Full Length Research Paper

Application of area to point Kriging to breast cancer incidence in Ashanti Region of Ghana

Ebenezer Bonyah[1*], L. Munyakazi[2], N.N.N. Nsowah-Nuamah[3], D. Asong[4] and Bashiru I.I. Saeed[1]

[1]Lecturers at Department of Mathematics and Statistics, Kumasi Polytechnic Institute, Kumasi, Ghana
[2]Associate Professor and Head Department of Mathematics and Statistics, Kumasi Polytechnic Institute
[3]Professor in Statistics and Rector of Kumasi Polytechnic Institute, Kumasi, Ghana.
[4]Komfo Anokye Teaching Hospital and Senior Lecturer at Kwame Nkrumah Universtiy of Sciences and Technology, Kumasi, Ghana

This paper provides a spatial analysis of breast cancer incidence in the Ashanti region area during the period of 2010 to 2011. Breast cancer disease has prevalence in Ghana particularly in the major cities including Ashanti region. Geographical units vary in shape and size and incidence count is non homogeneous in nature. For this reason, assigned area to point kriging approach is adopted as methodology. There is a large range of spatial autocorrelation in ages above 40 years than that of below 40 years in the various administrative units. The surrounding administrative units in the regional capital are less endemic for women whose ages are above 40 years. However, for those whose ages are below 40 years in all the surrounding administrative units are endemic but the capital is not. Most of the endemic districts share boundaries with Kumasi metropolis, the regional capital, where the only Teaching hospital is located. Most of the districts do not have good health facilities where women report for early treatment.

Key words: Area to point kriging, breast cancer, area to area, kriging.

INTRODUCTION

The leading malignancy in Ghana is breast cancer (Archampong, 1977). This is responsible for 15.4% of all malignancies and seems to be on the increase (Archampong, 1977). In 2011 there has been rise in admission for malignant neoplasms at the Komfo Anokye Teaching hospital of which most of the cases were breast cancers. Ghana has seen tremendous public education about breast cancer within the last few years and some of the non-governmental organizations such as Mammocare Ghana, Cancer Society of Ghana and others have been playing pivotal role in dissemination of information about this disease. More than fifty percent of Ghanaian women have reported the issue of breast cancer at the hospitals when the disease may be at its advanced stage (Badoe

and Baako 2000). In terms of average most of these women report eight months or more after observing a change in their breast (Biritwum et al., 2000). These patients are referred to the Komfo Anokye Teaching hospital where they go for treatment at the surgical outpatient's clinic.

Breast cancer disease incidence rate recorded at districts level are areal data which is good for areal data mapping in geostatistics. This has been implemented by several authors including Goovaerts and Jacquez (2005) and Kyriakidis (2004) to predict areal values. This approach is referred to as "area-to-point" (ATP) or "area to area" (ATA) kriging as following Kyriakidis (2004). The unique feature about ATP kriging is that it allows the mapping of variability within geographical unit (polygon) and at the same time ensuring the coherence of the prediction. For instance, disaggregated estimates of count data are non-negative and the sum is equal to the

*Corresponding author. E-mail: ebbonya@yahoo.com.

original aggregated count.

Kerry et al. (2010a) applied ATP and ATA for analyzing the geography of offenses and for identifying significant clusters of crimes on car-related thefts in the Baltic states. Shao el al. (2009) applied ATP to introduce sex for the cancer rates, and observed the difference between age-adjusted rates and age-sex-adjusted rates.

Goovaerts (2006a) used this technique for cancer data analysis. This approach applied areal supports to predict point values by taking into account the spatial support of data as well as the varying population size. ATP and ATA are capable of analyzing cancer count and mortality maps making it possible to incorporate the shape and size of administration units into the smoothing of choropleth maps and the creation of isopleths risk maps, respectively.

This paper presents a geostatistical analysis of breast cancer incidence data that consists of three steps: (1) filtering of noise in the data using Poisson kriging where the shape and size of administrative units is incorporated into the filtering, (2) the mapping of the corresponding risk at a fine scale and (3) geographical clustering of the disease at the administrative units.

METHODOLOGY

Study area

The Ashanti Region is centrally located in the middle belt of Ghana. It lies between longitudes 0.15° W and 2.25° W, and latitudes 5.50° N and 7.46° N. The region shares boundaries with four of the ten political regions, Brong-Ahafo in the north, Eastern Region in the east, Central Region in the south and Western Region in the South west. The region occupies a total land area of 24,389 km^2 representing 10.2% of the total land area of the country.

