller & Gotway(2004)) noté I et définie par :

$$I = \frac{1}{w_+} \cdot \frac{\sum_{i=1}^{i=K} \sum_{j=1}^{j=K} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{i=K} (y_i - \bar{y})^2 / K}$$

Où $K = nombre\ d'unités\ spatiale$

$w_{ij} = éléments\ de\ la\ matrice\ de\ proximité\ pour\ les\ unités\ spatiales\ i\ et\ j.$

$w_+ = \sum_{i,j=1}^{K} w_{ij}$

$$y_i = \frac{O_i}{n_i} = \frac{nombre\ de\ cas\ observés\ dans\ l'unité\ spatiale\ i}{effectif\ de\ l'unité\ \ spatiale\ i}$$

$$\bar{y} = \frac{\sum_{i=1}^{i=K} y_i}{K} = moyenne\ des\ proportions\ sur\ l'ensemble\ des\ K\ unités\ spatiales$$

La statistique I est donc une variable aléatoire. Sous l'hypothèse nulle H_0, I suit une loi asymptotiquement normale identique quel que soit l'unité spatiale i($I \rightarrow N(m, \sigma^2)$) avec :

$$\hat{m} = -1/K - 1$$

$$\widehat{\sigma^2} = \frac{(K^2 . \frac{1}{2} . \sum_{i \neq j}^{K} (w_{ij} + w_{ji})^2 - K . \sum_{i=1}^{i=K} (w_{i+} + w_{+i})^2 + 3{w_+}^2)}{(K-1)(K+1){w_+}^2} - \hat{m}^2$$

$$w_{i+} = \sum_{j=1}^{K} w_{ij} \ \ , w_{+j} = \sum_{i=1}^{K} w_{ij}$$

L'indice de Moran mesure la similitude entre les unités spatiales voisines, son interprétation est similaire a celle d'un coefficient de corrélation (Gaudart (2007)):

I< 0 ⇔ Autocorrelation spatiale négative, donc les unités spatiales voisines sont différentes.

I ≃ 0 ⇔ il n'y'a aucune corrélation entre les unités spatiales voisines, et le modèle spatial est parfaitement aléatoire.

I> 0 ⇔ Les unité spatiales voisines sont similaires (existence d'un pattern sous forme d'un cluster d'unités spatiales).

La statistique de Moran ne prend pas en compte l'hétérogénéité des effectifs de population. Une corrélation spatiale significative pourrait être expliquée par la proximité de zones fortement peuplées et non pas par un cluster de taux élevés. Des versions alternatives de la statistique de Moran ont été proposées pour prendre en compte des effectifs de population hétérogènes.

II-2-La statistique de Tango :

Tango a proposé une statistique pour l'évaluation d'une clusterisation globale(global clustering). La méthode de Tango teste si les cas de maladie sont regroupés dans des clusters à l'intérieur de la région d'étude.

Tango (1995) a généralisé spatialement le test précédent (test de chi-deux de Pearson) en pondérant les écarts par la proximité. La statistique dite de Tango (noté T) s'obtient de la manière suivante:

On considère $P_1 = \left(\frac{O_1}{O_+} \dots \frac{O_K}{O_+}\right)$ le vecteur des proportions observées où $O_+ = \sum_{i=1}^{k} O_i$ est le nombre total de cas observés, et $P_2 = \left(\frac{n_1}{n_+} \dots \frac{n_K}{n_+}\right)$ le vecteur des proportions attendues avec $n_+ = \sum_{i=1}^{k} n_i$ la population totale exposée au risque.

Sous l'hypothèse nulle H_0 de risque constant, les proportions attendues sont issues d'une distribution multinomiale. La statistique de Tango compare les proportions observées et les proportions attendues et elle est définie par :

$$T = \sum_{i,j}^{k} w_{ij} \left(\frac{O_i}{O_+} - \frac{n_i}{n_+}\right)\left(\frac{O_j}{O_+} - \frac{n_j}{n_+}\right)$$

Sous l'hypothèse H_0 de risque constant, la variable T est asymptotiquement normale

$T \to N(E(T), V(T))$ avec

$$E(T) = \frac{1}{O_+} tr(WV_{P_2}), \quad V(T) = \frac{1}{O_+^2} tr\left[(WV_{P_2})^2\right] \qquad V_{P_2} = diag(P_2) - P_2 \acute{P}_2$$

La forme standard de la statistique de Tango est donnée par :

$$T^* = \frac{T - E(T)}{\sqrt{V(T)}}$$

Tango (1990) a proposé une approximation par la loi de chi-deux de la statistique donnée par :

$v + T^* \sqrt{2v} \to \chi_v^2$

où $v = 8\left(2\sqrt{2}\frac{tr\left[(WV_{P_2})^3\right]}{\left(tr\left[(WV_{P_2})^2\right]\right)^{1,5}}\right)^{-2}$ est le degré de liberté.

III-Identification des clusters ou analyse de regroupement :

Le but d'une analyse de regroupements est de créer des groupes d'individus homogènes à partir d'un certain nombre de variables. Plus précisément, nous voulons combiner des sujets en groupes interprétables de telle sorte que les individus d'un même groupe soient semblables et que les groupes soient différents.

Plusieurs méthodes ont été développées pour le traitement de données continues et d'autres pour les données qualitatives, mais ces méthodes ne sont pas fiables pour les deux types de données à la fois (cas le plus fréquent dans la pratique).

La méthode de clusters en deux étapes (Two-step clustering) introduite par Chiu et al. (2001) a pour but l'identification de clusters pour les données mixtes. Donc les données seront organisées en groupes (clusters), les membres de chaque cluster sont très similaires entre eux et très dissimilaires avec les membres des autres clusters ; ceci sur la base d'une ou plusieurs variables discriminantes. Le nom de la méthode signifie que son algorithme s'applique en deux étapes :

Etape 1 : les individus sont attribués à des pré-clusters

Etape 2 : Les pré-clusters sont clustérisés une deuxième fois à l'aide de l'algorithme hiérarchique (Abu Abbas (2008)).

Cette méthode implémentée sur le package SPSS suppose toutes les variables indépendantes, les variables continues normalement distribuées, et les variables catégorielles de distribution multinomiale (Voir (Bacha et al. (2007)) pour plus de détails).

Conclusion :

Les méthodes I et II répondent aux objectifs suivants : tester si une maladie est distribuée aléatoirement dans la région étudiée ; détecter des zones à incidence élevée…Huang *et al.* comparent ces différents tests pour répondre aux questions suivantes : quelle est la méthode la plus appropriée ou (et) la plus puissante pour comprendre la distribution spatiale d'une maladie ? Est-il possible de fournir un guide pour l'utilisation de ces méthodes statistiques quand elles sont appliquées par exemple à des données de cancer ?

Parmi les tests de clusterisation globale considérés, le test de Tango semble le plus puissant. Parmi les méthodes de détection de cluster étudiées la statistique de Kulldorff avec fenêtres elliptiques semble être la plus puissante quant aux autres elles ne peuvent être considérés que comme des méthodes de "dépistage", devant être complétées par des tests d'analyse de clusters (analyse de regroupement ou two-step clustering par exemple), des études plus ciblées ; pour confirmer (ou pas) les hypothèses qu'elles permettent de dégager.

Dans notre étude nous avons privilégié l'utilisation de plusieurs tests, basés sur des hypothèses et des méthodes d'estimations différentes ; le but cherché, étant la convergence cohérente des résultats.

Présentation de la problématique :

Les cancers en Algérie sont devenus émergents et constituent une préoccupation majeure. Le cancer de la thyroïde augmente de manière spectaculaire depuis une quinzaine d'années,

c'est le troisième cancer chez la femme d'après le registre des tumeurs d'Alger. Entre 2007 et 2010 il a été multiplié par 6 et ce seulement au service endocrinologie du centre Pierre et Marie Curie d'Alger. Ces observations nous amènent à se poser plusieurs questions :

- Certaines wilayas ont-elles un nombre de cas de cancers excessif?
- Les cas de cancer de la thyroïde sont-ils anormalement concentrés?
- La distribution spatiale de ces cas, est-elle aléatoire ?

Répondre à ces questions revient à décrire l'hétérogénéité Spatiale. Un cluster est une organisation spatiale définie comme un agrégat, un regroupement de cas proches les uns des autres ; la proximité étant définie au sens d'une distance géographique. Les méthodes de détection de clusters identifient les regroupements de cas incohérents sous l'hypothèse nulle d'absence de clusters et évaluent leur niveau de significativité .Quant aux méthodes de détection globale d'agrégation de cas, elles étudient la corrélation spatiale et détectent la tendance des cas à la clustérisation.

Le but de ce travail est double tester la présence de clusters de thyroïde par les tests globaux (test de Moran et test Tango) et le test du Chi-deux de Pearson ; localiser les clusters pouvant exister par la méthode des clusters en deux étapes et confirmer la significativité des clusters potentiels par la méthode de balayage de Kulldorff

Méthodologie :

Différentes méthodes statistiques ont été développées (Besag & Newell(1991)) pour l'identification des « structures spatiales (clusters) », Pour vérifier l'existence ou la non existence d'agrégats spatiaux on a choisi les tests de détection globale (Gaudart et al 2007). Ces tests se caractérisent par la recherche d'une tendance générale à l'agrégation, ils reposent sur les hypothèses suivantes :

$$\begin{cases} H_0 : Hypothese\ du\ risque\ constant\ (la\ non\ existance\ de\ clusters). \\ H_1 : Il\ ya\ tendance\ \text{à}\ la\ clusterisation. \end{cases}$$

Ils estiment une statistique pour la zone étudiée dans le but de tester l'existence d'une hétérogénéité globale.

