Geospatial Health 8(2), 2014, pp. 517-526

Cluster detection of diseases in heterogeneous populations: an alternative to scan methods

Rebeca Ramis[1,2,3], Diana Gomez-Barroso[1,2], Gonzalo López-Abente[1,2]

[1]Department of Environmental Epidemiology and Cancer, National Centre for Epidemiology, Carlos III Institute of Health (CIBER en Epidemiología y Salud Pública - CIBERESP), Madrid, Spain; [2]Consortium for Biomedical Research in Epidemiology and Public Health, Madrid, Spain; [3]Faculty of Health and Medicine, Lancaster University, Lancaster, UK

Abstract. Cluster detection has become an important part of the agenda of epidemiologists and public health authorities, the identification of high- and low-risk areas is fundamental in the definition of public health strategies and in the suggestion of potential risks factors. Currently, there are different cluster detection techniques available, the most popular being those using windows to scan the areas within the studied region. However, when these areas are heterogeneous in populations' sizes, scan window methods can lead to inaccurate conclusions. In order to perform cluster detection over heterogeneously populated areas, we developed a method not based on scanning windows but instead on standard mortality ratios (SMR) using irregular spatial aggregation (ISA). Its extension, i.e. irregular spatial aggregation with covariates (ISAC), includes covariates with residuals from Poisson regression. We compared the performance of the method with the flexible shaped spatial scan statistic (FlexScan) using mortality data for stomach and bladder cancer for 8,098 Spanish towns. The results show a collection of clusters for stomach and bladder cancer similar to that detected by ISA and FlexScan. However, in general, clusters detected by FlexScan were bigger and include towns with SMR, which were not statistically significant. For bladder cancer, clusters detected by ISAC differed from those detected by ISA and FlexScan in shape and location. The ISA and ISAC methods could be an alternative to the traditional scan window methods for cluster detection over aggregated data when the areas under study are heterogeneous in terms of population. The simplicity and flexibility of the methods make them more attractive than methods based on more complicated algorithms.

Keywords: cluster detection, irregular shaped clusters, standard mortality ratios, cancer, Spain.

Introduction

The study of the spatial distribution of health events has become an important part of the agenda of epidemiologists and public health authorities. Identification of high- and low-risk areas helps in the definition of public health strategies and in the suggestion of potential risk factors. Furthermore, heterogeneous spatial patterns in incidence or mortality at the local level can point to environmental threats.

In the last decades, the advance of geographical information systems (GIS) and spatial statistics methods have enabled the development of cluster detection techniques, such as local indicator of spatial association (LISA) (Anselin, 1995), spatial scan statistic (SatScan) (Kulldorff, 1997; Kulldorff et al., 2006), upper level set scan statistic (UPS) (Patil and Taillie, 2004), flexible spatial scan statistic (FlexScan) (Tango and Takahashi, 2005) among other special spatial approaches (Assuncao et al., 2006; Yiannakoulias et al., 2007; Oliveira et al., 2011) including proposals from the Bayesian perspective (Knorr-Held and Rasser, 2000; Lawson, 2006). Most of these methods study the occurrence of health events scanning the region with windows of even (circular or ellipsoidal) or irregular shapes and use population under risk as denominator to compute the statistical significance of clusters by the maximum likelihood. Some techniques work with individual data, some with count or aggregated data and some with both (Schmiedel et al., 2012). Among the methods with evenly shaped scanning windows, the most popular is Kulldorff spatial scan statistics (Kulldorff, 1997), e.g. the SaTScan version 9.1.1 software (http://www.satscan.org). Regarding irregularly shaped clusters, Patil and Taillie (2004) developed UPS for detecting arbitrary shaped hotspots and Tango and Takahashi (2005) presented a flexible spatial scan (FlexScan) that allowed the study of irregular shapes (implemented in FlexScan software presented at http://www.niph.go.jp/soshiki/gijutsu/download/flexscan/index.html).

Corresponding author:
Rebeca Ramis
Department of Environmental Epidemiology and Cancer
National Centre for Epidemiology
Carlos III Institute of Health
Monforte de Lemos 5, Madrid 28029, Spain
Tel. +34 91 822 2664; Fax +34 91 387 7815
E-mail: rramis@isciii.es

In the identification of data clusters the unit of aggregation plays a key part for the accuracy of results (Jeffery et al., 2009). When we use homogeneous areas in terms of population sizes such as census tracts, methods based on scan windows like spatial scan statistic provide quite reliable results (Lawson, 2010; Goujon-Bellec et al., 2011; Torabi and Rosychuk, 2011; Schmiedel et al., 2012). Furthermore, according to Huang et al. (2008) who compared SatScan, LISA, FlexScan and UPS working with more realistic spatial patterns for health data, scan window methods were a suitable choice even for heterogeneous populations. Nevertheless, Waller et al. (2006) suggest that those methods could lead to inaccurate conclusions when areas are strongly heterogeneous with reference to population sizes. Interested readers should consult the cited papers to learn more about the performance of the various cluster detection techniques, which are available.