It is the third largest region after Northern (70,384 km^2) and Brong Ahafo (39,557 km^2) regions. The region has a population density of 194 persons per square kilometer, the third after Greater Accra and Central Regions.

The total population of the region is 4,725,046 made up of 2,288325 males and 2,436,721 females (Ghana Statistical Service, 2010). The average daily temperature is about 27°C. Much of the region is situated between 150 and 300 m above sea level. The region has one Teaching hospital situated at the regional capital Kumasi and serves the entire region and beyond for tertiary cases. The rest of the 26 administrative units mostly do not have district hospitals and where they exist there are not enough qualified personnel to manage these facilities.

Data sources

The Ashanti region has a Disease Control Units (DCU) to which all District Health Directorates (DHD) report confirmed cases of various diseases at the end of year. In addition to this, the Teaching Hospital within the region has a research unit where various database of diseases are kept. Some of the confirmed cases of cancer disease were obtained from this centre and compared to that of the various District Disease Control Units.

Data for the analysis were classified based on ages below and above 40 years within each administrative unit. This was to find out the incidence rates deference between the two groups of women.

Population data obtained from Ghana Statistical Service was used in computing the raw rates of cancer. Raw rates were calculated as the number of cancer cases in each district divided by the estimated Population in 2010. In order to put the risk better, the raw rates were rescaled by multiplying it by a factor of 100,000. This expresses the raw rates as per 100,000 people

Spatial and non-spatial data input

The basic data inputs were topographic map data, geographic location of the study area where breast cancer cases with patients ages been reported. Topographic map of Ashanti region at a scale of 1:25000 was obtained from Accra Survey and Mapping unit (Figure 1). This was georeferenced and digitized in ArcGIS version 10.0 where coordinates per polygon were extracted from the map.

Reported cases of breast cancer and ages of patients with confirmed breast cancer cases obtained from Komfo Anokye Teaching Hospital (KATH) and Disease Control Units (DCU) were entered as attributes of the polygon features (that is, the District) in the software. Application software are ArcGIS version 10.0 developed by ESRIL and SpaceStat 3.6.1 developed by BioMedware USA.

Geostatistical analysis

Area to Area (ATA) Poisson kriging

Given number N of geographical units v_α (administrative units), represent the observed mortality rates Cancer areal data as $v_\alpha = d(v_\alpha)/n(v_\alpha)$, where $d(v_\alpha)$ is the number of Cancer counts and $n(v_\alpha)$ is the size of the population at risk. The Cancer incidence is explained as realization of a random variable $D(v_\alpha)$ that obeys a Poisson distribution with one parameter (expected number of count) simply the product of the population size $n(v_\alpha)$ by the local risk $R(v_\alpha)$.

The risk is computed as a linear combination of rate $z(v_\alpha)$ and the rates observed in (K-1) neighouring entities v_i:

$$\hat{r}(v_\alpha) = \sum_{i=0}^{k} \lambda_i z(v_i) \tag{1}$$

We compute weights λ_i assigned to the K rates by solving the number of system of linear equations known as "Poisson kriging" system:

$$\sum_{j=1}^{K} \lambda_j \left[\overline{C}_R(v_i, v_j) + \delta_{ij} \frac{m^*}{n(v_i)} \right] + \mu(v_\alpha) = \overline{C}_R(v_i, v_\alpha)$$

