Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1087–1093.

CONTIGUOUS DISEASES OUTBREAK IN INDONESIA: THE APPLICATIONS OF SPATIAL SCAN STATISTICS METHOD

Yekti Widyaningsih[1], Asep Saefuddin[2]
Department of Statistics, Bogor Institute of Agriculture, Indonesia
E-mail: [1]yekti@ui.ac.id, [2]asaefuddin@gmail.com

ABSTRACT

The ability to detect disease outbreaks early is important in order to minimize morbidity and mortality through timely implementation of disease prevention and control measures. Many nationals, states, and local health departments are launching disease surveillance systems without statistical testing. Spatial scan statistic is a statistical tool to detect location of clusters (outbreak) of interest. This paper shows how scan statistic method detects diseases clusters. The applications are on detections of contiguous diseases (HIV/AIDS, Tuberculosis and Malaria) hotspot in 2001 - 2006. SaTScan software was used for the computation. As the result, for AIDS cases in Indonesia, there was a moving hotspot area of the mortality number from 2002 to 2006. In 2002, AIDS mortality hotspot is around east part, but in 2006, it was around the central part of Indonesia. The spreading of Tuberculosis in Indonesia was around east and central part of the country in 2001, whereas, in 2002 a reduction of the hotspot was in the central part. For malaria cases, the cluster regions of diseases cases were found, that the cluster regions of high malaria cases tend to decrease from year to year. The highest was in the east and central part of the country in 2001 and 2002.

Keywords: Spatial scan statistics, HIV- AIDS, tuberculosis, malaria, hotspot, SatScan

1. INTRODUCTION

Health officials are often asked to evaluate local disease clusters alarms. After the case definition is established, an early question is whether the cluster has occurred by chance or whether the outbreak is so great that it is probably due to some common elevated risk factor of limited geographical and/or temporal extension. Because of these needs, scan statistics and/or space-time scan statistics have become popular methods in disease surveillance for the detection of disease clusters. The standard approach is to look at a single disease, such as leukemia incidence, breast cancer mortality, HIV/AIDS mortality, bird flu, tuberculosis, and dengue fever. The aim of this paper is to identify the highest response areas of some contiguous diseases (HIV/AIDS, Tuberculosis and Malaria) occurred in Indonesia and to test whether those areas are significant statistically.

Kulldroff (2006), created the SaTScan software as a tool for data change detection within space and /or time. The properties of the data are the scan area geographic, probability distribution of the response under the null hypothesis, and the significance of the statistic test is evaluated with Monte Carlo simulation; Kulldorff (2006).

2. METHODS

2.1 Data

In the year 2005 the population of Indonesia was about 240 millions. As an archipelago, Indonesia consists of thousands small islands and five big islands; they are Sumatera, Kalimantan, Java, Sulawesi, and Papua. There are 33 provinces, 440 districts, and about 5900 sub-districts spread on the islands. Two seasons in Indonesia are the dry season and wet season with a transition period in September.

2.1.1 HIV/AIDS

HIV/AIDS disease is continuing to increase in number, and spreading through the regions. Recently, according to the Jakarta Post news paper, the number of AIDS cases was found in Yogyakarta and more number in Papua; Kulldorff (2007). Through the scan statistics calculation, we analyzed the cluster regions of the AIDS mortality cases in year 2002, which had the mortality number 379 of 1016 cases; and in year 2006, which had mortality number 1651 of 6987 cases; Depkes (2006). This application compared the two results, year 2002 and 2006 data. Based on the number of HIV/AIDS cases, assumed that the mortality number as Bernoulli distribution with possibilities of die or not die. The equation (2) is used as the model of the statistical test.

2.1.2 Ttuberculosis

Although tuberculosis had occurred in human since thousands years ago, this disease is still a big problem in the world and difficult to be combated. Nowadays, many people are still suffering of this disease. Estimated, one person contagious this disease in every second. According to the World Health Organization (WHO) data, more than 670,000 cases occurred each year with deadly number is 175,000 persons; Depkes (2007). Hotspot of TB cases in Indonesia was analyzed using Scan Statistics Method. The data was the number of TB cases of every province in 2001, 2002, 2004, and 2005.