Together, the papers mentioned above provide a good review of existing methodologies and it is not our purpose to launch a completely new one. Our specific target here is the study and monitoring of cancer mortality in Spain at the municipal level and its potential association with environmental factors. This scenario does not fit with the homogeneous areas criterion, because the population distribution is generally heterogeneous across the Spanish territory, which is composed of 8,098 municipalities. As well-known, and documented by the "Instituto Nacional de Estadística" (INE, 2011), the majority of the Spanish population live in a few big cities with large rural areas almost uninhabited. The population size varies along the longitudinal axis: the northern half of the Iberian Peninsula is characterised by a large number of towns and villages with small populations, while there are fewer towns with bigger populations in the south.

We first applied general spatial scan statistics for the cancer data, but the results did not seem reliable as already suggested by Waller et al. (2006). We found unusually large high-risk clusters, sometimes including a whole city, while many small towns appeared to be of no risk at all. In view of these results we decided to develop a different strategy for hotspot identification without windows, allowing irregular shapes and providing probabilities associated with the clusters found. We introduce this work based on two methods: irregular spatial aggregation (ISA) using standard mortality ratios (SMR) and irregular spatial aggregation with covariates (ISAC) that is derived from Poisson regression. We used these two approaches to analyse the spatial patterns of mortality due to stomach and bladder cancer in Spain and compared the results with those provided by FlexScan. We chose these two causes of cancer because they are not evenly distributed as seen in several Spanish mortality atlases (López-Abente et al., 2001, 2007). The counter-intuitive distribution pattern of these two cancers made us analyse their specific, geographical dissemination and possible association with environmental factors (Lopez-Abente et al., 2006; Aragonés et al., 2009).

Materials and methods

Study area and data source

The study included the whole of Spain (Fig. 1) and the data used for statistical analysis were provided by INE.

ISA method

We defined a set of areas spatially contiguous by a shared boundary, satisfying a specific condition regarding a risk indicator, as irregularly shaped cluster. As risk indicator we used SMR defined as the ratio between observed cases and expected cases, which is assumed to be Poisson distributed. We computed SMR's exact confidence interval (CI) (Ng et al., 2008). Initially, we searched for contiguous aggregations of towns with statistically significant values of SMR. We first selected the towns with SMR >1 and kept those with the lower CI limit above a value (k) previously set according to the characteristics of the data and the aims of the study. Among these areas, we identified sets of areas that were spatially contiguous with a minimum of two contiguous areas. For each town, we

Fig. 1. Administrative map of Spain.

computed the empirical distribution of probability of the lower limit of SMR above k under the null hypothesis of spatial independence by Monte Carlo simulations using model 1:

$$O_i \sim Poisson\ (E_i\ \theta_i)$$

where O_i (the number of observed cases) had a Poisson distribution of mean $E_i\ \theta_i$ and E_i is the number of expected cases with θ_i being the relative risk that is equal to 1 under the null hypothesis. We computed E_i by the indirect method of standardisation (Rothman and Greenland, 1998). This method can be implemented for a different risk indicator, i.e. rates, but in that case the probability function would have a different formulation. For this example we set the parameter k equal to 0.97 in order to include smaller towns with wider CIs and ran 100,000 Monte Carlo simulations.

ISA with covariates (ISAC)

In a second stage we searched for irregular clusters after controlling for the effect of some covariates or risk factors; in other words, we looked for clustering with higher than expected risk not related to the covariates. We fitted Poisson regression models including the covariates (model 2) and kept the Pearson residuals to study their spatial distribution. The spatial distributions of these residuals showed the risk as not associated with the covariates, therefore our interest was directed to the towns with the highest residual values, i.e. towns with a larger variance not explained by these risk factors. We selected the towns with Pearson residuals values above a given percentile (P) and identified the irregular clusters as in the previous case using model 2 as follows:

$$O_i \sim Poisson\ (E_i\ \lambda_i);\ \ log(\lambda_i) = \rho + \Sigma_j\ \beta_j\ Cov_j \ \Rightarrow$$
$$\hat{\beta}_j \text{ Estimated covariates effects}$$

Again, we used Monte Carlo simulation to build the empirical distribution needed to compute the probability. We simulated the observed cases under the null hypothesis of spatial independence plus the covariate effects using a third model (model 3).

Model 3. Simulation model:

$$O.sim_i \sim Poisson\ (E_i\ \hat{\lambda}_i);\ \ \hat{\lambda}_i = \exp(\rho + \Sigma_j\ \hat{\beta}_j\ Cov_j)$$

where λ was defined as the exponential of a sum of the terms: baseline risk (ρ) and the estimated parameters of the effect of the risk factors estimated by model 2 (beta.hat). As specific examples we used the following socio-demographic covariates: tobacco, illiteracy, unemployed, farmers, people aged >65 years, average number of persons per household (pph) and income. These covariates were chosen for their availability at municipal level and potential explanatory ability *vis-à-vis* certain geographic mortality patterns (see the Data section below for further information about these covariates). The sequence for the Monte Carlo simulations was the following:

(i) model 2 was fitted with the above mentioned covariates to estimate the corresponding parameters ($\hat{\beta}$);
(ii) 100,000 data sets (*O.sim*) from model 3 were simulated using the estimated parameters ($\hat{\beta}$);
(iii) model 2 was fitted for each set of simulated values (*O.sim*1, .., *O.sim*100,000) to identify the towns with the higher residuals. We defined high residuals as those in the percentile 0.95 (P = 0.95); and
(iv) the empirical distribution with the towns with high residuals was built and this empirical distribution used to compute the probability of the irregular clusters shown.