$$i=1,\ldots, K \tag{2}$$

$$\sum_{j=1}^{K} \lambda_j = 1$$

Where δ_{ij} if i=j and 0 otherwise and m* is the population- weighted mean of the N rates. The "error variance ", m*/n(v$_i$) help to locate
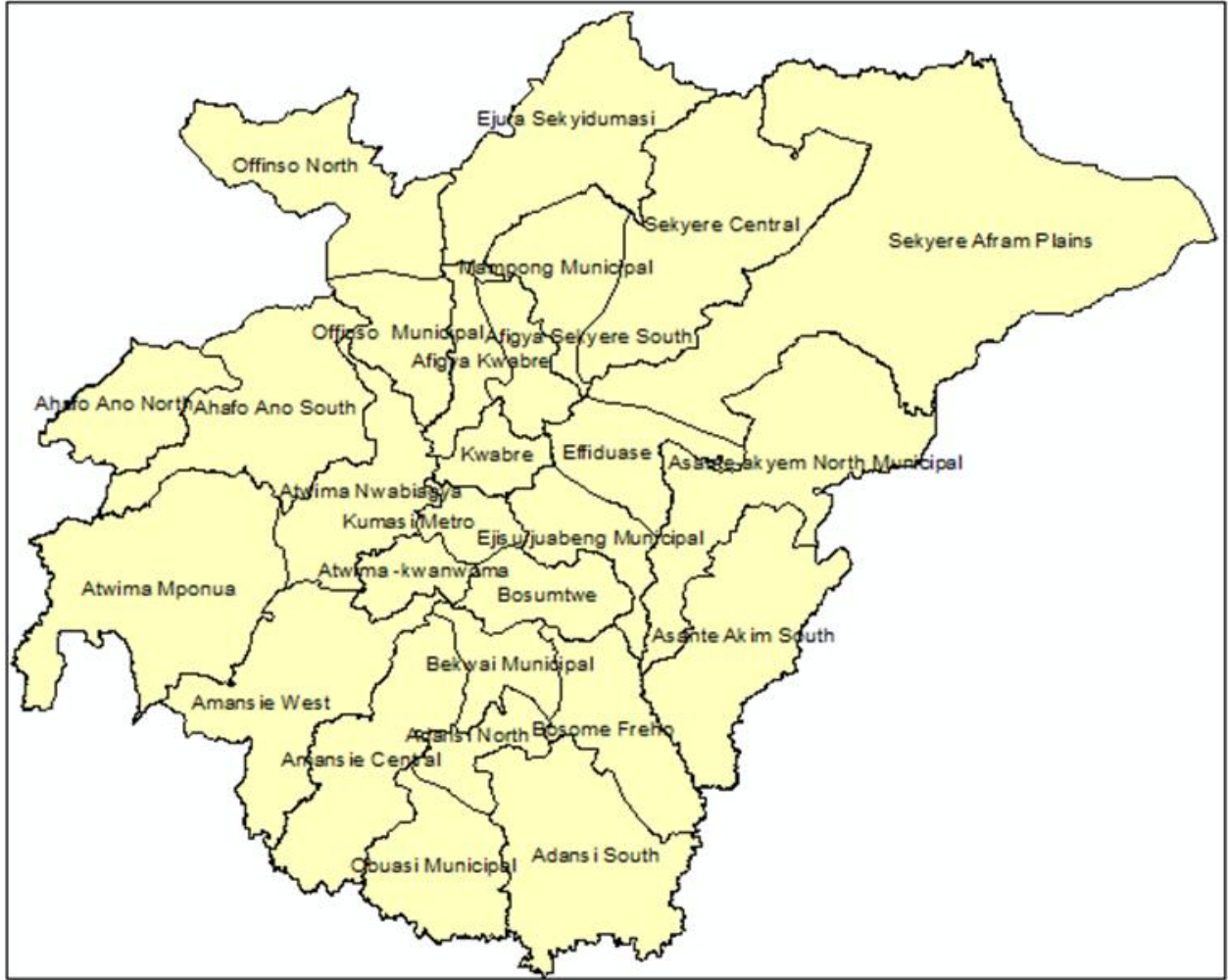
Figure 1. Map of Ashanti region of Ghana indicating District names.

small weightes for less reliable data. The spatial correlation among geographical units through the area-to-area covariance terms $\overline{C}_R(v_i, v_j) = Cov\{Z(v_i), Z(v_j)\}$ and $\overline{C}_R(v_i, v_\alpha)$.

Covariance is then between any two locations discretizing the area v_i and v_j:

$$\overline{C}_R(v_i, v_j) = \frac{1}{\sum\limits_{s=1}^{P_i}\sum\limits_{s'=1}^{P_j} wss'} \sum\limits_{s=1}^{P_i}\sum\limits_{s'=1}^{P_j} wss' C(u_s, u_{s'})$$

(3)

Where P_i and P_j are the number of points used to discretize the two areas v_i and v_j respectively. We compute the weights wss' as the product of two population sizes assigned to each discretizing point u_s and $u_{s'}$:

$$wss' = n(u_s) \times n(u_{s'}) \quad \text{with} \quad \sum\limits_{s=1}^{P_j} n(u_s) = n(v_i) \quad \text{and}$$

$$\sum\limits_{s=1}^{P_j} n(u_{s'}) = n(v_j)$$

(4)

The uncertainty about the cancer mortality risk prevailing within the geographical unit v_α can be modeled using the conditional cumulative distribution function (ccdf) of the risk variable R, Prob$\{R(v\alpha) \leq r|K\}$. Based on assumption of normality of the prediction errors, ccdf is modeled as a Gaussian distribution with the mean and variance corresponding to the Poisson kriging estimate and variance are computed as:

$$\sigma^2(v_\alpha) = \overline{C}_R(v_\alpha, v_\alpha) - \sum\limits_{i=1}^{K} \lambda_i \overline{C}_R(v_\alpha, v_\alpha) \mu(v_\alpha)$$