Pour effectuer les tests globaux, il est nécessaire de décrire la proximité entre les unités spatiales, cette dernière est donnée par la matrice de proximité noté W. W est une matrice carré et elle résume la relation entre chaque deux paires d'unités spéciales de la zone étudiée. Cette proximité peut être la distance entre deux unités spatiales ou par dichotomie en donnant la valeur 1 si la paire a des frontières en commun et zéro sinon (Waller & Gotway (2004)). Cela permet d'affecter un poids à chaque paire.

Trois tests ont été utilisés, le test de Moran, le test de chi-deux de Pearson et le test de Tango.

2-1- Test de Moran

Le test repose sur la statistique dite Indice de Moran (Waller & Gotway(2004)) noté I et définie par :

$$I = \frac{1}{w_+} \cdot \frac{\sum_{i=1}^{i=K} \sum_{j=1}^{j=K} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{i=K} (y_i - \bar{y})^2 / K}$$

Où $K = nombre\ d'unités\ spatiale$

$w_{ij} = éléments\ de\ la\ matrice\ de\ proximité\ pour\ les\ unités\ spatiales\ i\ et\ j.$

$w_+ = \sum_{i,j=1}^{K} w_{ij}$

$y_i = \frac{O_i}{n_i} = \frac{nombre\ de\ cas\ observés\ dans\ l'unité\ spatiale\ i}{effectif\ de\ l'unité\ \ spatiale\ i}$

$\bar{y} = \frac{\sum_{i=1}^{i=K} y_i}{K} = moyenne\ des\ proportions\ sur\ l'ensemble\ des\ K\ unités\ s$patiales

La statistique I est donc une variable aléatoire. Sous l'hypothèse H_0, I suit une loi asymptotiquement normale identique quel que soit l'unité spatiale i($I \to N(m, \sigma^2)$) avec :

$$\hat{m} = -1/K - 1$$

$$\widehat{\sigma^2} = \frac{(K^2 . \frac{1}{2} . \sum_{i \neq j}^{K} (w_{ij} + w_{ji})^2 - K . \sum_{i=1}^{i=K} (w_{i+} + w_{+i})^2 + 3{w_+}^2)}{(K-1)(K+1){w_+}^2} - \hat{m}^2$$

$$w_{i+} = \sum_{j=1}^{K} w_{ij} \quad , w_{+j} = \sum_{i=1}^{K} w_{ij}$$

L'indice de Moran mesure la similitude entre les unités spatiales voisines, son interprétation est similaire a celle d'un coefficient de corrélation (Gaudart (2007)):

I< 0 ⇔ Autocorrelation spatiale négative, donc les unités spatiales voisines sont différentes.

I ≃ 0 ⇔ il n'y'a aucune corrélation entre les unités spatiales voisines, et le model spatial est parfaitement aléatoire.

I> 0 ⇔ Les unité spatiales voisines sont similaires (existence d'un pattern sous forme d'un cluster d'unités spatiales)

2-2- Test d'ajustement de chi-deux de Pearson

Au lieu d'utiliser un coefficient d'autocréation spatiale, certains auteurs proposent des statistiques d'ajustement estimant l'écart entre les valeurs observées et les valeurs théoriques issues d'un modèle probabiliste.

La statistique d'adéquation la plus utilisé est la statistique du Chi-deux (χ^2) de Pearson donnée par :

$\chi^2 = \sum_{i=1}^{i=K} \frac{(O_i - E_i)^2}{E_i}$

où K= nombre d'unités spatiales.

$O_i = nombre\ de\ cas\ observés\ dans\ l'unité\ spatiale.$

$E_i = nombre\ de\ cas\ \ attendues\ sous\ H_0$

Sous l'hypothèse H_0 de risque constant, les nombres de cas attendus (E_i) sont issus d'une distribution de poisson et la statistique χ^2 suit une loi de chi-deux à m degré de liberté ($\chi^2 \to \chi^2_m$).

Le rejet de l'hypothèse de risque constant ou de modèle aléatoire de poisson suggère l'existence de clusters. La statistique du χ^2 fournit un test acceptable de détection globale de clusters mais elle n'est pas capable de repérer le caractère spatial des écarts au modèle théorique, c'est-à-dire si plusieurs unités spatiales présentent un écart important par rapport au modèle théorique, la statistique reste inchangée que ces u.s. soient contiguës (suggérant un cluster) ou non.

2-3- Test de Tango

Tango (1995) a généralisé spatialement le test précédent (test de chi-deux de Pearson) en pondérant les écarts par la proximité. La statistique dite de Tango (noté T) s'obtient de la manière suivante:

On considère $P_1 = \left(\frac{O_1}{O_+} \dots \frac{O_K}{O_+}\right)$ le vecteur des proportions observées où $O_+ = \sum_{i=1}^{k} O_i$ est le nombre total de cas observé, et $P_2 = \left(\frac{n_1}{n_+} \dots \frac{n_K}{n_+}\right)$ le vecteur des proportions attendues avec $n_+ = \sum_{i=1}^{k} n_i$ la population totale exposée au risque.

Sous l'hypothèse H_0 de risque constant, les proportions attendues sont issues d'une distribution multinomiale. La statistique de Tango compare les proportions observées et les proportions attendues et elle est définie par :

$T = \sum_{i,j}^{k} w_{ij} \left(\frac{O_i}{O_+} - \frac{n_i}{n_+}\right)\left(\frac{O_j}{O_+} - \frac{n_j}{n_+}\right)$

Sous l'hypothèse H_0 de risque constant, la variable T est asymptotiquement normale $T \to N(E(T), V(T))$ avec

$E(T) = \frac{1}{O_+} tr(WV_{P_2}), \quad V(T) = \frac{1}{O_+^2} tr\left[(WV_{P_2})^2\right] \qquad V_{P_2} = diag(P_2) - P_2 \acute{P}_2$

La forme standard de la statistique de Tango est donnée par :

$$T^* = \frac{T - E(T)}{\sqrt{V(T)}}$$

Tango (1990) a proposé une approximation par la loi de chi-deux de la statistique donnée par :

$v + T^*\sqrt{2v} \rightarrow \chi^2_v$

où $v = 8\left(2\sqrt{2}\frac{tr\left[\left(WV_{P_2}\right)^3\right]}{\left(tr\left[\left(WV_{P_2}\right)^2\right]\right)^{1,5}}\right)^{-2}$ est le degré de liberté.

1- Identification des clusters

Plusieurs méthodes ont été développées pour le traitement de données continues et d'autres pour les données qualitatives, mais ces méthodes ne sont pas fiables pour les deux types de données à la fois (cas le plus fréquent dans la pratique).

La méthode de clusters en deux étapes (Two-step clustering) introduite par Chiu et al. (2001) a pour but l'identification de clusters pour les données mixtes. Donc les données seront organisées en groupes (clusters), les membres de chaque cluster sont très similaires entre eux et très dissimilaires avec les membres des autres clusters ; ceci sur la base d'une ou plusieurs variables discriminantes. Le nom de la méthode signifie que son algorithme s'applique en deux étapes :

Etape 1 : les individus sont attribués à des pré-clusters

Etape 2 : Les pré-clusters sont clustérisés une deuxième fois à l'aide de l'algorithme hiérarchique (Abu Abbas (2008)).

Cette méthode implémentée sur le package SPSS suppose toutes les variables indépendantes, les variables continues normalement distribuées, et les variables catégorielles de distribution multinomiale (Voir (Bacha et al. (2007)) pour plus de détails).

Résultats obtenus

2- Application

4-1- Données

L'étude porte sur 527 malades hospitalisés (98% sont des femmes) et opérés du cancer de la thyroïde au service endocrinologie du centre Pierre et Marie Curie (CPMC) de l'hôpital Mustapha Bacha à Alger sur la période 2007-2011. Les données englobent le nombre de cas

par wilaya (Origine géographique), et la population exposée au risque et par conséquent l'incidence.

4-2- Méthodes et logiciels

Nous avons utilisé Excel pour calculer la statistique du Chi-deux de Pearson, l'indice de Moran et la statistique de Tango. La matrice de proximité utilisé est la matrice dichotomique car la seule information disponible est l'origine géographique, elle est définie par :

$$w_{ij} = \begin{cases} 1 & si\ la\ wilaya\ i\ a\ des\ frontieres\ avec\ la\ wilaya\ j \\ 0 & sinon \end{cases}$$

On a supposé que la wilaya i n'a pas de frontières avec elle-même ce qui implique

$w_{ii} = 0$. Cette procédure a facilité les calculs car la matrice résultante est symétrique.

Le logiciel SPSS 17 a été utilisé pour identifier les groupes identiques de cancers de la Thyroïde en appliquant la méthode décrite précédemment dite de clusters en deux étapes.

Le logiciel SatScan a été utilisé pour confirmer l'existence et la significativité des clusters. Ce logiciel nous permet d'appliquer la méthode de balayage de Kulldorff (méthode de détection locale,Waller &Gotway (2004)) qui consiste à regrouper les différentes unités spatiales voisines en clusters potentiels.

4-3- Résultats

Les trois méthodes utilisées rejettent l'hypothèse des risques constants et confirme la non existence d'un pattern aléatoire.