We used R software to perform all computations.

Flexible shaped spatial scan statistic (FlexScan)

Based on Kulldorff circular spatial scan, the flexible spatial scan statistic of Tango and Takahashi (http://www.niph.go.jp/soshiki/gijutsu/download/flexscan/index.html) was implemented in the FleXScan software for the detection of irregular shaped clusters. As with the spatial scan statistics, the flexible scan has a circular window but it fitted with a maximum opening covering 20 areas per window. In this method not only the whole window can be considered as a potential cluster, but also connected areas inside the window, which makes it possible to detect clusters of irregular shape. Due to the characteristic of allowing the maximum number of 20 areas per window the number of potential clusters to investigate increased considerably. Again, as in the spatial scan statistics, the alternative hypothesis (the risk inside the potential cluster is higher than the risk outside) was tested with a likelihood ratio test and Monte Carlo replications.

Data

We extracted mortality data from the INE records corresponding to deaths coded as stomach cancer

(ICD-10 code C16, ICD-9 code 151) and bladder cancer (ICD-10 code C67, ICD-9 code 188) covering the period 1997-2006 for all the 8,098 municipalities in Spain. We calculated expected cases using the specific Spanish rates, broken down by age (18 groups, 0-4, 5-9, ..., ≥85 years), sex and two five-year periods (1997-2001, 2002-2006), multiplying these by the person-years for each town, broken down into the same strata. We calculated the person-years for the two periods using the 1999 and 2004 populations.

We also collected information about known or potential risk factors that could affect the spatial distribution of the two causes. We did not have direct information about tobacco smoking at the municipal level, but used the SMR for lung cancer mortality during the study period in every municipality as proxy. Lung cancer mortality has been used as a proxy for tobacco smoking previously (Lopez-Abente et al., 2006). Socio-demographic data by percentile were obtained from the 1991 census: illiteracy (illiteracy), unemployed (unemployed), farmers, people >65s, and pph. We also used information about income levels extracted from Ayuso-Orejana et al. (1993). Before their inclusion in the model all covariates were standardised.

Effect of the population size in the towns included in the aggregations

We also studied the effect of the population size for the ISA method. By definition, SMR and its CI are linked to the population size (Rothman and Greenland, 1998); to evaluate this effect we performed a simulation study. Firstly, we divided the towns into four strata according to population size: (i) towns with <500 inhabitants; (ii) towns with 500-2,000 inhabitants; (iii) towns with 2,000-10,000 inhabitants; and (iv) towns with >10,000 inhabitants. Then we simulated observed cases from a Poisson distribution with mean E_i using the ISA method to detect the simulated clusters. We repeated this Monte Carlo simulation 1,000 times and, finally, we computed the proportion of times that towns of each stratum were included in the simulated clusters.

Results

ISA, ISAC and FlexScan

During the study period 36,754 men and 22,917 women died of stomach cancer, and 34,107 men and 7,175 women of bladder cancer. Among the 8,098 Spanish towns, 4,707 presented at least one death of stomach cancer for men and 3,733 for women; for bladder cancer the situation was 4,222 towns for men and 1,963 towns for women. Fig. 2 shows SMR maps based on the towns with statistically significant SMR for the two forms of cancer for both sexes.

Figs. 3-5 include maps for the detected clusters of towns by the ISA, ISAC and the flexible scan methods. Maps of clusters for stomach cancer showed a similar pattern for men (Fig. 3) and women (Fig. 4). For ISA the largest identified clusters were located in the western coastal area of Galicia in the northwest, and most of the remaining ones were located in the north-western half of the country. The ISAC maps give a similar picture but with less clusters for men and more for women. These clusters were seen inland and extended to the southeast. The flexible scan method showed a similar pattern of clusters with respect to the number of towns included, but these clusters were generally bigger.

The ISA maps of bladder cancer clusters were not the same for the two genders. There were clusters in the south (Andalusia) and the east (the Valencia region) for men (Fig. 5), while no statistically significant clusters showed up for women (map not shown). The ISAC maps were quite different; only the male clusters in the Valencia region remained, while the residual patterns for women showed a cluster of three towns in the middle of Spain. For men, the flexible scan method showed a similar pattern to the ISA map, but again clusters were bigger with a higher number of towns included. For women, the flexible scan approach did not find any statistically significant clusters at all.