(5)

Where $\overline{C}_R(v_\alpha, v_\alpha)$ is the within-area covariance that is calculated

according to Equation (3) with $v_i = v_j = v_\alpha$

Area- to- Point (ATP) Point kriging

The prediction support being small as point u_s resulting area-to-point Poisson kriging estimator and kriging variance:

$$\hat{r}_{PK}(u_s) = \sum_{i=1}^{K} \lambda_i(u_s) z(v_i)$$

(6)

$$\sigma^2_{PK}(u_s) = C_R(0) - \sum_{i=1}^{K}(u_s)\overline{C}_R(v_i, u_s) - \mu(u_s)$$

(7)

We compute the kriging weights and the Langrange parameter $\mu(u_s)$ by solving system similar to the ATA kriging system (2), apart from the right-hand –side term where the area-to-area covariance $\overline{C}_R(v_i, u_\alpha)$ are replaced by area-to-point point covariance $\overline{C}_R(v_i, u_s)$ that are simplified as:

$$\overline{C}_R(v_i, u_s) = \frac{1}{\sum_{s'=1}^{P_i} wss'} \sum_{s'=1}^{P_j} wss' C(u_{s'}, u_s)$$

(8)

Where P_j and wss' are defined as in expression (3). ATP reduces visual bias and has coherence property. The population-weighted average of the risk values estimated at the P_α points μ_s discretizing a given entity v_α produces the ATA risk estimate for this entity:

$$\hat{r}_{PK}(v_\alpha) = \frac{1}{n(v_\alpha)}\sum_{s=1}^{P_\alpha} n(u_s)\hat{r}_{PK}(u_s)$$

(9)

Constraint (8) is fulfilled if the same K areal data used for the ATA $\hat{r}_{PK}(v_\alpha)$ are also used for the ATP kriging of the P_α risk values.

Deconvolution of the semivariogram of the risk

In ATA and ATP kriging we need knowledge of point support covariance of the risk **C(h)** or similarly the semivariogram $\gamma(h)$. We cannot obtain this straightly since only the areal data is available. The regularized semivariogram of the risk can be estimated as:

$$\hat{r}_R(h) \frac{1}{2\sum_{\alpha,\beta}^{N(h)} \frac{n(v_\alpha)n(v_\beta)}{n(v_\alpha)+n(v_\beta)}} \sum_{\alpha,\beta}^{N(h)}\left\{ \frac{n(v_\alpha)n(v_\beta)}{n(v_\alpha)+n(v_\beta)}[z(v_\alpha)-z(v_\beta)]^2 - m* \right\}$$

(10)

where N (h) is the number of pairs of administrative units (v_α, v_β) whose population-weighted centroids are separated by the vector h. The varying spatial increment $[z(v_\alpha) - z(v_\beta)]^2$ are weighted by a function of their respective population size $n(v_\alpha)n(v_\beta)/[n(v_\alpha)+n(v_\beta)]$, a term which is inversely proportional to their standard deviations (Monetiez et al., 2006).

Determination of a point-support semivariogram from the semivariogram γ R (**h**) fitted to areal data is known as "deconvolution", a common operation in geostatistics and it typically involves regular areas or blocks (Journel and Huijbregts , 1978). In this paper, we adopted the iterative procedure introduced for rate data measured over irregular geographical units in which one seeks the point-support model that, once regularized, is the closest to the model fitted to areal data; more details and simulation studies are found in Goovaerts (2006b).

Cluster analysis

A common task in disease analysis is to examine administrative units in adjacent geographical locations that are significantly similar or different. Similarity between the breast cancer incidence rate observed within area v_β and those recorded in the $j(v_\alpha)$ neighboring areas v_α can be computed by the local *Moran statistic* (Anselin et al., 2000) as:

$$l(v_\alpha) = \left[\frac{z(v_\alpha)-m}{s} \right] \times \left(\sum_{j=1}^{j(v_\alpha)} \frac{1}{j(v_\alpha)} \times \left[\frac{z(v_j)-m}{s} \right] \right)$$

(11)

where m and s are the mean and standard deviation of the set of N area incident rates respectively. This Local Indicator of Spatial Association (LISA) is simply the product of the kernel rate and the average of the neighboring rates.

The distribution of the local *Moran statistic* under the null hypothesis of complete spatial randomness is usually obtained through a random of shuffling all the count(s) except at v_α each time calculating (10) to get the distribution of simulated LISA values.