2.1.3 Malaria

Malaria is presently endemic in a broad band around the equator, in areas of the Americas, many parts of Asia, and much more of Africa. The geographic distribution of malaria within large regions is complex, and malaria-free areas are often found close to each other. The global endemic levels of malaria have not been mapped since the 1960s. However, the Welcome Trust, UK, has funded the Malaria Atlas Project to rectify this, providing a more contemporary and robust means with which to assess current and future malaria disease burden; Oemijati (1992).

In Indonesia, malaria has spread to all provincial areas. Malaria often emerges as an outbreak with relatively high morbidity and mortality rates. As an archipelago nation, malaria condition in Indonesia is various for each island. Java and Bali Islands which populated of 70% from total Indonesian population is categorized as a hypo-endemic area. While in other islands that sparsely outer of Java and Bali consist of Sumatera, Kalimantan, Sulawesi, Nusa Tenggara, Maluku and Papua, malaria is found at much higher levels. These areas are categorized from hypo- to hyper endemic.

2.2 Scan Statistics

2.2.1 Purely Spatial Scan Statistics

Purely spatial scan statistics concern in two-dimensional space. Three basic properties of the scan statistic are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window. Depending on the application, different models will be chosen, and depending on the model, the test statistics may be evaluated either through explicit mathematical derivations and approximations or through Monte Carlo sampling. In the latter case, random data sets are generated under the null hypothesis, and the scan statistics is calculated in each case, comparing the values from the real and random data sets to obtain a hypothesis test; Kulldorff (1999).

Bernoulli Model. Under the Bernoulli model, the null hypothesis is $H_o : p = q$, $N(A) \sim$ Binomial $(\mu(A), p)$ for all sets A. And the alternative hypothesis is $H_1 : p > q$, $N(A) \sim$ Binomial $(\mu(A), p)$ for set $A \subset Z$, and $N(A) \sim$ Binomial $(\mu(A), q)$ for set $A \subset Z'$. N(A) is the number of cases in A. Z' is hotspot areas. Probability density functions of an event is

$$f(x) = \begin{cases} p(1-p) & A \subset Z \\ q(1-q) & A \subset Z' \end{cases} \quad (1)$$

If n_Z is point (number of cases in zone Z), n_G is the number of observation, and G is the study area, Likelihood for Bernoulli model is

$$L(Z, p, q) = p^{n_Z}(1-p)^{\mu(Z)-n_Z} q^{n_G-n_Z}(1-q)^{(\mu(G)-\mu(Z))-(n_G-n_Z)} \quad (2)$$

The test statistic λ of the likelihood ratio test can be written as

$$\lambda = \frac{\sup_{Z\in\Omega, p>q} L(Z, p, q)}{\sup_{p=q} L(Z, p, q)} = \frac{L(\hat{Z})}{L_o} \quad (3)$$

L_0 is likelihood under null Hypothesis; Kulldorff (1997).

Poisson Model. Under the Poisson model, points are generated by an inhomogeneous Poisson process. There is exactly one zone $Z \subset G$ such that $N(A) \sim$ Poisson$(p\mu(A\cap Z) + q\mu(A\cap Z^c)) \forall A$. The null hypothesis is $H_o : p = q$, while the alternative hypothesis states that $H_1 : p > q$, $Z \in$ **Z.** Under H_o, $N(A) \sim$ Poisson $(p\mu(A)) \forall A$. Note that one of the parameters, Z disappears under the null hypothesis. The probability of n_G number of points in the study area is

$$\frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}[p\mu(Z)+q(\mu(G)-\mu(Z))]^{n_G}}{n_G!} \quad (4)$$

The likelihood function for the Poisson model is

$$L(Z, p, q) = \frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}}{n_G!} p^{n_Z} q^{(n_G-n_Z)} \prod_{a_i} \mu(a_i) \quad (5)$$