Although the flexible scan method generally showed similar results as ISA, the main difference was the inclusion of towns with not statistically significant SMRs. However on some occasions the flexible scan method did not include important towns with high risks and high populations within the detected clusters, i.e. for stomach cancer in men it did not locate a cluster in the neighbouring cities of Badajoz (Obs = 73, Exp = 58.04) and Merida (Obs = 50, Exp = 36.14) (blue ring on the map in Fig. 3). In addition, the flexible scan located clusters in big cities and included small neighbouring towns with excess of risks based on one or two cases. For example, for bladder cancer in men, the flexible scan method detected a cluster of eight towns including a medium-size city (Burgos) with 206 observed cases and 134.54 expected cases with the remaining towns showing just one case each.

Tables 1-3 show the sizes and probabilities (P-values) of the detected clusters. Each table is divided into three sub-tables including the results for ISA, flexible

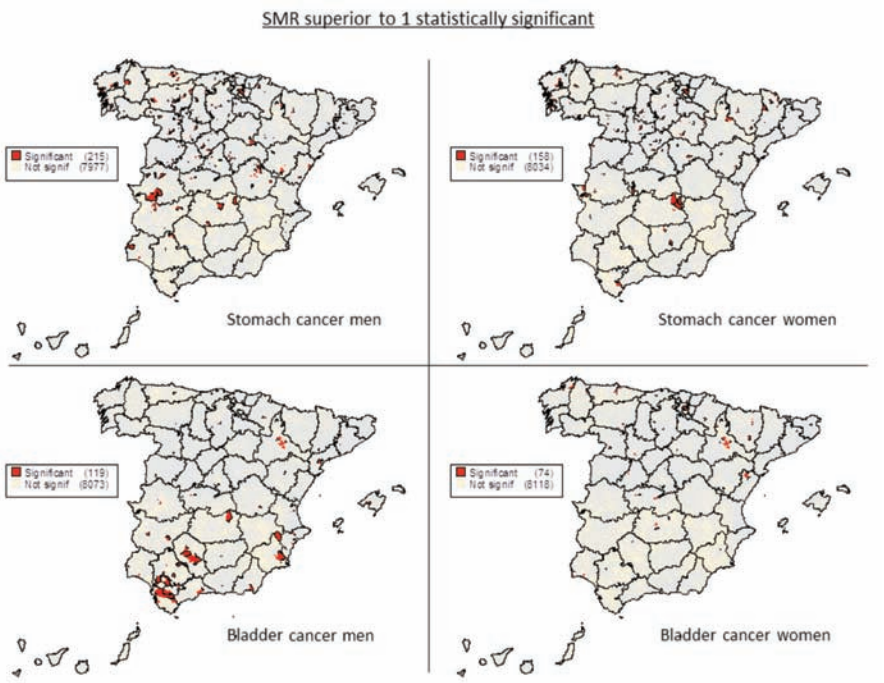

Fig. 2. Maps of statistically significant SMR superior to 1. Stomach cancer in men (top left); stomach cancer in women (top right); bladder cancer in men (bottom left); and bladder cancer in women (bottom right). Red, towns with statistically significant SMR; yellow, towns with SMR not statistically significant. Number of towns in brackets.

scan and ISAC. Table 1 shows the results for stomach cancer in men. The ISA method detected 13 clusters with a statistically significant SMR (1a), while the flexible scan algorithm detected 16 clusters with P-values below 0.01 (1b). Seven clusters of this last group matched nine of those detected by our method, in two cases joining two clusters in one. After controlling for the covariates by the ISAC method we detected six clusters that matched clusters already detected by the two previous methods (1c).

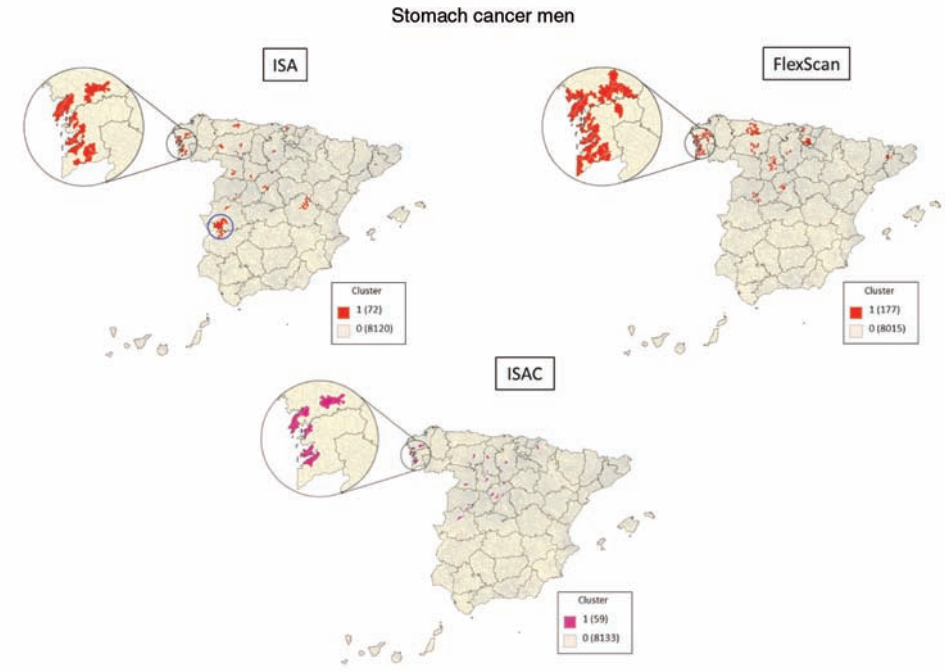

Fig. 3. Cluster maps for stomach cancer in men. Top left, ISA; top right, FlexScan; bottom, ISAC. Clusters in red. Blue ring cities of Badajoz and Merida.