The empirical values of (10) are compared with this distribution to compute the P value for the rest. This randomization ignores the population size associated with each areal unit (Goovaerts and Jacquez, 2005).

RESULTS AND DISCUSSIONS

The Figures 2 and 3 show the omnidirectional variogram of breast cancer below and above 40 years risk computed from district-level rates using estimator (10). The experimental variogram was fitted using a Cubic model with a range of 13.4 km for breast cancer below 40 year and 32.52 km for above 40 years (Table 1).

However, breast cancer incidence above 40 years has better range of spatial autocorrelation than incidence below 40 years for each administrative unit. Each model was deconvoluted using the iterative procedure. In two
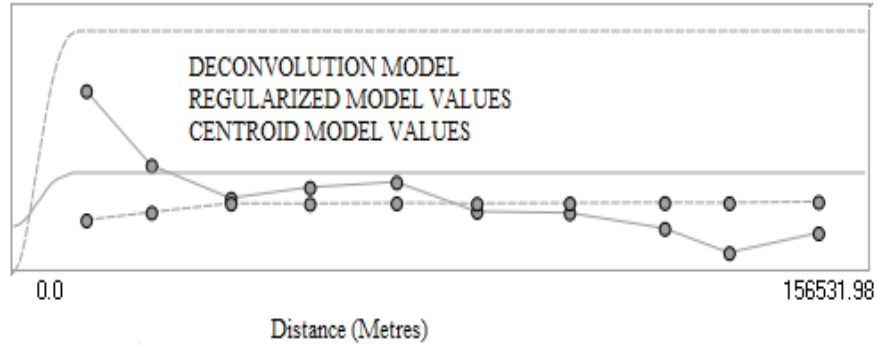
Figure 2. Experimental variogram and model from areal data; and theoretically regularized variogram and deconvoluted model for breast cancer below 40 years.
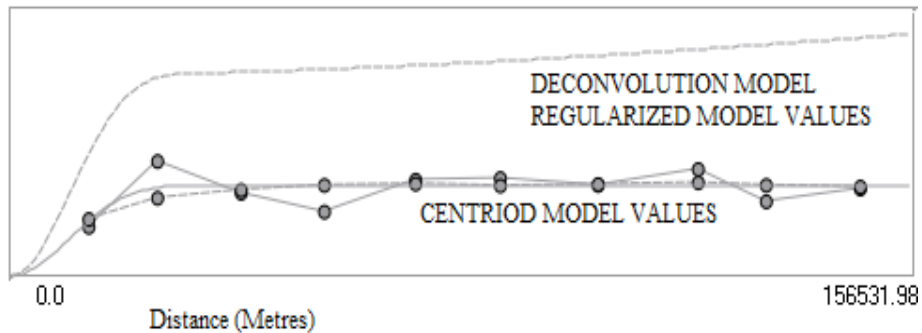

Figure 3. Experimental variogram and model from areal data; and theoretically regularized variogram and deconvoluted model for breast cancer above 40 years.

Table 1. Semivariogram parameters for both below and above 40 years breast cancer disease.

Name	Model type	Sill	Nugget	Range (m)	MSS Error
Breast Cancer Below 40 years	Cubic	0.1669	0	13435.76	2.667
Average Age Above 40 years	Cubic	0.237	0	32522.30	0.018

situations, the procedure ended once a small (that is, <2%) decrease in D statistic occurred four times, after 24 iterations for breast cancer 40 year and 13 for above 40 years.

The deconvoluted variogram model was used to estimate aggregated risk values at the district level in both region (ATA and ATP kriging) (Figure 4). In all cases, the estimation was based on K=32 closest observations which were selected according to the population- weighted district for ATA kriging. All maps are smoother than the map of raw rates since the noise due to small population sizes is filtered.

The breast cancer incidence rate below 40 years (Figure 4 A) indicates that the disease is more endemic around the regional capital Kumasi. The regional capital has the only teaching hospital in the northern sector of the country. There is similarity between the breast cancer

incidence rate and ATA (Figure 4 A and B).