La statistique de chi-deux de Pearson ($\chi^2 = 150{,}598$) est strictement supérieure à la valeur de chi-deux tabulée ($\chi^2_{0.95}(2) = 5.598$) donc le nombre de cas de malades observés ne se distribue pas aléatoirement et on rejette l'hypothèse de risque constant.

L'indice de Moran I=0.4913 est positif et supérieur à sa valeur moyenne théorique (E(I)=0.0212), ceci suggère une hétérogénéité spatiale (tendance à une clusterisation).

Pour la statistique de Tango (T=0.043), l'approximation par la loi de chi-deux donne une valeur de 307,902 qui est strictement supérieure au quantile d'ordre 95% d'une gamma (6,3, 0.5).

L'utilisation de la méthode de clusters en deux étapes met en évidence cinq groupes homogènes par rapport aux variables taux d'incidence et population qui sont :

Cluster 1	2, 6, 9, 10, 15, 16, 28, 33

Cluster 2	14, 26, 30, 34, 35, 42, 44, 47

Cluster 3	4, 7, 12, 18, 21, 22, 23, 27, 29, 32, 39, 43, 48

Cluster 4	5, 13, 17, 19, 25, 31

Cluster 5	1, 3, 8, 11, 20, 24, 36, 37, 38, 40, 41, 45, 46

L'analyse par le logiciel SatScan a donné les résultats suivants

- La recherche de cluster spatiaux montre l'existence d'un cluster significatif (p-value <0.00000000000000010) regroupant les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6. Il contient 389 cas.
- La recherche de cluster temporels montre un cluster significatif entre 2007 /2011 regroupant 441 cas, avec une p-value égale à0,001.
- La recherche de clusters spatiaux- temporels montre un cluster significatif (p-value <0.00000000000000010) entre 2007 et 2011 regroupant les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6, 19, 14, 18, 17.Il contient 371 cas.

Discussion des résultats :

Nous rappelons que nous avons travaillé sur les seuls malades hospitalisés au CPMC sur la période 2007-2011. Face à l'augmentation du nombre des cancers thyroïdiens, vu l'impossibilité de respecter les délais de traitement adéquats et vu que les structures de médecine nucléaire de Constantine et de Tlemcen étaient fonctionnelles, le CPMC a délimité les wilayas prises en charge : au nord la région centre et tout le sud Algérien (à l'Est BBA, à l'ouest Tiaret).

Hors sur la même période d'autres hôpitaux ont aussi recruté des malades du cancer de la thyroïde. Donc le taux d'incidence ne pourrait qu'augmenter et d'autres clusters potentiels pourraient émerger si on généralise l'étude au niveau national.

Il ressort de cette étude que certaines willayas (2, 6, 15) appartenant au cluster significatif sont déjà connues comme régions endémiques du goitre, le principal facteur de risque étant la carence d'iode. Est-ce que c'est toujours le cas aujourd'hui ? Peut-on considérer qu'un goitre conduit nécessairement à un cancer de la thyroïde ?

Est-ce que l'utilisation d'une autre matrice de proximité (changement de distance) conduirait aux mêmes résultats ?

Les réponses à ces questions nécessitent un complément d'informations et donc d'autres investigations par le biais d'enquêtes épidémiologiques par exemple. C'est pour cela que cette analyse pourrait être considérée comme un test de dépistage de la maladie et devrait être suivie par des études plus ciblées, avec des informations plus détaillées pour confirmer ou infirmer les hypothèses qu'elle dégage.

Originalité des résultats

C'est la première étude de ce type qui est menée en Algérie et donc les résultats obtenus sont nouveaux.

Impact des résultats socio-économiques :

-Les résultats obtenus pourraient orienter les différents centres de décisions dans les mesures à prendre (enquêtes épidémiologiques,…) pour cerner l'évolution de cette maladie qui prend des ampleurs inquiétantes.

_Les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6**,** devraient retenir l'attention des autorités de la santé publique dans la mesure où on a démontré dans cette étude qu'il y'a une concentration anormale de cas de cancer de la thyroïde dans ces régions ; dans l'espace et dans le temps.

Présentation des principales conclusions

1. L'ensemble des méthodes statistique utilisées rejettent l'hypothèse d'un risque constant et suggère une tendance à une concentration de cas de cancer de la thyroïde.
2. La recherche de clusters spatiaux montre l'existence d'un cluster significatif regroupant les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6. Il contient 389 cas.
3. La recherche de clusters temporels montre un cluster significatif entre 2007 /2011 regroupant 441 cas, avec une p-value égale à0,001
4. La recherche de clusters spatiaux- temporels montre un cluster significatif entre 2007 et 2011 regroupant les wilayas 16, 35, 9, 10, 26, 42, 15, 44, 34, 28, 2, 38, 6, 19, 14, 18, 17, il contient 438 cas.
5. Nous faisons remarquer que la constitution de l'échantillon se poursuit dans le but d'atteindre les 1000 malades. Par ailleurs le développement des outils statistiques dans le cadre du balayage statistique et l'analyse bayésienne afin d'identifier et étudier l'émergence de clusters se fait dans le cadre d'une thèse de doctorat d'Etat en cours de réalisation

Pistes de recherche future :

- Elargir cette étude à l'ensemble es hôpitaux traitant de cette pathologie. Mais pour cela il faudrait permettre l'accès à l'information. Par la suite des enquêtes épidémiologiques pourraient être envisagées afin d'infirmer ou confirmer les résultats émanant de ce projet.

- la même problématique pourrait être adoptée pour étudier l'évolution des cas des autres types de cancer.

-le développement des outils statistiques dans le cadre du balayage statistique et l'analyse bayésienne afin d'identifier et étudier l'émergence de clusters se poursuit dans le cadre d'une thèse de doctorat d'Etat en cours de réalisation.

Difficultés rencontrées lors de la réalisation de la recherche:

1. L'accès à l'information dans les hôpitaux autres que le CPMC
2. Problèmes financiers pour recruter du personnel de saisie, se déplacer pour des communications, acheter des logiciels etc…

Ressources bibliographiques

1- **Abu Abbas O. "Comparison beteen data clustering algorithms". The international Arab Journal of Information Technology , 5 (2008) N°3, 320-325.**

2- **Bacha P. Brunck T., Delany, J. Clustering methods and their uses daylight CIS, INC, 2007.**

3- **Besag, J & Newell, J. "The detection of clusters in rare diseases". J. R. Statist. Soc. A (1991) 154, Part 1, 143-155.**

4- **Chiu T., Fang D., Chen J., Wang Y., Jeris C. "A robust and scalabale clustering algorithm for mixed type attributes in large database environment". In proceeding for the 7th ACM SIGKDD international conference in knowledge discovery and data mining association for computing machinery. SanFrancisco CA (2001), PP 263-268.**

5- **Abid L. Epidémiologie des cancers en Algérie. Problématique des registres des cancers. Af J of Cancer 2009.vol 1(2):9-103**

6- **L. Amokrane cancer vésiculaire de la thyroïde en Algérie: aspects epidemiologiques, cliniques, et pronostiques. Thèse de DESM 1999**

7- **Bakiri F, Djemli FK, Amokrane L, Djidel FK. Relative roles of endemic goiter and socioeconomic developement status in prognosis of thyroid carcinoma. Cancer 1998: 1146-53**

8- Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. Environ Health Perspect 2008;116(8):1105-10.

9- FZ Benserai. Carcinomes primitifs de la thyroide. Thèse de DESM Alger 2004

10- N Berber le cancer thyroidien différencié surveillance biologique. Thèse de DESM Alger 1998

11- FK Hamzaoui Djemli. Aspects anatomo-cliniques, thérapeutiques et pronostiques des cancers thyroïdiens en Algérie.1981. thèse de DESM

12- Davies L, WelchHG. Increasing incidence of thyroid cancer in the United States, 1973-2002. JAMA 2006; 295(18) 2164-7.

13- Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial epidemiology: methods and applications. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications.Oxford: Oxford University Press;2000.

14- Gaudart, J. et al. " Détection de clusters spatiaux sans point source prédéfini : utilisation de cinq méthodes et comparaison de leurs résultats". Revue d'épidémiologie et de Santé Publique 55(2007) 297-306.

15- Tango T., "A class of tests for detecting a general and focused clustering of rare diseases". Statistics in Medecine 14(1995); 2323-2334.

16- Waller L.A.,&Gotway, C.A. "Applied Spatial Statistics for Public Health ". John Willey &sons, Inc. (2004).

17- Goria S, Le Tertre A. Les études locales autour d'un point source - Les différentes méthodes statistiques, leurs avantages et leurs inconvénients. Note méthodologique. http://www.invs.sante.fr

18- S Hadjarab. Les cancers différenciés d la thyroide (cancers papillaires et vésiculaires) caractéristiques cliniques et pronostics. Thèse de DESM Alger 2008

19- Lawson AB, Biggeri A, Williams FLR. A review of modelling approaches in health risk assessment around putative sources. In: Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R, (dir.). Disease mapping and risk assessment for public health. Chichester: Wiley; 1999. p. 231-45.

20- Guihenneuc-Jouyaux C. Statistical modelization of geographic variations: a major challenge in epidemiology and statistics. Rev Epidemiol Santé Publique 2002;50(5):409-12.p.

21- Richardson S. Problèmes méthodologiques dans les études écologiques santé-environnement. CR Acad Sci Paris,Sciences de la Vie/Life Sciences 2000;323:611-6.