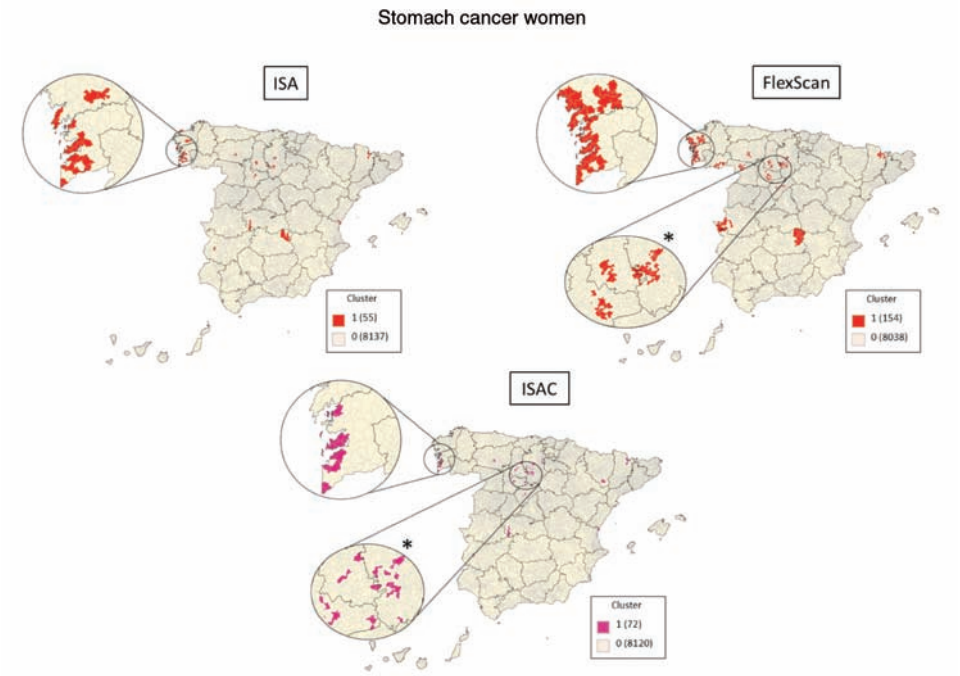

Fig. 4. Cluster map for stomach cancer in women. Top left, ISA; top right, FlexScan; bottom, ISAC. Clusters in red. *Burgos area.

Table 2 shows the results for stomach cancer in women. The ISA method detected seven clusters showing statistically significant SMR (2a), while the flexible scan algorithm detected 13 clusters with P-values <0.1 (2b), six of them matching nine of those detected by our method (in one case joining two clusters in one). After controlling for the covariates by the ISAC method we detected 10 clusters with five of them matching clusters already detected by the two previous methods (2c).

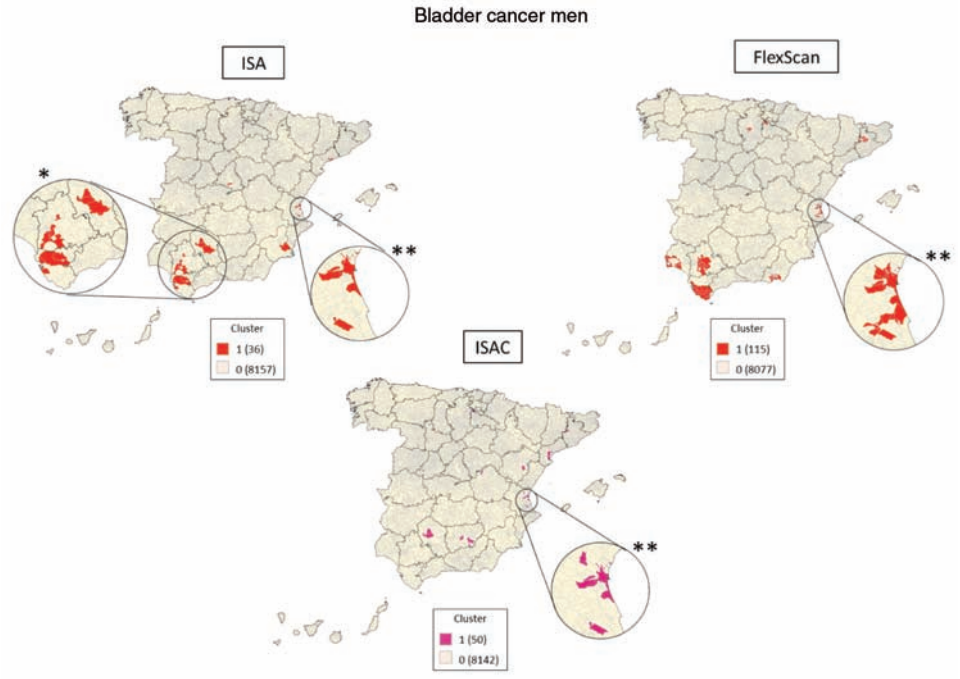

Fig. 5. Cluster map for bladder cancer in men. Top left, ISA; top right, FlexScan; bottom, ISAC. Clusters in red. *Seville and Cadiz. **Valencia.