This provides the isopleths map which does not reflect the viability at various administrative units. Almost all the administrative units surrounding the regional capital are more endemic and other areas (Figure 4 A and C) for ages below 40 years. The ATP provides the variability within each administrative unit also shows that all the administrative units surrounding Kumasi metropolis have high risk of cancer disease except Atwima Nwabiagya and Kwabre. These two places have seen good infra-structure development in terms of health facilities. There are other places such as Ofinso North and Asante Akim South which is also endemic and these areas are poorly developed in terms of infrastructure and human resource for them to manage the health facilities even if they exist. In Obuasi municipality there is variability of risk of breast cancer and there is a need to do further research to
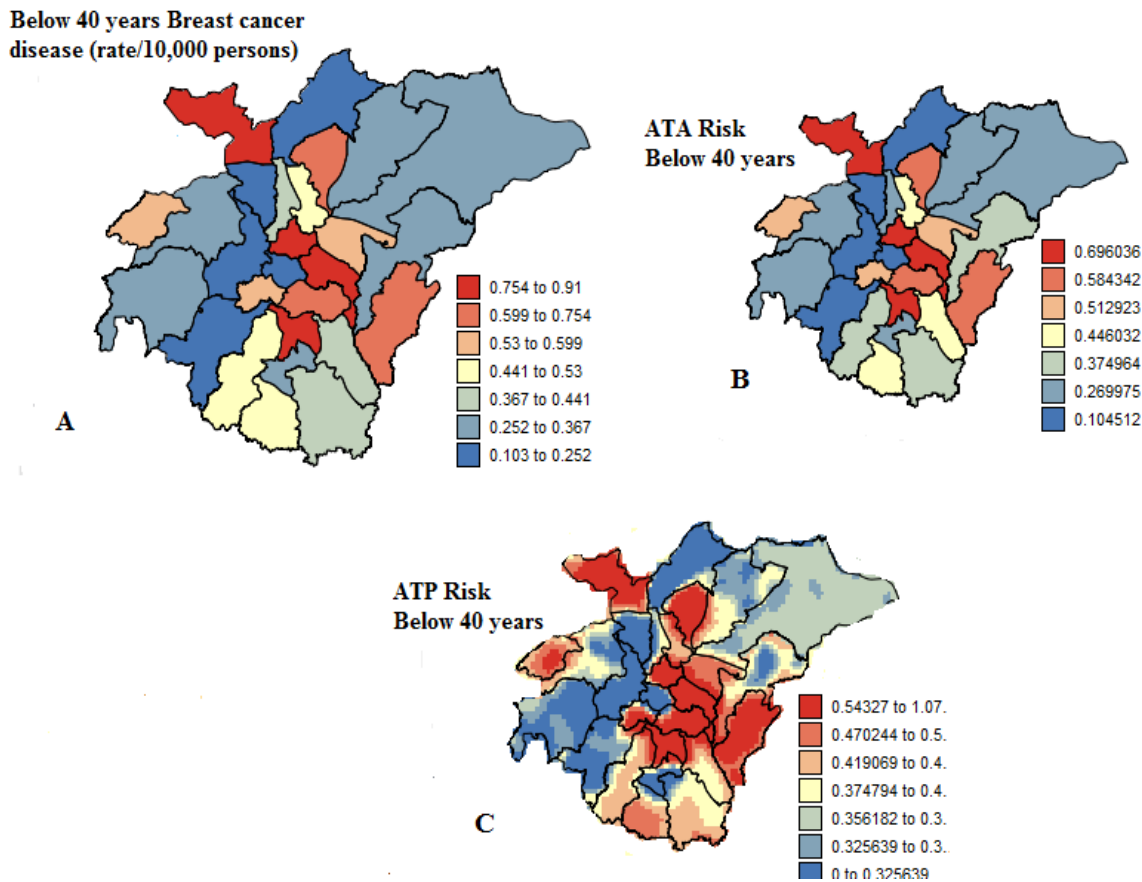
Figure 4. Maps of breast cancer incidence rate estimated by breast cancer rate per 10,000 person, ATA Poisson kriging and ATP kriging on ages below 40 years at various administrative units (A, B and C) respectively.

identify these specific places so that enough education could be conducted for early attendance to health facilities. There are some remote administrative units such as Amansie West, Sekyere Afram Plains and others have very low risk of breast cancer disease within the region.

The breast cancer incidence rate above 40 years for rate per 10,000 persons and ATA risk are similar (Figure 5 A and B). There are risk in most of the areas including Sekyere Afram Plain, Offinso North, Ahafo Ano North, Kwabre, Adansi North and Adansi South. Some of these administrative units do not have even health post for primary medical care. Therefore these women whose ages are above 40 do not report the disease to the health facilities for early treatment.