22- Best NG, Cockings S, Bennett JE, Wakefield JC, Elliott P. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. Journal of the Royal Statistical Society, Series A 2001;164:155-74.

23- Cordier S, Chevrier C, Robert-Gnansia E, Lorente C, Brula P, Hours M. Risk of congenital anomalies in the vicinity of municipal solid waste incinerators. Occup Environ Med 2004;61(1):8-15.

24- aheswaran R, Haining RP, Pearson T, Law J, Brindley P, Best NG. Outdoor NOx and stroke mortality: adjusting for small area level smoking prevalence using a Bayesian approach. Statistical methods in medical research 2006;15(5):499-516.

25- Nieuwenhuijsen MJ, Toledano MB, Bennett J, Best N, Hambly P, de HCet al . Chlorination disinfection by-products and risk of congenital anomalies in England and Wales. Environ Health Perspect 2008;116(2):216-22.

26- Richardson S, Monfort C, Green M, Draper G, Muirhead C. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. Stat Med 1995;14(21-22):2487-501.

27- Fabre P, Daniau C, Goria S, de Crouy-Chanel P, Empereur-Bissonnet P. Étude d'incidence des cancers à proximité des usines d'incinération d'ordures ménagères; http://www.invs.sante.fr

28- Wakefield J. Disease mapping and spatial regression with count data. Biostatistics 2007;8(2):158-83.

29- Salway R. Statistical issues in the analysis of ecological studies, Ph.D. Thesis Imperial College School of Medicine,University of London; 2003.

30- Sassolas G, Hafdi-Nejjari Z Remontet L, Bossard N, Belo A, Berger-Dutrieux N, Decaussin–Petrucci M, Bournaux C, Peix JL, Orgiazzi J, Borson- Chazot F. Thyroid cancer : is the incidence abating ? Eur J Endoc 2009

31- Wakefield JC, Salway R. A statistical framework for ecological and aggregate studies. Journal of the Royal Statistical Society, series A 2001;164:119-37.

32- Salway R, Wakefield J. A hybrid model for reducing ecological bias. Biostatistics 2008;9(1):1-17.

33- Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. Journal of the Royal Statistical Society, Series A 2008;171(1):159-78.

34- Best N, Ickstadt K, Wolpert R. Spatial Poisson regression for health and exposure data measured at disparate resolutions. Journal of the American Statistical Society 2000;95:1076-88.

35- Fortunato L, Guihenneuc-Jouyaux C, Tirmarche M, Laurier D, Hémon D. Misspecification of within-area exposure distribution in ecological Poisson models. Environ Ecol Stat 2009;16:341-53.

36- Fleuret S, Thouez JP. Géographie de la santé, un panorama. Paris : Economica; 2007.

37- Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect 2004;112(9):1007-15.

38- Béguin M, Pumain D. La représentation des données géographiques: statistique et cartographie. Armand Colin éd.;1994. 192 p.

39- Bertin J. Sémiologie graphique: les diagrammes, les réseaux, les cartes. Paris : EHESS; 1999.

40- Jenks GF, Caspall FC. Error on choroplethic maps: definition, measurement, reduction. Annals of the Association of American Geographers 1971;61(2):217-44.

41- Colonna M. Habilitation à diriger des recherches Université Joseph Fourier, Grenoble; 2006.

42- Pumain D, Saint-Julien T. L'analyse spatiale, localisation dans l'espace. Armand Colin éd. Paris: 2008. 166 p.

43- Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis 1995;27(4):286-306.

44- Vandentorren S. Exposition environnementale à l'amiante chez les personnes riveraines d'anciens sites industriels et affleurements naturels. Étude cas-témoins à partir des données du Programme national de surveillance du mésothéliome (2009). http://www.invs.sante.fr

45- Counil E, Daniau C, Isnard H. Étude de santé publique autour d'une ancienne usine de broyage d'amiante : le Comptoir des minéraux et matières premières à Aulnay-sous-Bois (Seine-Saint-Denis) - Pollution environnementale entre 1938 et 1975 : impacts sanitaires et recommandations (2007).254 p. http://www.invs.sante.fr.

46- De Crouy-Chanel P. Étude SIG de la corrélation entre exposition indirecte à l'amiante et asbestose. Geomatique Expert 2007;54:28-32

47- Poulstrup A, Hansen HL. Use of GIS and exposure modeling as tools in a study of cancer incidence in a population exposed to airborne dioxin. 2004;1129:1032.

48- Yu CL, Wang SF, Pan PC, Wu MT, Ho CK, Smith TJet al. Residential exposure to petrochemicals and the risk of leukemia: using geographic information system tools to estimate individual-level residential exposure. Am J Epidemiol 2006;164(3):200-7.

49- Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer Pet al. A review of land-use regression models toassess spatial variation of outdoor air pollution. Atmos Environ 2008;42:7561-78.

50- Best N, Ickstadt K, Wolpert R, Briggs D. Combining models of health and exposure data: the SAVIAH study. In:Elliott P, Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications.Oxford: Oxford University Press; 2000. p. 393-414.

51-]Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. Hoboken, New Jersey: Wiley; 2004.

52- Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial epidemiology: methods and applications. Oxford: Oxford University Press; 2000.

53- Huang L, Pickle LW, Das B. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Stat Med 2008;27(25):5111-42.

54- Demattei C. Détection d'agrégats temporels et spatiaux, Ph.D. Thesis Université Montpellier 1 UFR de médecine,Montpellier; 2006.

55- Wakefield JC, Kelsall JE, Morris SE. Clustering, cluster detection, and spatial variation in risk. In: Elliott P,Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications. Oxford: Oxford University Press; 2000. p. 128-52.

56- Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med 1995;14(8):799-810.

57- Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med 2006;25(22):3929-43.

58- Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 1997;26(6):1481-96.

59- Kulldorff M. SaTScan User Guide for version 7.0; 2006.

60- Bivand RS, Pebesma EJ, Gomez-Rubio V. Applied spatial data analysis with R. Springer; 2008.

61- Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 2005;4:11.

62- Assuncao R, Costa M, Tavares A, Ferreira S. Fast detection of arbitrarily shaped disease clusters. Stat Med 2006;25(5):723-42.

63- Morris SE, Wakefield JC. Assessment of disease risk in relation to a pre-specified source. In: Elliott P, Wakefield JC,Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications.Oxford: Oxford University Press;2000;153-84.

64- Bithell JF, Stone RA. On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. J Epidemiol Community Health 1989;43(1):79-85.

65- Stone RA. Investigations of excess environmental risks around putative sources: statistical problems and aproposed test. Stat Med 1988;7(6):649-60.

66- Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford Jet al. Cancerincidence near municipal solid waste incinerators in Great Britain. Br J Cancer 1996;73(5):702-10

67- Yaker A. Algeria Histopathology Study. Algiers Oran Constantine 1966-1975.Cancer occurence in developing countries 1986 Parki DM Ed. IARC Scientf Pub N75:27-31.

ANNEXE 1 :

RESULTATS PRELIMINAIRES (présentés en session ordinaire de travail du groupe, Avril 2012)

LA PREMIERE PHASE DE CE TRAVAIL ETAIT DESCRIPTIVE ET EXPLORATOIRE: IL S'AGISSAIT DE CONSTRUIRE L'ECHANTILLON DE TRAVAIL, IDENTIFIER ET SELECTIONNER LES VARIABLES SIGNIFICATIVES. 77 VARIABLES ONT ETE RETENUES POUR LA CONSTRUCTION DU TABLEAU STATISTIQUE.

AINSI IL A ÉTÉ PROCEDE A LA SAISIE D'UN ECHANTILLON DE 368 MALADES RELEVANT DU SERVICE ENDOCRINOLOGIE AU CPMC DE L'HOPITAL MUSTAPHA SOUS LA DIRECTION ET RESPONSABILITE DU PROFESSEUR MOURAD SEMROUNI.

LES INFORMATIONS ENREGISTREES ONT ÉTÉ FAITES SUR LA BASE DE LA FICHE : ITEMS DES CANCERS DIFFERENCIES DE LA THYROIDE ELABOREE AU NIVEAU DU SERVICE DU PROFESSEUR M.SEMROUNI.

APRES CORRECTION DES REDONDANCES NOUS AVONS OBTENU UN ECHANTILLON DE 307 MALADES ET 70 VARIABLES.

