18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Area-to-point Poisson kriging analysis for lung cancer incidence in Perth areas

Shao, C.Y. [1], **U. Mueller** [1] **and J. Cross** [1]

[1] *School of Engineering, Edith Cowan University*
100 Joondalup Drive, Joondalup, WA, 6027, Australia
Email: cshao@student.ecu.edu.au

Abstract: This paper provides an analysis of lung cancer incidence in the Perth metropolitan area during the period 1990-2005. The Statistical Local Areas (SLAs) in Perth vary in size and shape, and the incidence case count follows an inhomogeneous Poisson process. For this reason the geostatistical methodology area-to-point (ATP) Poisson kriging is adopted.

For lung cancer in Western Australia (WA) incidence rates are higher for males than for females and it is more common among old people than younger people. In this paper, sex is first taken into consideration for geostatistical cancer analysis. Crude rates, age-adjusted rates and age-sex-adjusted rates are compared. ATP Poisson kriging estimates based on the two adjusted rates are quite similar. However, the ATP Poisson kriging variance based on the age-sex-adjusted rates is much lower than that based on the age-adjusted rates. Slight difference exists between the semivariogram of these two adjusted rates.

The age-sex-adjusted rates will facilitate the correlation between lung cancer and population distribution and composition. Higher cancer rates with low variance are normally found close to the Perth metropolitan area. These areas are not only densely populated but also the average age is relatively higher than in other areas.

Keywords: *Area-to-point (ATP) Poisson kriging, Age-adjusted rates, Age-sex-adjusted rates, Population at risk, semivariogram, nugget effect*

1. INTRODUCTION

The data used in this study is cancer data, and the cancer count data is discrete. To model these spatial data, the standard kriging algorithm like ordinary kriging is not appropriate for the discrete distribution. It is important to take into consideration the binomial or Poisson nature of the count data. The methodology for estimating a spatial Poisson distribution was introduced by Kaiser et al. (1997). They developed the spatial "auto-models" based on the Poisson distribution to be used to incorporate spatial dependencies among the variables. Binomial cokriging was employed to produce a map of childhood cancer risk by taking account of the discrete nature of cancer data in the West Midlands of England by Oliver et al. (1998). Monestiez et al. (2004; 2006) developed Poisson kriging to model spatially heterogeneous observation effort. The approach applied by Monestiez is similar to binomial cokriging proposed by Oliver et al. (1998) except that count data followed a Poisson distribution. Poisson kriging was then generalised by Goovaerts (2005) to analyse cancer data under the assumption that all geographic units are the same size. However, this centroid-based kriging ignored the spatially varying size and shape of each geographic unit. The framework for ATP Poisson kriging was presented by Kyriakids (2004) for interpolating point values from available areal data, and then Goovaerts (2006) 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.

Age-adjusted rates are normally applied in the traditional geostatistical analysis of the cancer data. Since lung cancer affects people not only by age but also by sex, it is very important to remove the bias from sex differences in the population structures. The objective of this paper is to introduce sex for the cancer rates, and see the difference between age-adjusted rates and age-sex-adjusted rates. Also, the adjusted rates, crude rates and ATP Poisson kriging estimates are compared. Software STIS is mainly applied for the data analysis.

2. THEORY

The number of cancer cases follows a Poisson process. The cancer count in the entities (e.g. SLA or suburb) can be viewed as a realisation of a random variable which has a Poisson distribution. This Poisson process has a parameter that is the product of the population size by the local cancer risk (Goovaerts, 2005).