The test statistic λ of the likelihood ratio test can now be written as

$$\lambda = \frac{\sup_{Z\in\Omega,\, p>q} L(Z,p,q)}{\sup_{p=q} L(Z,p,q)} = \frac{L(\hat{Z})}{\frac{e^{-n_G}}{n_G!}\left(\frac{n_G}{\mu(G)}\right)^{n_G}\prod_{a_i}\mu(a_i)}$$

$$= \sup_{Z\in\Omega} \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_G - n_Z}{\mu(G)-\mu(Z)}\right)^{n_G - n_Z}}{\left(\frac{n_G}{\mu(G)}\right)^{n_G}} I\left(\frac{n_Z}{\mu(Z)} > \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}\right) \qquad (6)$$

if there is at least one zone Z such that $\frac{n_Z}{\mu(Z)} > \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}$, and λ=1 otherwise. I() is the indicator function. The ratio λ, is used as the test statistic, and its distribution is find through Monte Carlo repetition as described below; Kulldorff (1997).

In order to find the value of the test statistic, we need a way to calculate the likelihood ratio as it is maximized over the collection of zones in the alternative hypothesis. Once the value of the test statistic has been calculated, it is easy to do the inference. We cannot expect to find the distribution of the test statistic in closed analytical form. Instead we rely on Monte Carlo simulation. Because we know the underlying measure μ, we can obtain replications of the data set generated under the null hypothesis when we condition on the total number of points n_G. With 9999 such replications, the test is significant at the 5 percent level if the value of the test statistic for real data set is among 500 highest values of the test statistic coming from the replications; Kulldorff (1997).

In addition, the most likely cluster for the real data has the significantly test statistic at α = 0.05, that is, its likelihood ratio is on 5% highest among the values of replicated data. The p-value of the Monte Carlo hypothesis is defined as p-value = *r /(1 + sim)*; *r* is the ranking and *sim* is the number of repetitions of the data simulation under the null hypothesis; Kulldorff (2006).

3. APPLICATIONS AND THE RESULTS

The datasets; case and geographical datasets are appended to a master archive using SaTScan. The goal of data analysis is to detect the cluster region (outbreak) of some contiguous diseases. The performance of the purely spatial scan statistic evaluated HIV/AIDS mortality, Tuberculosis (TB) and Malaria data. TB and Malaria cases were assumed as Poisson distribution, whereas HIV/AIDS mortality was assumed as Bernoulli distribution with possibilities of die or not die as mentioned before.

Figure 1 shows the atlases of the spatial scan statistics results for HIV/AIDS mortality cases through SatScan software; ESRI (1996 – 2000). In 2002, the regions of the Most Likely Cluster of aids mortality were Maluku, Central Sulawesi, Papua, North Sulawesi, South Sulawesi, East Nusatenggara, and East Kalimantan. This cluster was statistically significant with relative risk = 1.305, Likelihood ratio (LLR) = 5.175177, and p-value= 0.020. Secondary clusters (Central Java and some parts of Sumatera areas) were not statistically significant.

In 2006, the regions of the Most Likely Cluster of aids mortality were the central part of the country. The provinces were East Nusatenggara, West Nusatenggara, South Sulawesi, Southeast Sulawesi, Bali, West Sulawesi, Central Sulawesi, South Kalimantan, East Java, Gorontalo, Yogyakarta, Central Kalimantan, Maluku, North Sulawesi, East Kalimantan, Central Java, and North Maluku. This cluster was statistically significant with relative risk = 1.578, LLR = 50.587, and p-value= 0.001. The regions of the first secondary cluster were some parts of Sumatera. The

provinces were West Sumatera, Riau, and Riau Islands. This secondary cluster was statistically significant with relative risk = 2.024, LLR = 41.15658, and p-value= 0.001.