These women who are above 40 years within the various administrative units are mostly unemployed to get money for treatment. In ATP risk (Figure 5F) which explains the variability within each administrative unit indicates the regional capital Kumasi which is endemic but the surrounding administrative units are less endemic for women above 40 years. Notwithstanding, there are some that are far from the regional capital and endemic including Sekyere Afram Plains, Offinso North, Asante

Akim South and others. These places share boundaries with other regions such Brong Ahafo and Eastern region. The proximity to health facility may account for the low reporting of the disease and some resort to herbal and spiritual healing. Kumasi has a renowned breast cancer NGO known as Breast Care International. They have been organizing free breast cancer screen for women within and around Kumasi metropolis. In most cases special attention is given to women above 40 years who are known to be prone to the disease.

The Local Moran statistics (Figure 6) shows that only Ofinso North and the regional capital Kumasi are significant for women with ages below 40 years. However, it is only significant in Amansie West for women with ages above 40 years in Ashanti region (Figure 6H). This could be one of the reasons why there is a low awareness in most of the administrative units especially those that share boundaries with other region. The regional capital Kumasi which has the teaching hospital and well trained personnel for screening for breast cancer and frequent outreach programme for various suburbs of the cities has improved the awareness of this disease.

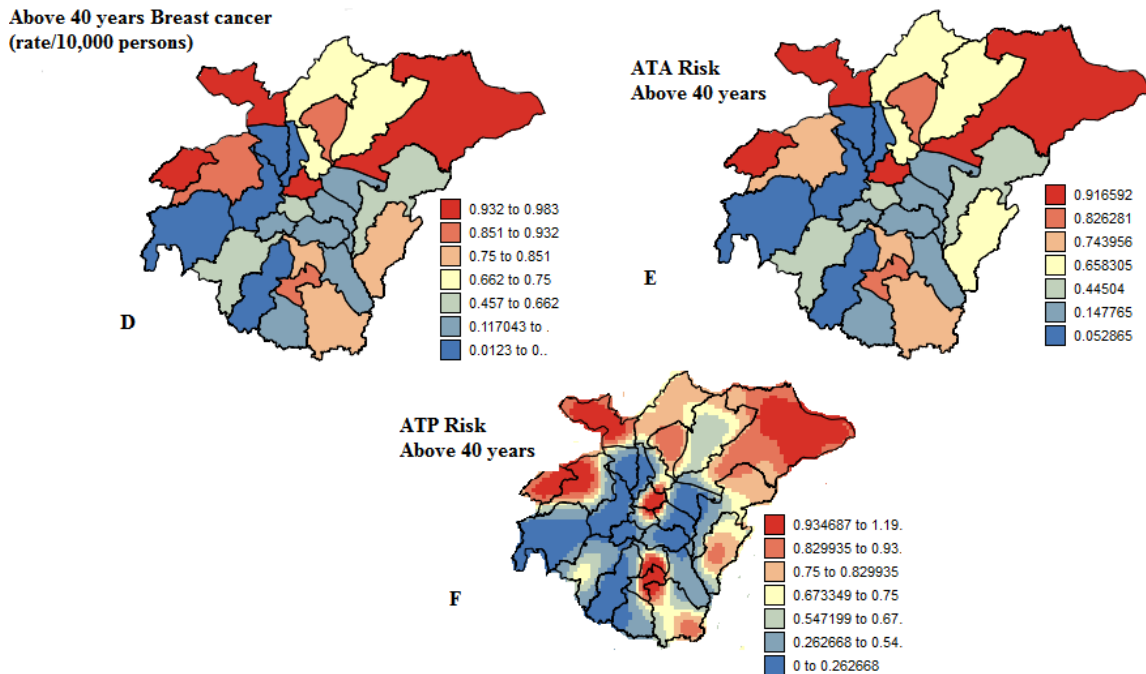However, the majority of the administrative units are not significant (p-value > 0.05). This does not imply that

Figure 5. Maps of breast cancer incidence rate estimated by breast cancer rate per 10,000 person, ATA Poisson kriging and ATP kriging on ages above 40 years at various administrative units (D, E and F) respectively.
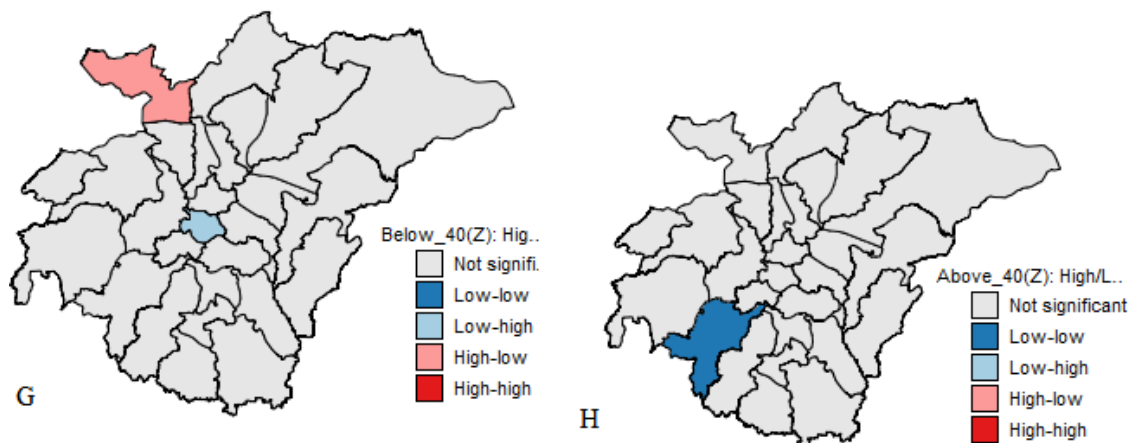