Let $\mathbf{u}_\alpha$ represent the centroid coordinates for each areal supports v_α and $z(\mathbf{u})$ denote the (unknown) point value of the attribute z at location $\mathbf{u}$ within a study domain *D*. In a geostatistical framework, the set of all point support values $\{z(\mathbf{u}), \mathbf{u} \in D\}$ is regarded as a joint particular realization of random variables $\{Z(\mathbf{u}), \mathbf{u} \in D\}$. Area-to-point spatial interpolation is to predict any point value $z(\mathbf{u})$ using K areal data $\{z(v_i), i = 1,...,K\}$:

$$z^*(\mathbf{u}) = \sum_{i=1}^{K} \lambda_i(\mathbf{u}) z(v_i) \tag{1}$$

where the areal supports v_i are disjoint and the prediction locations are arbitrary, that is, they need not be located on a regular grid and they can lie inside or outside v_i (Kyriakids, 2004). The weights $\lambda_i(\mathbf{u})$ are computed to ensure the minimization of prediction mean square error under the condition of the unbiasedness of $z^*(\mathbf{u})$, and they are the solution of the following equations:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(\mathbf{u}) \left[C_{ij} + \delta_{ij} \frac{m^*}{n_i} \right] + \mu(\mathbf{u}) = C(v_i, \mathbf{u}) \quad i = 1,...,K \\ &\sum_{i=1}^{K} \lambda_i(\mathbf{u}) = 1 \end{aligned} \tag{2}$$

where $\mu(\mathbf{u})$ is the Lagrange parameter, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. m^* is the population-weighted mean, C_{ij} is the covariance between area v_i and v_j, and n_i is the size of the population at risk in area v_i. The term $\frac{m^*}{n_i}$ accounts for the variability resulting from the population size. The variance is calculated as:

$$\sigma_{PK}^2(\mathbf{u}) = C(0) - \sum_{i=1}^{K} \lambda_i(\mathbf{u}) C(v_i, \mathbf{u}) - \mu(\mathbf{u}) \tag{3}$$

where $\mathrm{C}(0) = \mathrm{Var}(\mathrm{Z}(\mathbf{u}))$ and $C(v_i, \mathbf{u})$ is the covariance between area v_i and location $\mathbf{u}$ and is inferred from the experimental semivariogram by using $\hat{\gamma}(\mathbf{h}) = \mathrm{C}(0) - \mathrm{C}(\mathbf{h})$ when the variance $\mathrm{Var}(\mathrm{Z}(\mathbf{u}))$ is finite. The experimental semivariogram is calculated as (Goovaerts, 2005):

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2\sum_{\alpha=1}^{\mathrm{N(h)}} \frac{\mathrm{n}(\mathbf{u}_\alpha)\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})}{\mathrm{n}(\mathbf{u}_\alpha)+\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})}} \sum_{\alpha=1}^{\mathrm{N(h)}} \left\{ \frac{\mathrm{n}(\mathbf{u}_\alpha)\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})}{\mathrm{n}(\mathbf{u}_\alpha)+\mathrm{n}(\mathbf{u}_\alpha+\mathbf{h})} [\mathrm{z}(\mathbf{u}_\alpha) - \mathrm{z}(\mathbf{u}_\alpha+\mathbf{h})]^2 - \mathrm{m}^* \right\} \tag{4}$$

3. DATA

The data used in this paper is from the Department of Health of WA. There is a great difference in size and shape in SLAs in the Perth region, and cancer case count does not follow a uniform Poisson process. The population density in different SLAs is highly variable (see Figure 1 for the population density in 2000), and areas near the Perth metropolitan area have higher population density than other areas, so it is more appropriate to apply ATP Poisson kriging than the centroid-based Poisson kriging (Goovaerts, 2005).

The quadrat counting test from the R program Spatstat (Baddeley, 2008), was used to test the null hypothesis of Complete Spatial Randomness (CSR) against the hypothesis that the lung case count follows an inhomogeneous Poisson process. The p-value was lower than 2.2e-16, indicating that the lung case count process is not CSR.