POUR CETTE PREMIERE PHASE NOUS N'AVONS EXPLOITE QUE QUATRE VARIABLES : SEXE, AGE, ORIGINE GEOGRAPHIQUE, ET ANTECEDENTS FAMILIAUX**.**

Tableau N0 1 : Sexe

	Fréquence	Pourcentage	Pourcentage valide	Pourcentage Cumulatif
Femme	275	89,6	89,6	89,6
Homme	32	10,4	10,4	100,0
Total	307	100,0	100,0	

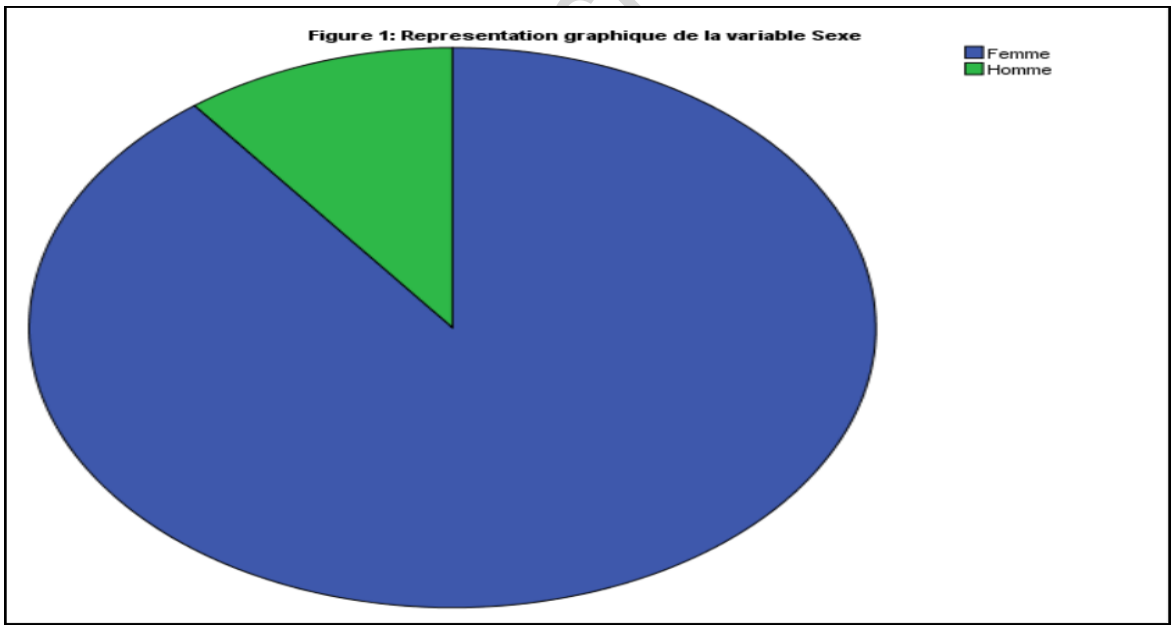


Figure 2: Representation graphique de la variable 'Origine géographique'

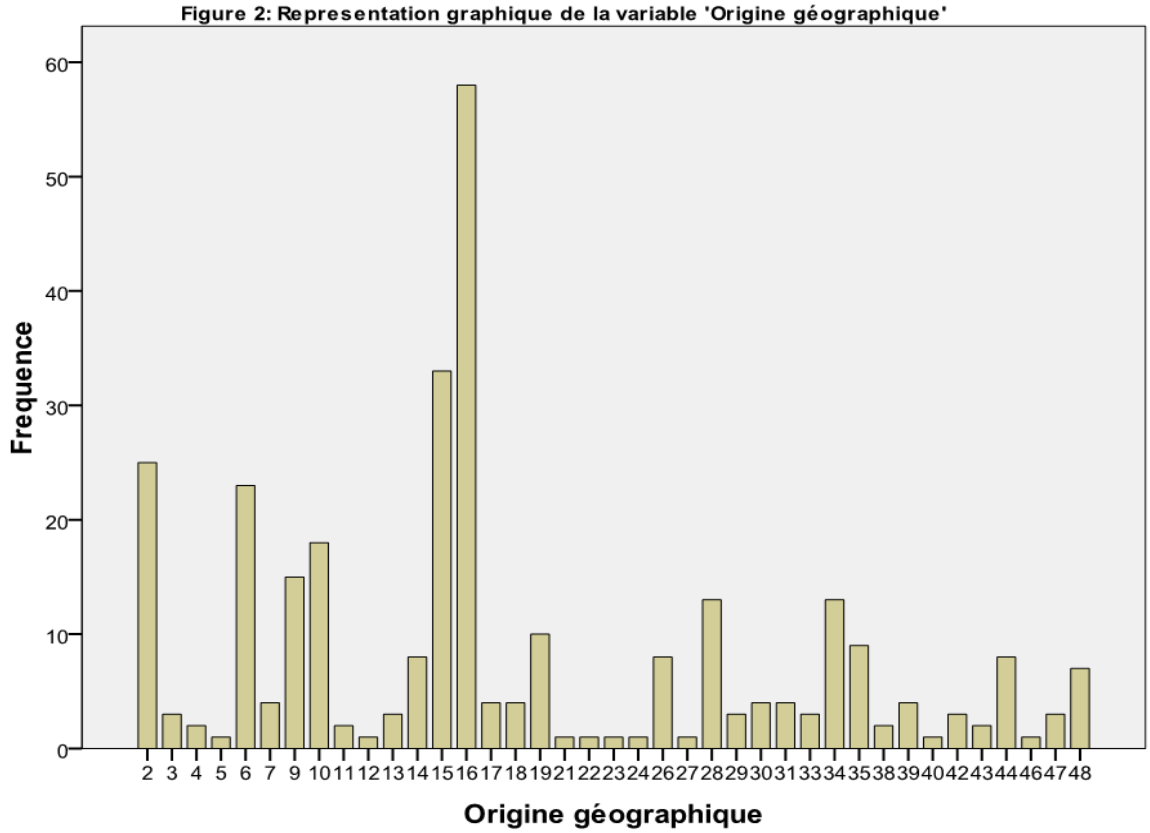


Tableau N 4 ; les fréquences des malades par tranche d'age

Age	Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulatif
0-10	2	,7	,7	,7
10-20	11	3,6	3,6	4,2
20-30	56	18,2	18,2	22,5
30-40	69	22,5	22,5	45,0
40-50	72	23,5	23,5	68,4
50-60	62	20,2	20,2	88,6
60-70	22	7,2	7,2	95,8
70-80	13	4,2	4,2	100,0
Total	307	100,0	100,0	

Figure 3: repartition des frequences de malades par tranche d'age

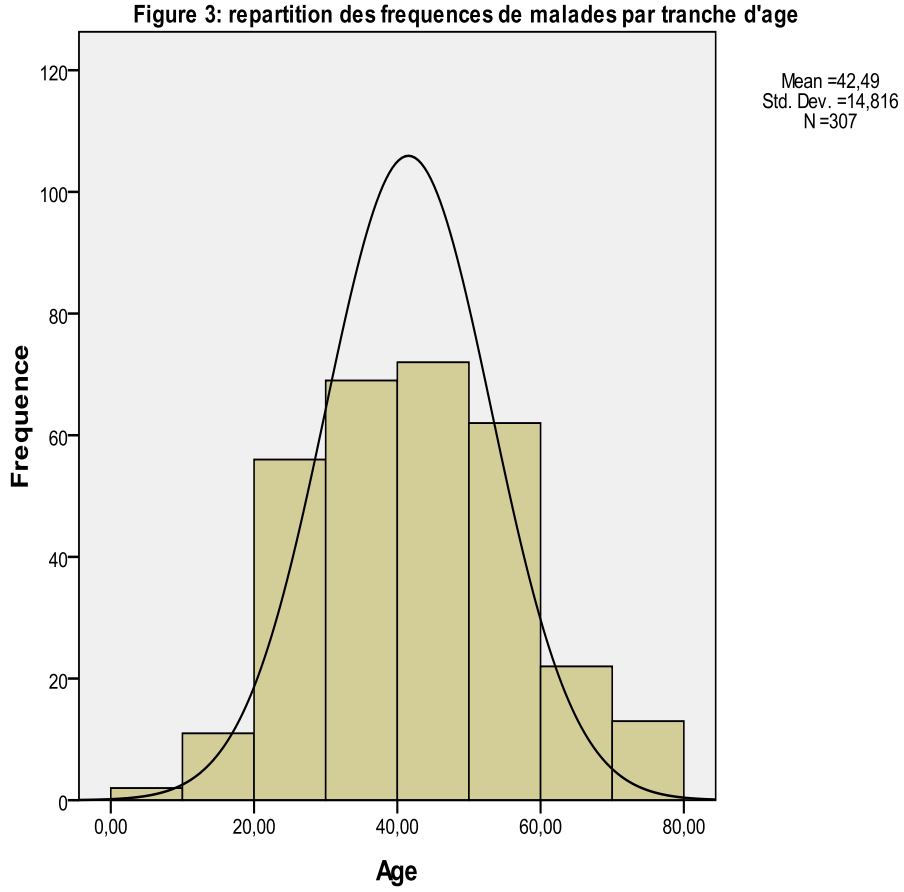
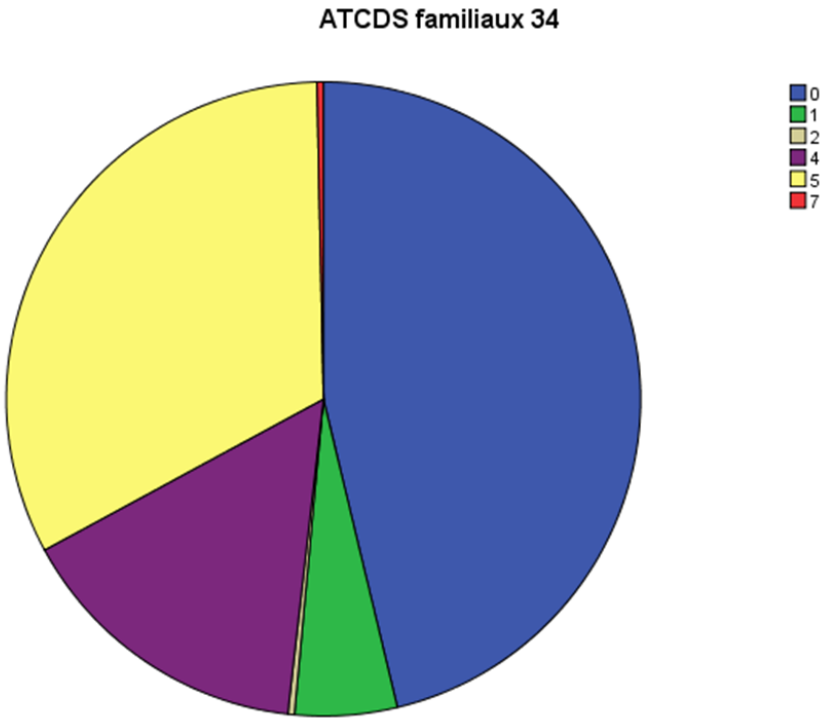