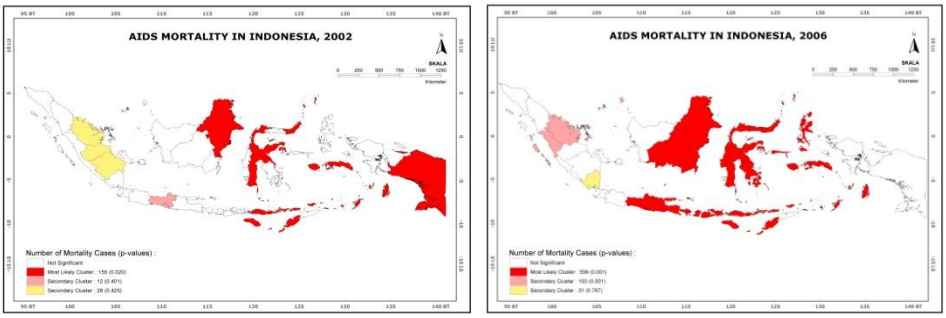

Figure 1. Hotspot of Aids Mortality cases, 2002 and 2006

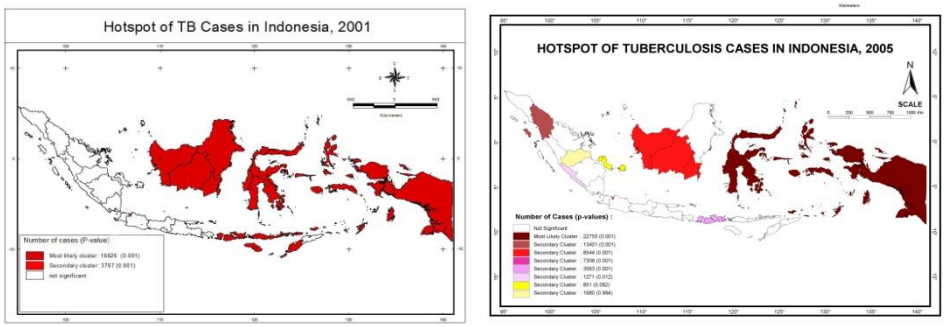

Figure 2. Hotspot of TB Cases, 2001 and 2005

Figure 2 shows the hotspot of Tuberculosis cases, 2001 and 2005. The cases was high in Kalimantan, Sulawesi, Maluku, Nusa Tenggara Islands, and Papua in year 2001. In 2002 these areas were still high, except West and Central Kalimantan (not showed). Followed by Jakarta and West Java, also had significantly high TB cases. Furthermore, in 2004, North Sulawesi, North Maluku, South East Sulawesi, Gorontalo, South Sulawesi, West Sulawesi, and South Kalimantan, the number of TB cases was high. There were 17,288 cases in those areas. Furthermore, in 2005, the highest cases were in Maluku, Sulawesi, and Papua. As a result, hotspots of TB cases were relatively moving just around East and central part of Indonesia and some areas in Sumatera. In addition, hotspot of TB cases also high in North Sumatra in 2004 and 2005, whereas it was not significant in 2001 and 2002. In 2005, the most likely cluster of TB consists provinces in *the east part* of Indonesia, they were Maluku, North Maluku, the whole Sulawesi, and Papua. This cluster was statistically significant as a hotspot with the number of cases was 22755. The secondary clusters were North Sumatera with 13401 cases, followed by Central Kalimantan, South Kalimantan, West Kalimantan with 8544 cases; Jakarta, 7308 cases; West Nusa Tenggara, 3563 cases; and Bengkulu, 1271 cases. All these clusters were statistically significant as the hotspots of Tuberculosis case in Indonesia. Whereas, the estimation number of tuberculosis cases in Indonesia were 296,381 cases spread around Indonesia Islands.

In 2001, hotspot areas of malaria were Kalimantan, Sulawesi, West Nusa Tenggara, East Nusa Tenggara, North Maluku, and Papua. In 2002, hotspot areas were still in the same areas,

except Central Kalimantan and West Kalimantan. In 2003 to 2004, hotspot area is held out in East Nusa Tenggara. In 2005, hotspots areas back to Gorontalo, Southeast Sulawesi, North Sulawesi, and South Sulawesi.