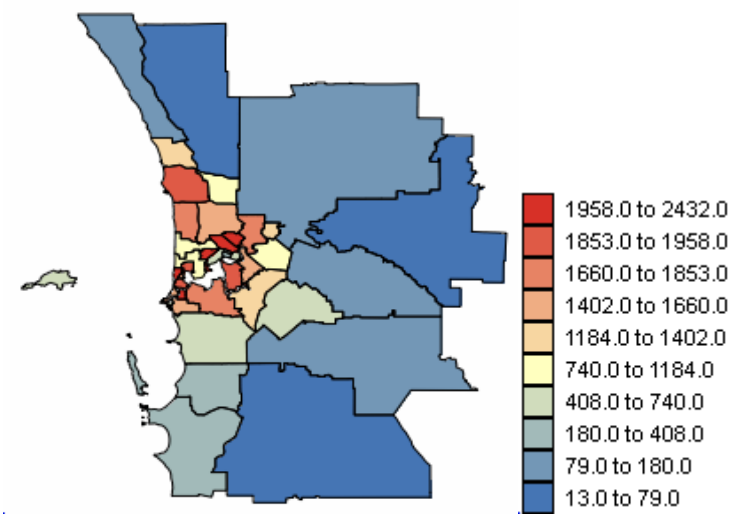

Figure 1. The population density in Perth in 2000. It was calculated by square kilometres using the 2000 population and area size for each SLA.

4. DATA ANALYSIS AND RESULTS

In this study, the areal supports are SLAs and the rates were adjusted using the 2000 WA population composition. For a given SLA v the age-adjusted rates $r_\alpha(v)$ per 100,000 person-years during the 16-year period 1990-2005 were computed as:

$$100{,}000 \times \sum_{i=1}^{9} \left(d_i(v) \Big/ 16 \times N_i(v) \right) \times s_i \tag{5}$$

where s_i is the percentage of age group i in WA 2000, and $d_i(v)$ is the number of incidence cases over the 16-year period for age group i and $N_i(v)$ is the population of age group i in SLA v in 2000. The

population at risk was computed as: $100{,}000\times$ the number of incidence cases over the 16-year period divided by the corresponding adjusted rate. The age-sex-adjusted rates per 100,000 person-years were computed similarly to that of the age-adjusted rates:

$$100{,}000\times\sum_{i=1}^{9}\left(d_{1i}(v)\Big/16\times N_{1i}(v)\times s_{1i}+d_{2i}(v)\Big/16\times N_{2i}(v)\times s_{2i}\right)/2 \tag{6}$$

where the subscripts *1i* and *2i* represent the male and the female age group respectively.

According to Figure 2, crude rates show slight more variability than the two adjusted rates which are quite similar. It can be seen that the population at risk based on these two adjusted rates are nearly the same (see Figure 3).

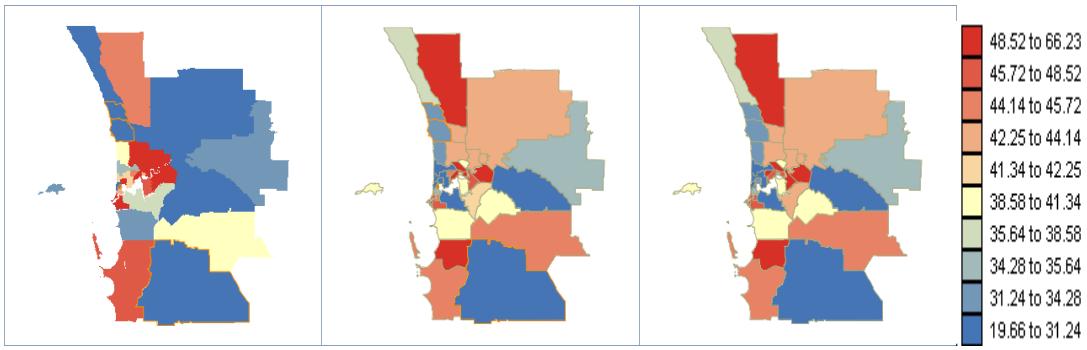

Figure 2. Lung cancer incidence rates per 100,000 person years during the period 1990-2005 in Perth: crude rates (left), age-adjusted rates (middle) and age-sex-adjusted rates (right).