Tableau n 3 :Statistiques de base pour la variable 'Age'

N Taille de l'échantillon	307
La moyenne	42,07
Erreur standard de la moyenne	,821
La médiane	42,00
Le mode	26
L'écart type	14,392
La variance	207,129
Skewness	,194
Std. Error of Skewness	,139
Kurtosis	-,348
Std. Error of Kurtosis	,277
Range	77
Minimum	3
Maximum	80

ATCDS familiaux 34

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	142	46,3	46,3	46,3
	1	16	5,2	5,2	51,5
	2	1	,3	,3	51,8
	4	47	15,3	15,3	67,1
	5	100	32,6	32,6	99,7
	7	1	,3	,3	100,0
	Total	307	100,0	100,0	



DANS UNE SECONDE ETAPE NOUS AVONS ESSAYE D'IDENTIFIER DES GROUPES HOMOGENES EN FONCTION DE LEURS SIMILARITE SUR UN ENSEMBLE DE CARACTERISTIQUES.
POUR CELA NOUS AVONS EU RECOURS A L'ANALYSE DE REGROUPEMENT LAQUELLE EST UNE TECNIQUE QUI PERMET D'IDENTIFIER DES REGROUPEMENTS D'INDIVIDUS OU D'OBJETS QUI SE RESSEMBLENT OU PARTAGENT DES ATTRIBUTS COMMUNS.
1- QUELLES SONT LES WILAYAS QUI SE RESSEMBLENT PAR RAPPORT AU TAUX D'INCIDENCE du cancer de la thyroïde?
POUR CELA NOUS AVONS CONSTRUIT LE TABLEAU SUIVANT :

WILAYA	MALADES	POPULATION	Taux incidence
1		3997.14	
2	25	10020.88	.02
3	3	4556.02	.01
4	2	6216.12	.00
5	1	11197.91	.00
6	23	9125.77	.03
7	4	7213.56	.01
8		2700.61	

9	15	10029.37	.01
10	18	6955.83	.03
11	2	1766.37	.01
12	1	6487.03	.00
13	3	9491.35	.00
14	8	8468.23	.01
15	33	11276.08	.03
16	58	29881.45	.02
17	4	10921.84	.00
18	4	6369.48	.01
19	10	14899.79	.01
20		3306.41	
21	1	8986.80	.00
22	1	6047.44	.00
23	1	6094.99	.00
24	1	4824.30	.00
25		9384.75	
26	8	8199.32	.01
27	1	7371.18	.00
28	13	9905.91	.01
29	3	7840.73	.00
30	4	5585.58	.01
31	4	14540.78	.00
32		2286.24	
33	3	523.33	.06
34	13	6284.75	.02
35	9	8020.83	.01
36		4084.14	
37		491.49	
38	2	2944.76	.01
39	4	6475.48	.01
40	1	3866.83	.00
41		4381.27	
42	3	5910.10	.01
43	2	7668.86	.00
44	8	7660.13	.01
45		1928.91	
46	1	3712.39	.00
47	3	3635.98	.01
48	7	7261.80	.01

L'APPLICATION DE CETTE TECHNIQUE NOUS APERMIS D'IDENTIFIER DEUX GROUPES DISTINCTS **:**

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	32	82,1%	66,7%
	2	7	17,9%	14,6%
	Combined	39	100,0%	81,3%
Excluded Cases		9		18,8%
Total		48		100,0%

Centroids

		tauincidence	
		Mean	Std. Deviation
Cluster	1	,5727	,38837
	2	2,8955	1,29273
	Combined	,9896	1,09656

CLUSTER 1: (3, 4, 5, 7, 11, 12, 13, 14, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29, 31, 35, 38, 39, 40,42, 43, 44, 46, 47, 48)
CLUSTER 2:(2, 6, 9, 10, 15, 16,33, 34)

ANNEXE 2 :

Communication faite lors des journées statistiques organisées par la Société Française de Statistique à Toulouse : Détection et identification des clusters de cancer de la thyroïde en Algérie (Mai 2013)

Résumé

Les cancers en Algérie sont devenus émergents et constituent une préoccupation majeure. Le cancer de la thyroïde augmente de manière spectaculaire depuis une quinzaine d'années, c'est le troisième cancer chez la femme d'après le registre des tumeurs d'Alger. Entre 2007 et 2010 il a été multiplié par 6 et ce seulement au service endocrinologie du centre Pierre et Marie Curie d'Alger. es observations nous amènent à se poser plusieurs questions :

- Certaines wilayas ont-elles un nombre de cas de cancers excessif?
- Les cas de cancer de la thyroïde sont-ils anormalement concentrés?
- La distribution spatiale de ces cas, est-elle aléatoire ?

Un cluster est une organisation spatiale définie comme un agrégat, un regroupement de cas proches les uns des autres ; la proximité étant définie au sens d'une distance géographique.

Répondre à ces questions revient à décrire l'hétérogénéité Spatiale.

Les méthodes de détection de clusters identifient les regroupements de cas incohérents sous l'hypothèse nulle d'absence de clusters et évaluent leur niveau de significativité .Quant aux méthodes de détection globale d'agrégation de cas, elles étudient la corrélation spatiale et détectent la tendance des cas à la clustérisation.

Dans ce travail nous allons utiliser des méthodes de détection d'agrégats de cas , dont les statistiques sont souvent basées sur les distances afin d'analyser l'existence des clusters de cancers de la thyroïde en Algérie. Ce sont le test du coefficient de corrélation de Moran, le test de Tango et la méthode de balayage de Kulldorff. Le test d'ajustement a été utilisé pour vérifier l'hypothèse des risques constants de l'incidence du cancer de la thyroïde.

Abstract

Cancers in Algeria are emerging and are becoming a major concern. The thyroid cancer increases dramatically over the past fifteen years, it is the third most common cancer of women as indicated by the tumor registry of Algiers. It was multiplied by 6 between 2007 and 2010 and this is only at Pierre and Marie Curie center of Algiers. These observations lead us to ask several questions:

- Some states have they an excessive number of cancer cases?
- are the cases of thyroid cancer unusually concentrated?
- the spatial distribution of these cases, is it random?

A cluster is defined as a spatial organization defined as an aggregate, a combination of cases close to each other, the proximity is defined in the sense of geographical distance.

Answering these questions is equivalent to describe the spatial heterogeneity.

Detection methods of clusters detect the groups of cases inconsistent under the null hypothesis of no clusters (constant risks) and evaluate their significance level, while global detection methods study the overall aggregation of cases and they investigate the spatial correlation and detect trends of cases to clustering.

In this work we will use methods of detecting clusters of cases where the statistics are often based on distances in order to analyze the existence of clusters of thyroid cancer in Algeria. It is the test of the correlation of Moran, the test of Tango and Kulldorff scan method. The goodness of fit test of chi-square was used also to test the hypothesis of constant risk of the incidence of thyroid cancer.

Mot clés: Cluster, taux d'incidence, proximité, indice de Moran, statistique de Tango.

1- Introduction

Les cancers en Algérie sont devenus émergents et constituent une préoccupation majeure. Le cancer de la thyroïde augmente de manière spectaculaire depuis une quinzaine d'années, c'est le troisième cancer chez la femme d'après le registre des tumeurs d'Alger. Entre 2007 et 2010 il a été multiplié par 6 et ce seulement au service endocrinologie du centre Pierre et Marie Curie d'Alger. Ces observations nous amènent à se poser plusieurs questions :

- Certaines wilayas ont-elles un nombre de cas de cancers excessif?

- Les cas de cancer de la thyroïde sont-ils anormalement concentrés?
- La distribution spatiale de ces cas, est-elle aléatoire ?

Répondre à ces questions revient à décrire l'hétérogénéité Spatiale. Un cluster est une organisation spatiale définie comme un agrégat, un regroupement de cas proches les uns des autres ; la proximité étant définie au sens d'une distance géographique. Les méthodes de détection de clusters identifient les regroupements de cas incohérents sous l'hypothèse nulle d'absence de clusters et évaluent leur niveau de significativité .Quant aux méthodes de détection globale d'agrégation de cas, elles étudient la corrélation spatiale et détectent la tendance des cas à la clustérisation.

Le but de ce travail est double tester la présence de clusters de thyroïde par les tests globaux (test de Moran et test Tango) et le test du Chi-deux de Pearson ; localiser les clusters pouvant exister par la méthode des clusters en deux étapes ; confirmer la significativité des clusters potentiels par la méthode de balayage de Kulldorff.