Table 1. Detected clusters for stomach cancer mortality in men. (1a) shows the clusters detected by ISA method; (1b) clusters by the FlexScan; and (1c) the clusters by ISAC. Columns: C = cluster Id; Size = clusters size; Prob = cluster probability; ISA = cluster id detected by ISA; Flex = cluster id detected by FlexScan.

(1a) ISA cluster			(1b) FlexScan cluster				(1c) ISAC cluster				
C	Size	Prob	C	Size	P-value	ISA	C	Size	Prob	ISA	Flex
1	6	0	1	7	0.001		1	5	0	1	
2	5	0	2	17	0.001	2, 5	2	4	0	2	
3	4	2.00E-05	3	10	0.001	1	3	3	0	5	
4	4	0	4	15	0.001	4	4	3	1.00E-05	10	
5	4	0	5	10	0.004	8	5	3	0		12
6	3	3.00E-05	6	11	0.004	3	6	3	1.00E-05		
7	3	2.00E-05	7	14	0.004	6					
8	3	0	8	9	0.006						
9	3	0	9	7	0.010	8, 9					
10	3	1.00E-05	10	12	0.014						
11	3	2.00E-05	11	11	0.021						
12	3	2.00E-05	12	11	0.021						
13	3	3.00E-05	13	13	0.031						
			14	10	0.031						
			15	10	0.069						
			16	12	0.074						

Table 3 shows the results for bladder cancer in men. The ISA method detected six clusters showing statistically significant SMR (3a), while the flexible scan algorithm detected nine clusters with P-values <0.1 (3b). Only two of these clusters matched three of those detected by our method (as before once joining two clusters in one). After controlling for the covariates by the ISAC method we detected four clusters (3c), one of which matched a cluster already detected by flexible scan.

Finally, for bladder cancer in women ISA method

Table 2. Detected clusters for stomach cancer mortality in women. (2a) shows the clusters detected by ISA method; (2b) clusters by the FlexScan; and (2c) the clusters by ISAC. Columns: C = cluster Id; Size = clusters size; Prob = cluster probability; ISA = cluster id detected by ISA; Flex = cluster id detected by FlexScan.

(2a) ISA cluster			(2b) FlexScan cluster				(2c) ISAC cluster				
C	Size	Prob	C	Size	P-value	ISA	C	Size	Prob	ISA	Flex
1	7	0	1	9	<0.001	5	1	6	0	5	1
2	7	0	2	15	<0.001	1, 2	2	5	0		
3	3	0	3	8	<0.001		3	5	0	1	3
4	3	2.00E-05	4	9	<0.001	7	4	4	0		
5	3	1.00E-05	5	8	<0.001		5	3	0	3	10
6	3	0	5	6	<0.001		6	3	0	4	5
7	3	2.00E-05	7	13	<0.001		7	3	0		
			8	8	0.004	4	8	3	2.00E-05		
			9	9	0.010		9	3	1.00E-05		
			10	7	0.015	3	10	3	0	6	
			11	11	0.020						
			12	8	0.020						
			13	7	0.026						

Table 3. Detected clusters for bladder cancer mortality in men. (3a) shows the clusters detected by ISA method; (3b) clusters by the FlexScan; and (3c) the clusters by ISAC. Columns: C = cluster Id; Size = clusters size; Prob = cluster probability; ISA = cluster id detected by ISA; Flex = cluster id detected by FlexScan.

(3a) ISA cluster			(3b) FlexScan cluster				(3c) ISAC cluster				
C	Size	Prob	C	Size	P-value	ISA	C	Size	Prob	ISA	Flex
1	6	0	1	16	0.001	5, 6	1	5	0	1	2
2	5	0	2	15	0.001	2	2	3	1.00E-05		
3	3	0.00034	3	15	0.001		3	3	3.00E-05		
4	3	5.00E-05	4	14	0.004		4	3	0		
5	3	8.00E-05	5	10	0.007						
6	3	0.00023	6	9	0.010						
			7	15	0.024						
			8	9	0.043						
			9	12	0.057						

did not detected any clusters. However after controlling for the covariates we found one cluster of three towns located in the North with a probability <0.001.