Figure 6. Results of the local cluster analysis conducted by breast cancer incidence rate for ages below and above 40 years (G and H).

these places are breast cancer free. The clustering of the disease in the central part for women with ages below 40 year (Figure 6G) is where we have the regional capital and is a densely populated area.

Conclusions

This study has demonstrated how the breast cancer incidence data can be analysed by considering average ages. ATP kriging is used to create a continuous risk surface that reduces the visual bias associated with large administrative units. This approach of ATP kriging may also give insight into more localized potential "hot spots" that are not evident when areal count on rates are employed. There is large spatial dependency which exist in breast cancer data (Figures 1 and 2). The risk associated with breast cancer (Figures 3 and 4) is centered in the regional capital and administrative units that share boundary with Kumasi the regional capital. In both situations ages below and above 40 years the disease is endemic in administrative units that are far away from

Kumasi. The risk of people developing breast cancer in Ashanti is heterogeneous during the period 2010 to 2011.

REFERENCES

Archampong EQ (1977). Breast Cancer. Ghana Med. J. 16(2): 63.

Asumanu E, Vowotor R, Naaeder SB (2000). Pattern of breast diseases in Ghana. Ghana Med. J. 34:206-209.

Anselin L, Cohen J, Cook D, Gorr W, Tita G (2000). 'Spatial Analyses of Crime. In Criminal Justice 2000: Volume 4. Measurement and Analysis of Crime and Justice, 213–62, edited by D. Duffee. Washington, DC: National Institute of Justice

Badoe EA, Baako BN (2000). The Breast. In Badoe EA, Archampong EQ and da Rocha- Afodu (eds) Principles and Practice of Surgery including pathology in the tropics. Accra: Department of Surgery, University of Ghana Medical School 449-477.

Biritwum RB, Gulaid J, Amaning AO (2000). Pattern of diseases or conditions leading to hospitalisation at the Korle Bu Teaching hospital. Ghana Med. J. 34:197-205.

Cressie NAC (1993). Statistics for Spatial Data. Revised edition. New York: John Wiley &Sons.

Goovaerts P (2006a). Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. Int. J. Health Geographics 5:7.

Goovaerts P (2006b). Geostatistical Analysis of Disease Data: Accounting for Spatial Support and Population Density in the Isopleth Mapping of Cancer Mortality Risk Using Area to Point Poisson Kriging. Int. J. Health Geographics 5: 52.

Goovaerts P, Jacquez GM (2005). Detection of Temporal Changes in the Spatial Distribution of Cancer Rates Using Local Moran's I and Geostatistically Simulated Spatial Neutral Models. J. Geogr. Syst. 7: 137–59.

Journel AG, Huijbregts CJ (1978). Mining geostatistics. Academic Press, London, 600 p

Kyriakidis P (2004). A Geostatistical Framework For Area-To-Point Spatial Interpolation. Geogr. Anal. 36: 259-89.

Kerry R, Goovaerts P, Haining V, Ceccato RP (2010a) Applying geostatistical analysis to crime data: car-related thefts in the Baltic States. Geogr Anal 42:53-75.

Monestiez P, Dubroca L, Bonnin E, Durbec JP, Guinet C (2006). Geostatistical Modelling of Spatial Distribution of *Balenoptera physalus* in the Northwestern Mediterranean Sea from Sparse Count Data and Heterogeneous Observation Efforts, Ecological Modelling, vol. 193, no. 3-4, p. 615-628.

Shao CY, Mueller U, Cross J (2009). Area-to-point Poisson kriging analysis for lung cancer incidence in Perth areas. 18th World IMACS / MODSIM Congress, Cairns, Australia.