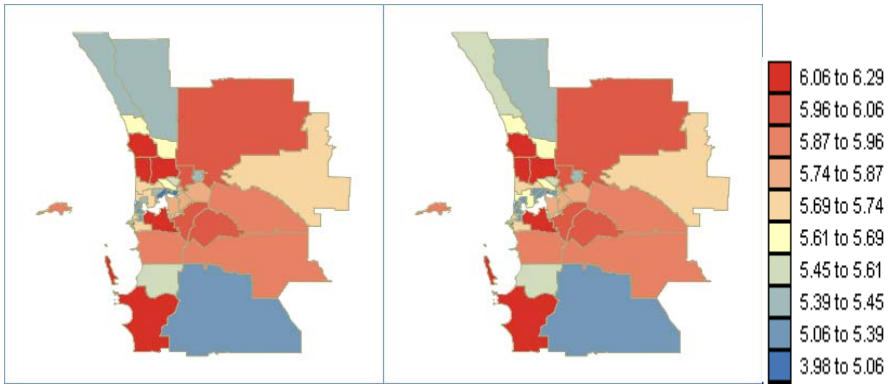

Figure 3. Population at risk (lognormal scale) for lung cancer incidence during 1990-2005 in Perth: based on age-adjusted rates (left) and based on age-sex –adjusted rates (right).

According to Figure 4, the directional semivariogram computed for lung cancer incidence in the four directions (azimuth are measured in degrees clockwise from the NS axis) shows more rapid variation in the east-west (azimuth 90˚) than in the north-south direction (azimuth 0˚). The solid curve represents the model fitted using weighted least square regression. Poisson kriging program will check all possible combinations of one or two basic models from three semivariogram models: spherical, exponential and cubic. The selected model should be the one which minimizes the weighted sum of squares (WSS) of differences between the experimental value $\gamma_e(\mathrm{h}_\ell)$ and model value $\gamma_m(\mathrm{h}_\ell)$:

$$WSS=\sum_{\ell=1}^{L}w(\mathrm{h}_\ell)[\gamma_e(\mathrm{h}_\ell)-\gamma_m(\mathrm{h}_\ell)] \tag{7}$$

where L is the number of classes of distance. There are five weights options of semivariogram modelling:

1. $w(\mathrm{h}_\ell)=1$
2. $w(\mathrm{h}_\ell)=\sqrt{N(\mathrm{h}_\ell)}\Big/\gamma_m(\mathrm{h}_\ell)$

3. $w(h_\ell) = 1/\gamma_m(h_\ell)^2$
4. $w(h_\ell) = N(h_\ell)$
5. $w(h_\ell) = N(h_\ell) \big/ \log|h_\ell|$

where $N(h_\ell)$ is the number of data pairs within the distance h_ℓ. To obtain a good fit, option 2 is used for the fitted semivariogram model of the age-adjusted rates and option 1 is used for that of the age-sex-adjusted rates (see Figure 4).

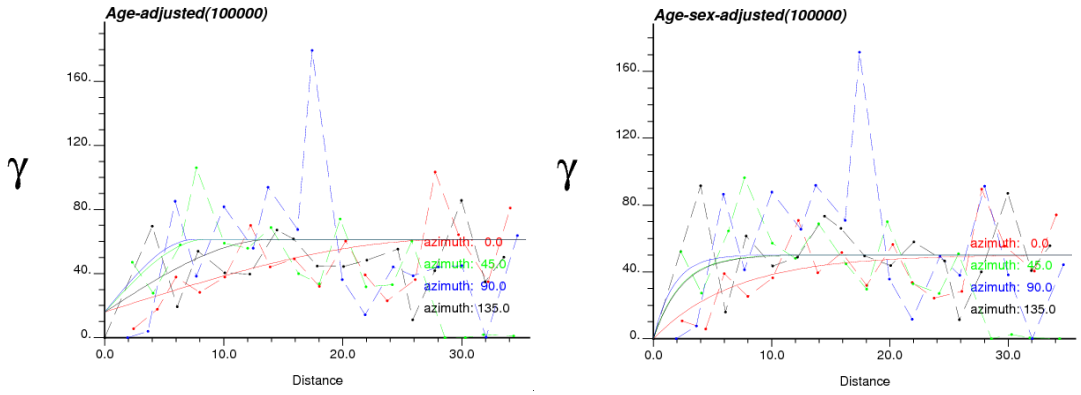

Figure 4. Directional semivariogram of poisson kriging estimator (equation 4) for lung cancer incidence age-adjusted rates (left) and age-sex-adjusted rates (right) during the period 1990-2005. The graphs were created by poisson_kriging.exe.