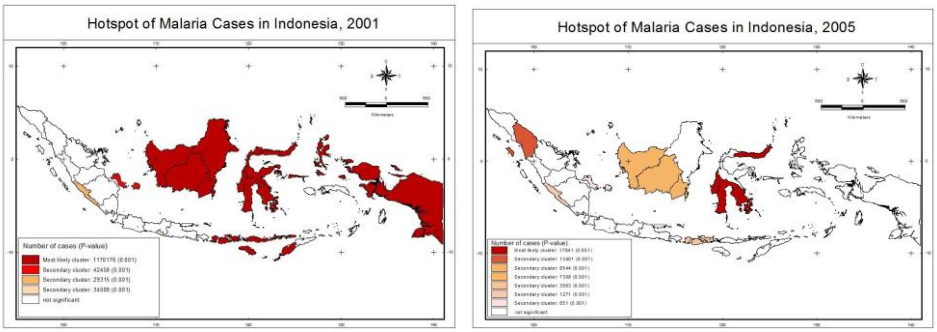

Figure 3. Hotspot of Malaria Cases, 2001 and 2005

There are different results about hotspot areas between the health department report and scan statistics results. The results were based on the reported data. It is believed that many cases are not reported to Health Department. It could be seen, as an example, in 2004 and 2005 malaria cases were high because the government received global fund to investigate malaria cases. It was possible that before 2004 the cases also high but they were not reported.

4. COMMENTS AND CONCLUSIONS

According to the result of calculation for HIV/AIDS mortality in Indonesia, there was a moving hotspot area of the mortality number from year 2002 to 2006. In 2002, AIDS mortality hotspot is around the east part of Indonesia, but in 2006, it was around the central part, include Central Java and Yogyakarta. The news about rising AIDS cases in Yogyakarta in 2007 is relevant with the analysis results. The second highest was in the Central part of Sumatera.

In 2001, the hotspot regions of TB cases were Papua, Sulawesi, and Kalimantan. Furthermore, in 2005, the highest TB cases were in Maluku, Sulawesi, and Papua. As a result, hotspots of TB cases were relatively moving just around east and central part of Indonesia and some areas in Sumatera. In addition, hotspot of TB cases also high in North Sumatra in 2004 and 2005, whereas it was not significant in 2001 and 2002.

Hotspot area of malaria cases was narrower from 2001 to 2005. In 2001, the hotspot areas of malaria cases were Kalimantan, Sulawesi, North Maluku, and Papua, while in 2005 hotspot areas were North Sulawesi, Gorontalo, Southeast Sulawesi, and South Sulawesi. The malaria cases was move from the east to some areas of Sumatera, Kalimantan, and Sulawesi.

As a conclussion, contiguos diseases tended to move from the east part to the central part in period 2001 to 2006 On the basis of study, prevention strategies are recommended that focus on these hotspot areas. The present study analyzed the association between human population and diseases cases. Gathering and including vector population data (including species, population

density, distribution, and infection prevalence rate) and environmental variables in the risk analysis of a disease in these areas provide a more comprehensive view of the disease risk.

ACKNOWLEDGMENTS

Our thanks are due to Prof. Martin Kulldorff for allowing us to use the software and methods. Also thanks are due to The American University in Cairo (AUC), New Cairo, Egypt for the conference facilities and to The University of Indonesia for the support and fund.

REFERENCES

Depkes. Departemen Kesehatan Republik Indonesia. (2006) Ditjen PP & PL. Jakarta.

Depkes. Departemen Kesehatan Republik Indonesia. (2007). *Indonesia Berada di Urutan Tiga Besar Kasus TBC,* http://www.Kapanlagi.com

ESRI. 1996 – 2000. ArcGIS V3.3, Using ArcView GIS.

Kulldorff, M. (1997). A Spatial Scan Statistic. *Communication in Statistics: Theory and Methods*, 26:1481—1496.

Kulldorff, M. (1999). Spatial Scan Statistics: Models, Calculations, and Application. National Cancer Institute, Bethesda, MD

Kulldorff, M. (2006). *SatScan User Guide version 6.1*.

Kulldorff, M., (2007). HIV/AIDS Cases on the Rise in Yogya. *The Jakarta Post*.

Oemijati, S. (1992). *Risk Behavior in Malaria Transmission in Indonesia*. Southeast Asian J Trop Med Public Health 1992; 23: 47-50.