2- détection de clusters (agrégats spatiaux)

Différentes méthodes statistiques ont été développées (Besag & Newell(1991)) pour l'identification des « structures spatiales (clusters) », Pour vérifier l'existence ou la non existence d'agrégats spatiaux on a choisi les tests de détection globale (Gaudart et al 2007). Ces tests se caractérisent par la recherche d'une tendance générale à l'agrégation, ils reposent sur les hypothèses suivantes :

$\begin{cases} H_0 : Hypothese\ du\ risque\ constant\ (la\ non\ existance\ de\ clusters). \\ H_1 : Il\ ya\ tendance\ \text{à}\ la\ clusterisation. \end{cases}$

Ils estiment une statistique pour la zone étudiée dans le but de tester l'existence d'une hétérogénéité globale.

Pour effectuer les tests globaux, il est nécessaire de décrire la proximité entre les unités spatiales, cette dernière est donnée par la matrice de proximité noté W. W est une matrice carré et elle résume la relation entre chaque deux paires d'unités spéciales de la zone étudiée. Cette proximité peut être la distance entre deux unités spatiales ou par dichotomie en donnant la valeur 1 si la paire a des frontières en commun et zéro sinon (Waller & Gotway (2004)). Cela permet d'affecter un poids à chaque paire.

Trois tests ont été utilisés, le test de Moran, le test de chi-deux de Pearson et le test de Tango.

2-1- Test de Moran

Le test repose sur la statistique dite Indice de Moran (Waller & Gotway(2004)) noté I et définie par :

$$I = \frac{1}{w_+} . \frac{\sum_{i=1}^{i=K} \sum_{j=1}^{j=K} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{i=K} (y_i - \bar{y})^2 / K}$$

Où $K = nombre\ d'unités\ spatiale$

$w_{ij} = éléments\ de\ la\ matrice\ de\ proximité\ pour\ les\ unités\ spatiales\ i\ et\ j.$

$w_+ = \sum_{i,j=1}^{K} w_{ij}$

$y_i = \frac{o_i}{n_i} = \frac{nombre\ de\ cas\ observés\ dans\ l'unité\ spatiale\ i}{effectif\ de\ l'unité\ \ spatiale\ i}$

$\bar{y} = \frac{\sum_{i=1}^{i=K} y_i}{K} = moyenne\ des\ proportions\ sur\ l'ensemble\ des\ K\ unités\ s$patiales

La statistique I est donc une variable aléatoire. Sous l'hypothèse H_0, I suit une loi asymptotiquement normale identique quel que soit l'unité spatiale i($I \to N(m, \sigma^2)$) avec :

$\widehat{m} = -1/K - 1,$

$$\widehat{\sigma^2} = \frac{(K^2 . \frac{1}{2} . \sum_{i \neq j}^{K} (w_{ij} + w_{ji})^2 - K . \sum_{i=1}^{i=K} (w_{i+} + w_{+i})^2 + 3{w_+}^2)}{(K-1)(K+1){w_+}^2} - \widehat{m}^2$$

$$w_{i+} = \sum_{j=1}^{K} w_{ij} \ \ , w_{+j} = \sum_{i=1}^{K} w_{ij}$$

L'indice de Moran mesure la similitude entre les unités spatiales voisines, son interprétation est similaire a celle d'un coefficient de corrélation (Gaudart (2007)):

I< 0 ⇔ Autocorrelation spatiale négative, donc les unités spatiales voisines sont différentes.

I ≃ 0 ⇔ il n'y'a aucune corrélation entre les unités spatiales voisines, et le model spatial est parfaitement aléatoire.

I> 0 ⇔ Les unité spatiales voisines sont similaires (existence d'un pattern sous forme d'un cluster d'unités spatiales)

2-2- Test d'ajustement de chi-deux de Pearson

Au lieu d'utiliser un coefficient d'autocréation spatiale, certains auteurs proposent des statistiques d'ajustement estimant l'écart entre les valeurs observées et les valeurs théoriques issues d'un modèle probabiliste.

La statistique d'adéquation la plus utilisé est la statistique du Chi-deux (χ^2) de Pearson donnée par :

$\chi^2 = \sum_{i=1}^{i=K} \frac{(O_i - E_i)^2}{E_i}$

où K= nombre d'unités spatiales.

$O_i = nombre\ de\ cas\ observés\ dans\ l'unité\ spatiale.$

$E_i = nombre\ de\ cas\ \ attendues\ sous\ H_0$

Sous l'hypothèse H_0 de risque constant, les nombres de cas attendus (E_i) sont issus d'une distribution de poisson et la statistique χ^2 suit une loi de chi-deux à m degré de liberté ($\chi^2 \rightarrow \chi^2_m$).

Le rejet de l'hypothèse de risque constant ou de modèle aléatoire de poisson suggère l'existence de clusters. La statistique du χ^2 fournit un test acceptable de détection globale de clusters mais elle n'est pas capable de repérer le caractère spatial des écarts au modèle théorique, c'est-à-dire si plusieurs unités spatiales présentent un écart important par rapport au modèle théorique, la statistique reste inchangée que ces u.s. soient contiguës (suggérant un cluster) ou non.

2-3- Test de Tango

Tango (1995) a généralisé spatialement le test précédent (test de chi-deux de Pearson) en pondérant les écarts par la proximité. La statistique dite de Tango (noté T) s'obtient de la manière suivante:

On considère $P_1 = \left(\frac{O_1}{O_+} \ldots \frac{O_K}{O_+}\right)$ le vecteur des proportions observées où $O_+ = \sum_{i=1}^{k} O_i$ est le nombre total de cas observé, et $P_2 = \left(\frac{n_1}{n_+} \ldots \frac{n_K}{n_+}\right)$ le vecteur des proportions attendues avec $n_+ = \sum_{i=1}^{k} n_i$ la population totale exposée au risque.

Sous l'hypothèse H_0 de risque constant, les proportions attendues sont issues d'une distribution multinomiale. La statistique de Tango compare les proportions observées et les proportions attendues et elle est définie par :

$$T = \sum_{i,j}^{k} w_{ij} \left(\frac{O_i}{O_+} - \frac{n_i}{n_+}\right)\left(\frac{O_j}{O_+} - \frac{n_j}{n_+}\right)$$

Sous l'hypothèse H_0 de risque constant, la variable T est asymptotiquement normale $T \rightarrow N(E(T), V(T))$ avec

$E(T) = \frac{1}{O_+} tr(WV_{P_2}), \quad V(T) = \frac{1}{O_+^2} tr\left[(WV_{P_2})^2\right] \qquad V_{P_2} = diag(P_2) - P_2 \acute{P}_2$

La forme standard de la statistique de Tango est donnée par :

$T^* = \frac{T - E(T)}{\sqrt{V(T)}}$

Tango (1990) a proposé une approximation par la loi de chi-deux de la statistique donnée par :

$v + T^* \sqrt{2v} \rightarrow \chi^2_v$

où $v = 8\left(2\sqrt{2}\frac{tr\left[(WV_{P_2})^3\right]}{\left(tr\left[(WV_{P_2})^2\right]\right)^{1,5}}\right)^{-2}$ est le degré de liberté.

3- Identification des clusters

Plusieurs méthodes ont été développées pour le traitement de données continues et d'autres pour les données qualitatives, mais ces méthodes ne sont pas fiables pour les deux types de données à la fois (cas le plus fréquent dans la pratique).

La méthode de clusters en deux étapes (Two-step clustering) introduite par Chiu et al. (2001) a pour but l'identification de clusters pour les données mixtes. Donc les données seront organisées en groupes (clusters), les membres de chaque cluster sont très similaires entre eux et très dissimilaires avec les membres des autres clusters ; ceci sur la base d'une ou plusieurs variables discriminantes. Le nom de la méthode signifie que son algorithme s'applique en deux étapes :

Etape 1 : les individus sont attribués à des pré-clusters

Etape 2 : Les pré-clusters sont clustérisés une deuxième fois à l'aide de l'algorithme hiérarchique (Abu Abbas (2008)).

Cette méthode implémentée sur le package SPSS suppose toutes les variables indépendantes, les variables continues normalement distribuées, et les variables catégorielles de distribution multinomiale (Voir (Bacha et al. (2007)) pour plus de détails).

4- Application

4-1- Données

L'étude porte sur 527 malades hospitalisés (98% sont des femmes) et opérés du cancer de la thyroïde au service endocrinologie du centre Pierre et Marie Curie (CPMC) de l'hôpital Mustapha Bacha à Alger sur la période 2007-2011. Les données englobent le nombre de cas par wilaya (Origine géographique), et la population exposée au risque et par conséquent l'incidence.

4-2- Méthodes et logiciels

Nous avons utilisé Excel pour calculer la statistique du Chi-deux de Pearson, l'indice de Moran et la statistique de Tango. La matrice de proximité utilisé est la matrice dichotomique car la seule information disponible est l'origine géographique, elle est définie par :

$$w_{ij} = \begin{cases} 1 & si\ la\ wilaya\ i\ a\ des\ frontieres\ avec\ la\ wilaya\ j \\ 0 & sinon \end{cases}$$

On a supposé que la wilaya i n'a pas de frontières avec elle-même ce qui implique $w_{ii} = 0$. Cette procédure a facilité les calculs car la matrice résultante est symétrique.

Le logiciel SPSS 17 a été utilisé pour identifier les groupes identiques de cancers de la Thyroïde en appliquant la méthode décrite précédemment dite de clusters en deux étapes.

Le logiciel SatScan a été utilisé pour confirmer l'existence et la significativité des clusters. Ce logiciel nous permet d'appliquer la méthode de balayage de Kulldorff (méthode de détection locale,Waller &Gotway (2004)) qui consiste à regrouper les différentes unités spatiales voisines en clusters potentiels.