Effect of population size

Among the 8,098 Spanish towns 47% of them fell in the first stratum (<500 inhabitants), 26% in the second (500-2,000 inhabitants), 19% in the third (2,000-10,000 inhabitants) and 8% in the fourth (>10,000 inhabitants). The results from this analysis showed that the towns in the first stratum were present in 33% of the detected clusters (underrepresented); towns in the second stratum were present in 25% of the detected clusters; towns in the third were present in 25% of the detected clusters (overrepresented); and the biggest towns were present in 12% of the detected clusters (overrepresented).

Discussion

The proposed methodology made it possible to identify clusters of towns with excess risk aggregated in irregular shapes like coast lines or rivers and to compute their probability under the null hypothesis of no spatial clustering. Furthermore, this approach permitted inclusion of covariates to control the spatial distribution of the disease, stomach and bladder cancer in this case. It is a simple methodology that combines statistical analysis and GIS to show the geographical location of hotspots in order to compute cluster probability. By definition, this approach shows the true location of the high-risk areas because it uses the SMR and its CI. Incidentally, we computed the CI of ISA and ISAC using Monte Carlo simulations; however, we could also have used for ISA the Wald continuity correction definition (Barker, 2002). Furthermore, the method used could be adapted to different risk indicators. The methodological flexibility is based on the possibility to set the parameters k, P and the number of aggregated areas at various levels, which allows detecting a varying range of clusters. The main difference between ISA and ISAC, on the one hand, and other methodologies on the other, is that they do not use windows and therefore do not depend on scanning the study area.

To assess the performance of ISA and ISAC we compared the results with those identified by the flexible scan method from the FlexScan software developed by Tango and Takahashi (2005). Several reviews of cluster detection methods (Huang et al., 2008; Lawson, 2010; Goujon-Bellec et al., 2011; Torabi and Rosychuk, 2011) suggest that scan statistics could be a good option under specific conditions such as area homogeneity. The flexible scan is able to identify irregular clusters, but it does not allow the inclusion of covariates in the analysis; therefore, we could only compare the results with ISA before controlling for covariates. In general, the statistically significant clusters detected by FlexScan matched locations of our results; however, the flexible scan had the tendency to include more towns in each cluster, mostly small towns with a very low number of cases (SMR>1, i.e. not statistically significant). With FlexScan not statistically significant clusters also matched locations already highlighted by ISA, and again, more small towns were included. In addition, some high risk areas composed by medium and big size cities were not detected as clusters by the flexible scan, while they were detected by ISA. These results suggest that the

flexible scan could include false positive units within the detected clusters, while not including other, false negatives; whereas ISA cannot include false positive units within the detected clusters, nor generate false negative clusters.

Computational times for ISA and ISAC are negligible, even with a large scenario like ours (>8,000 areas). The scanning window process is computationally demanding, especially when the study area contains many areas, but neither ISA nor ISAC scan the region. Even though ISA and ISAC used Monte Carlo simulations to compute the probabilities, computation time is short because the simulations include only those areas within the detected cluster, substantially reducing the computational demand.

Some authors have already mentioned the weaknesses of scan windows methods. For instance, Wakefield et al. (2000) expressed their concern about multiple testing and the choice of the maximum exposed population or distance. A more recent study that compared methods for cluster detection suggestes that elliptic scan (Kulldorff et al., 2006) and flexible scan (Tango and Takahashi, 2005) are particularly good at detecting clusters in large territories; however the statistical power of these methods is low and often fail to detect the last unit of the true cluster (Goujon-Bellec et al., 2011).

The main weakness of scan window methods is the aggregation of areas. Risk estimation used in cluster analysis for aggregated data is based on the population under risk within the area. When we use a window we build an artificial area by aggregating all the areas within the window to estimate the risk. This estimated risk is based on the total number of cases and total population within that artificial area, but it does not account for the individual risk of the original areas. When areas are homogeneous in terms of population, methods using scanning windows perform well (Torabi and Rosychuk, 2011); however, when neighbouring areas are heterogeneous in population, these methods are not fully reliable since the location of the cluster can impact the statistical power of the test (Waller et al., 2006). The most populated area within the window outlines the direction of the estimated risk and this effect increases with higher heterogeneity along the aggregated areas. Extreme cases appear when working with towns and cities: the windows include a big city thousands of times bigger than its neighbours, i.e. a city with more than a million inhabitants by towns with a thousand or less. In these cases the contribution of inhabitants and cases by the small towns would not affect the risk within the window defined by the risk of the big city. If the big city shows an excess of risk, windows including it would show that excess of risk whatever the risk of the small towns. The opposite case is also true; if the big city does not have an excess of risk and the small towns have, the window would not be considered as a potential cluster. For the present study, flexible scan results showed this phenomenon a few times and towns showing no cases were included in statistically significant clusters. The special case of Spanish administrative organization makes it more prone to this phenomenon when we perform cluster detection using towns.

By definition, the SMR is conditioned by population size, specially its CI (Rothman and Greenland, 1998). Since our proposal ISA is based on these values, SMR and CI, we decided to perform a sensitivity analysis to evaluate the influence of the population size of a town on its probability of appearing in a cluster. This analysis shows that towns with low population had a marginally reduced probability of being included in a cluster, while high populated towns had a small increase in their probability. This suggests that population size had an influence in the probability of being included in a cluster but not much. In contrast, we should point out that the method with covariates (ISAC) is not affected by population size because it uses the residuals form a Poisson regression instead of SMR.