The variogram of the risk for age-adjusted rates appears same in direction 45° and 90°, and the effective distance for spatial dependence is 8 km. However, there is divergence in the variogram from lag 0 among direction 0°, 45° (90°) and 135°, and almost all of the variation ceases to be autocorrelated after the distance 26 km. Only the correlation in direction 0° is strong and extends to longer distance than other directions: 26 km. The variogram of the risk for age-sex adjusted rates appears same except direction 0°, and the effective distance for spatial dependence is 10 km. The correlation is strong in direction 0° and extends to longer distance than other directions: 26 km. The semivariogram at zero lag is strictly 0. It tells us that the risk from age-sex is likely more continuous than the risk only from age.

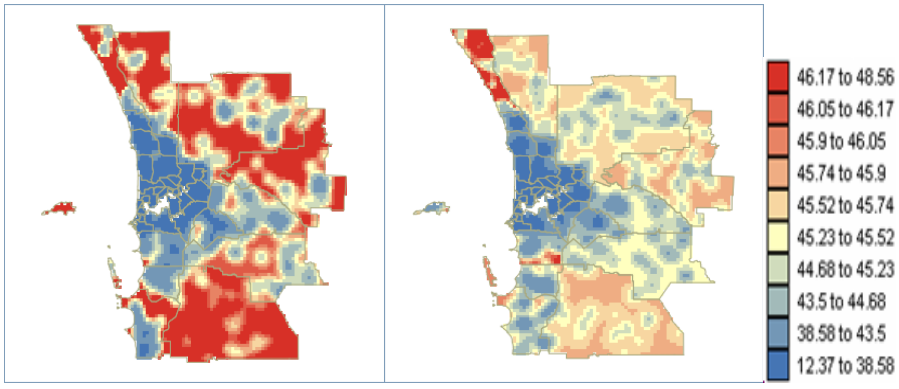

Figure 5. ATP Poisson kriging variance based on age-adjusted rates (left) and ATP Poisson kriging variance based on age-sex-adjusted rates (right). The variance is calculated using equation (3).

ATP Poisson kriging estimates based on the age-adjusted rates indicate more variability than those based on age-sex-adjusted rates (see Figure 5) and higher kriging variance is normally found in sparsely populated areas (see Figure 1). ATP Poisson kriging estimates show much less variability than other rates according to Figure 6 and Table 1. The mean of age-adjusted rates are slightly lower than that of age-sex-adjusted rates, and it is same to kriging estimates based on these two adjusted rates. Although there is difference among

these rates, they indicate a similar trend: higher incidence rates are generally found in areas around Perth metropolitan.