4-3- Résultats

Les trois méthodes utilisées rejettent l'hypothèse des risques constants et confirme la non existence d'un pattern aléatoire.

La statistique de chi-deux de Pearson ($\chi^2 = 150{,}598$) est strictement supérieure à la valeur de chi-deux tabulée ($\chi^2_{0.95}(2) = 5.598$) donc le nombre de cas de malades observés ne se distribue pas aléatoirement et on rejette l'hypothèse de risque constant.

L'indice de Moran I=0.4913 est positif et supérieur à sa valeur moyenne théorique (E(I)=0.0212), ceci suggère une hétérogénéité spatiale.

Pour la statistique de Tango (T=0.043), l'approximation par la loi de chi-deux donne une valeur de 307,902 qui est strictement supérieure à au quantile d'ordre 95% d'une gamma (6,3, 0.5).

L'utilisation de la méthode de clusters en deux étapes met en évidence cinq groupes homogènes par rapport aux variables taux d'incidence et population qui sont :

Cluster 1	2, 6, 9, 10, 15, 16, 28, 33
Cluster 2	14, 26, 30, 34, 35, 42, 44, 47
Cluster 3	4, 7, 12, 18, 21, 22, 23, 27, 29, 32, 39, 43, 48
Cluster 4	5, 13, 17, 19, 25, 31
Cluster 5	1, 3, 8, 11, 20, 24, 36, 37, 38, 40, 41, 45, 46

L'analyse par le logiciel SatScan a confirmé les wilayas 2,6, 9, 10, 15, 16 comme clusters spatiaux-temporels significatifs au seuil 1% par rapport aux même variables cités précédemment avec en plus les coordonnées géographiques de la wilaya (latitude et longitude) et l'année d'hospitalisation.

5- Conclusion

Nous rappelons que nous avons travaillé sur les seuls malades hospitalisés (527 malades) au CPMC sur la période 2007-2011. Hors sur la même période d'autres hôpitaux ont aussi recruté des malades du cancer de la thyroïde. Donc le taux d'incidence ne pourrait qu'augmenter et d'autres clusters potentiels pourraient émerger si on généralise l'étude au niveau national.
Il ressort de cette étude que certaines willayas (2, 6, 15) confirmées comme clusters spatiaux significatifs sont déjà connues comme régions endémiques du goitre, le principale facteur de risque étant la carence d'iode. Est-ce que c'est toujours le cas aujourd'hui ? Peut-on considérer qu'un goitre conduit nécessairement à un cancer de la thyroïde ? Pourquoi Alger grande métropole est un cluster spatial significatif ? Est-ce que l'utilisation d'une autre matrice de proximité (changement de distance) conduirait aux mêmes résultats ?
C'est pour cela que cette analyse pourrait être considérée comme un test de dépistage de la maladie et devrait être suivie par des études plus ciblées, avec des informations plus détaillées (adresse exacte du malade) pour confirmer ou infirmer les hypothèses qu'elle dégage.

ANNEXE 3 : Communication présentée lors de la journée cancer de la thyroïde organisée par la Société Algérienne d'Endocrinologie et de Métabolisme (Avril 2013).

ETUDE de L'INCIDENCE du CANCER de la THYROIDE par L'Analyse de regroupements (cluster analysis)

- **Le but d'une analyse de regroupements est de créer des groupes d'individus homogènes à partir d'un certain nombre de variables. Plus précisément, nous voulons combiner des sujets en groupes interprétables de telle sorte que :**
- **Les individus d'un même groupe soient semblables.**
- **Les groupes soient différents**

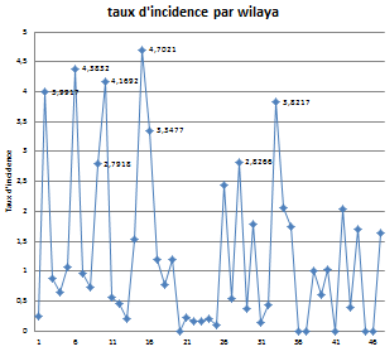
L'étude porte sur 527 malades hospitalisés et opérés du cancer de la thyroïde au service endocrinologie du Centre Pierre et Marie Curie (CPMC) de l'hôpital Mustapha Bacha à Alger sur la période 2007-2011.

Description des cas par sexe et par âge

Age (en année)	Cas	Minimum	Médiane	Maximum	Moyenne et $IC_{95\%}$
Homme	86 (16,3%)	16	39	72	39,58 [36,66; 42,51]
Femme	441 (83,7%)	2	40	80	41,41[«39,97,42,85]

classes d'âge	Femme	Homme
0-4 ans	3 (0,7%)	0
5-9 ans	5 (1,1%)	0
10-14 ans	4 (0,9%)	0
15-19 ans	21(4,8%)	5 (5,8%)
20-24 ans	20 (4,5%)	5 (5,8%)
25-29 ans	60 (13,6%)	15 (17,4%)
30-34 ans	43 (9,8%)	12 (14%)
35-39 ans	53 (12,0%)	7(8,1%)
40-44 ans	45 (10,2%)	10(11,6%)
45-49 ans	41 (9, 3%)	10 (11,6%)
50-54 ans	51 (11,6%)	11 (12,8%)
55-59 ans	42 (9,5%)	3 (3,5%)
60-64 ans	25 (5,7%)	4(4,7%)
65-69 ans	14 (3,2%)	3 (3,5%)
70-74 ans	8 (1,8%)	1 (1,2%)
75-79 ans	5 (1,1%)	0
80-84 ans	1 (0,2%)	0
TOTAL	441	86

- La répartition des cas par groupes d'âge montre qu'il n'y'a pas de différence significative entre les deux sexes



- Les taux d'incidence importants se situent dans les wilayas 2,6,9,10,15,16 28, 33
- Certaines wilayas absentes dans l'échantillon affichent un taux d'incidence égal à zéro

OBJECTIFS de L'ETUDE

- Tester une présence potentielle de clusters de thyroïde par les tests globaux (test de Moran et test Tango) et le test du Chi-deux de Pearson
- Localiser les clusters pouvant exister par la méthode des clusters en deux étapes
- Confirmer la significativité des clusters potentiels par la méthode de balayage de Kulldorff.

Résultats clusterisation

ind Moran	E(I)	var(I)	ecartyp	Z
0,491357118	-0,021276596	0,007977572	0,089317254	5,739470129

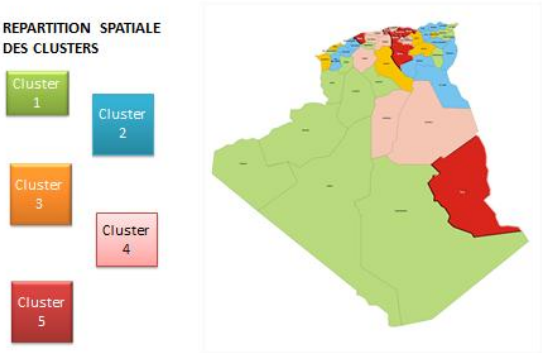
traceWVPR	E(T)	V(T)	NU	Ecartype T	Tango	Tg centréduit
0,105804063	0,00020347	5,00437E-07		0,000707416	0,04282939	58,8256

INTERPRETATION desRESULTATS

- Les trois méthodes utilisées rejettent l'hypothèse de risque constant et confirme la non existence d'un pattern aléatoire.
- L'utilisation de la méthode de clusters en deux étapes met en évidence cinq groupes homogènes par rapport aux variables taux d'incidence et population

Clusters

- Cluster 1: 1,3,8,11,20,24,32,36,37,38,40,41,45,46
- Cluster 2: 4,7,12,18,21,22,23,27,29,39,43,48
- Cluster 3: 5, 13, 17, 19, 25, 31
- Cluster 4: 14, 26, 30, 34, 35, 42, 44, 47
- Cluster 5: 2, 6, 9, 10, 15, 16, 28, 33



- L'analyse par le logiciel SatScan a confirmé les wilayas Chlef, Bougie, Blida, Bouira, Tizi-Ouzou , Alger, M'Sila, Illizi , comme clusters spatiaux-temporels significatifs au seuil 1% par rapport aux même variables cités précédemment avec en plus les coordonnées géographiques de chaque wilaya (latitude et longitude) et l'année d'hospitalisation.

CONCLUSION?

- Nous rappelons que nous avons travaillé sur les seuls malades hospitalisés au CPMC sur la période 2007-2011(certaines wilayas comptabilisent un taux d'incidence nul dans notre échantillon). Or sur la même période d'autres hôpitaux ont aussi recruté des malades du cancer de la thyroïde. Donc le taux d'incidence ne pourrait qu'augmenter et d'autres clusters potentiels pourraient émerger si on généralise l'étude au niveau national
- Il ressort de cette étude que certaines willayas (2, 6, 15) confirmées comme clusters spatiaux significatifs sont déjà connues comme régions endémiques du goitre, le principal facteur de risque étant la carence d'iode. Est-ce que c'est toujours le cas aujourd'hui ? Peut-on considérer qu'un goitre conduit nécessairement à un cancer de la thyroïde ?
- Cette analyse pourrait être considérée comme un test de dépistage de la maladie et devrait être suivie par des études plus ciblées, avec des informations plus détaillées (adresse exacte du malade,...) pour confirmer ou infirmer les hypothèses qu'elle dégage