Conclusion

The ISA and ISAC method could be a viable alternative to the traditional windows methods for cluster detection over aggregated data when the areas under study are heterogeneous in terms of population. The simplicity and flexibility of the methods make them more attractive to use than methods based on more complicated algorithms.

Acknowledgements

This study was funded by Spain's Health Research Fund (Fondo de Investigación Sanitaria - FIS) Grant PI11/00871.

References

Anselin L, 1995. Local indicators of spatial association-LISA. Geogr Anal 27, 93-115.

Aragonés N, Pérez-Gómez B, Pollán M, Ramis R, Vidal E, Lope V, García-Pérez J, Boldo E, López-Abente G, 2009. The striking geographical pattern of gastric cancer mortality in Spain: environmental hypotheses revisited. BMC Cancer 9, 316.

Assuncao R, Costa M, Tavares A, Ferreira S, 2006. Fast detection of arbitrarily shaped disease clusters. Stat Med 25, 723-742.

Ayuso-Ojerana J, Fernandez-Cuesta J, Plaza-Ibeas J, 1993. Anuario del mercado español. Madrid: Banesto.

Barker L, 2002. A comparison of nine confidence intervals for a Poisson parameter when the expected number of events is ≤ 5. Am Stat 56, 85-89.

Goujon-Bellec S, Demoury C, Guyot-Goubin A, Hémon D, Clavel J, 2011. Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. Int J Health Geogr 10, 53.

Huang L, Pickle LW, Das B, 2008. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Stat Med 27, 5111-5142.

INE, 2011. Censos de Población y Viviendas 2011. Madrid: Instituto Nacional de Estadística. Available at: http://www.ine.es/censos2011_datos/cen11_datos_inicio.htm (accessed on December 2013).

Jeffery C, Ozonoff A, White LF, Nuno M, Pagano M, 2009. Power to detect spatial disturbances under different levels of geographic aggregation. J Am Med Inform Assoc 16, 847-854.

Knorr-Held L, Rasser G, 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56, 13-21.

Kulldorff M, 1997. A spatial scan statistic. Commun Stat Theory Methods 26, 1481-1496.

Kulldorff M, Huang L, Pickle L, Duczmal L, 2006. An elliptic spatial scan statistic. Stat Med 25, 3929-3943.

Lawson AB, 2006. Disease cluster detection: a critique and a Bayesian proposal. Stat Med 25, 897-916.

Lawson AB, 2010. Hotspot detection and clustering: ways and means. Environ Ecol Stat 17, 231-245.

Lopez-Abente G, Aragones N, Ramis R, Hernandez-Barrera V, Perez-Gomez B, Escolar-Pujolar A, Pollan M, 2006. Municipal distribution of bladder cancer mortality in Spain: possible role of mining and industry. BMC Public Health 6, 17.

López-Abente G, Pollan M, Escolar-Pujolar A, Errezola M, Abraira V, 2001. Atlas de mortalidad por cáncer y otras causas en España 1978-1992. Madrid: Instituto de Salud Carlos III.

López-Abente G, Ramis R, Pollán M, Aragonés N, Pérez-Gómez B, Gómez-Barroso D, Carrasco JM, Lope V, García-Pérez J, Boldo E et al., 2007. Atlas municipal de mortalidad por cáncer en España 1989-1998. Madrid: Instituto de Salud Carlos III.

Ng HKT, Filardo G, Zheng G, 2008. Confidence interval estimating procedures for standardized incidence rates. Comput Stat Data Anal 52, 3501-3516.

Oliveira FLP, Duczmal LH, Cancado ALF, Tavares R, 2011. Nonparametric intensity bounds for the delineation of spatial clusters. Int J Health Geogr 10, 1.

Patil GP, Taillie C, 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environ Ecol Stat 11, 183-197.

Rothman KJ, Greenland S, 1998. Modern epidemiology. Lippincott Williams & Wilkins.

Schmiedel S, Blettner M, Schüz J, 2012. Statistical power of disease cluster and clustering tests for rare diseases: a simulation study of point sources. Spat Spatiotemporal Epidemiol 3, 235-242.

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

Torabi M, Rosychuk RJ, 2011. An examination of five spatial disease clustering methodologies for the identification of childhood cancer clusters in Alberta, Canada. Spat Spatiotemporal Epidemiol 2, 321-330.

Wakefield J, Kelsall J, Morris S. 2000. Clustering, cluster detection, and spatial variation in risk. In: Spatial epidemiology: methods and applications. Elliott P, Wakefield J, Best N, Briggs D (eds). Oxford: Oxford University Press.

Waller LA, Hill EG, Rudd RA, 2006. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Stat Med 25, 853-865.

Yiannakoulias N, Rosychuk RJ, Hodgson J, 2007. Adaptations for finding irregularly shaped disease clusters. Int J Health Geogr 6, 28.