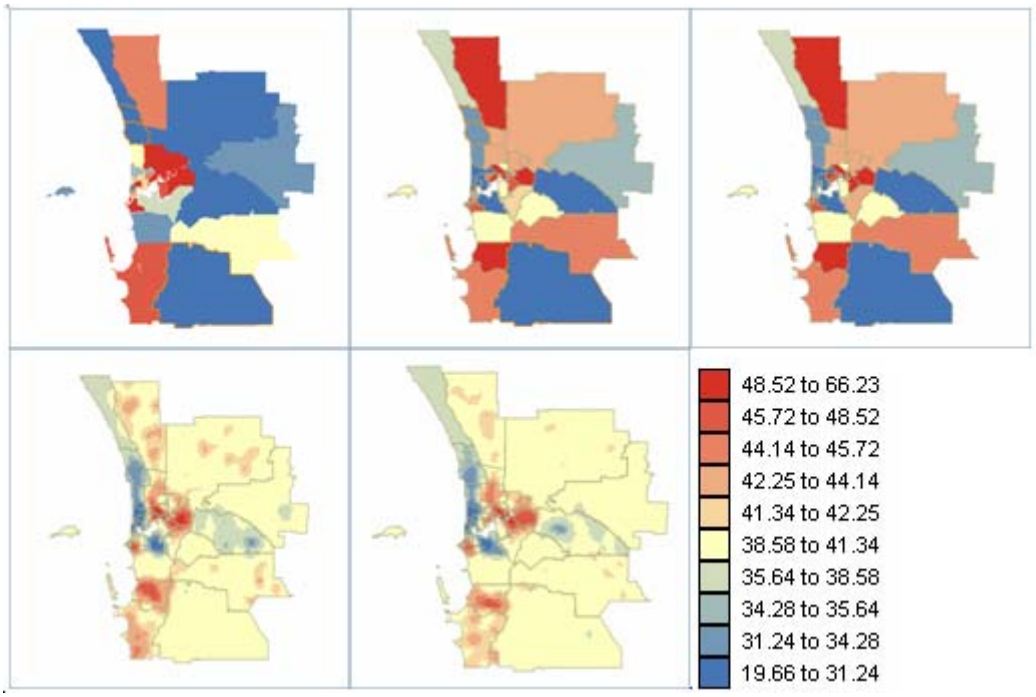

Figure 6. Lung cancer incidence rates per 100,000 person years during the period 1990-2005 in Perth: crude rates (top left), age-adjusted rates (top middle), age-sex-adjusted rates (top right), area-to-point (ATP) Poisson kriging estimates of lung cancer incidence based on age-adjusted rates (bottom left) and based on age-sex-adjusted rates (bottom right).

Table **1.** Summary statistics for lung cancer incidence estimates during the period 1990-2005.

Estimator	Lung cancer incidence					
	Mean	Variance	Kurtosis	Skewness	Minimum	Maximum
Crude rates	42.37	148.56	-1.01	0.06	19.66	66.23
Age-adjusted rates	39.78	82.89	-0.27	-0.07	21.64	60.74
Age-sex-adjusted rates	39.97	83.16	-0.15	0.09	22.94	62.94
ATP (age) estimator	39.73	52.18	1.51	-0.35	21.51	59.19
ATP (age-sex) estimator	39.88	33.39	2.08	-0.14	25.57	56.90

5. CONCLUSIONS

We have shown that how the lung cancer incidence data can be analysed by taking into account sex and age. Slight difference can be found between the two adjusted rates and between the two kriging estimates (see Figure 6 and Table 1), but the variance map (see Figure 5) of the age-sex adjusted rates show much less variability and relatively lower value than the other. In general, close observations are more alike than observations farther apart. The variation (see Figure 4) is spatially dependent and the risk from age-sex (with zero nugget effect) is likely more continuous than the risk only from age.

In short, the risk of people developing lung cancer in Perth is heterogeneous during the period 1990-2005. People living around Perth metropolitan had a higher chance of cancer risk than people living in other areas. According to Figure 4, the semivariogram of the two adjusted rates indicated a similar trend: there was more variability in cancer rates along east-west direction than north-south direction and the cancer rates appeared to vary more continuously (smaller semivariogram values) in the north-south direction.

REFERENCES

Baddeley, A. (2008), Analysing spatial point patterns in R.

Goovaerts, P. (2005), Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging, International Journal of Health Geographics.

Goovaerts, P. (2006), 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, International Journal of Health Geographics, 5, 52.

Kaiser, M.S. and N. Cressie (1997), Modeling Poisson variables with positive spatial dependence Statistical & Probability Letters, 35(4), 423-432.

Kyriakids, P. (2004), A geostatistical framework for area-to-point spatial interpolation Geographical Analysis, 36(3).

Monestiez, P., L. Dubroca and E. Bonin (2004), Comparison of model based geostatistical methods in ecology: application to fin whales spatial distribution in northwestern Mediterranean Sea, Geostatistics Banff, 2, 777-786.

Monestiez, P., L. Dubroca and E. Bonin (2006), Geostatistical modelling of spatial distribution of Balaenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts, Ecological Modelling, 193, 615-628.

Oliver, M.A., R. Webster, C. Lajaunie and K.R. Muir (1998), Binomial cokriging for estimating and mapping the risk of childhood cancer, IMA Journal of Mathematics Applied in Medicine and Biology, 15, 279-297.