Methodology

Neighborhood size and local geographic variation of health and social determinants
Mohammad Ali*1, Jin-Kyung Park1, Vu Dinh Thiem2, Do Gia Canh2, Michael Emch3 and John D Clemens1

Address: 1International Vaccine Institute, SNU Research Park, San 4–8 Bongcheon-7 dong, Kwanak-gu, Seoul, Korea, 2National Institute of Health and Epidemiology, No. 1 Yersin Street, Hanoi, Vietnam and 3Robert Wood Johnson Foundation Health & Society Scholar, Columbia University, USA

Email: Mohammad Ali* - mali@ivi.int; Jin-Kyung Park - jkpark@ivi.int; Vu Dinh Thiem - vudinhthiem@hn.vnn.vn; Do Gia Canh - vnccdp2@hn.vnn.vn; Michael Emch - me2159@columbia.edu; John D Clemens - jclemens@ivi.int

* Corresponding author

Abstract

Background: Spatial filtering using a geographic information system (GIS) is often used to smooth health and ecological data. Smoothing disease data can help us understand local (neighborhood) geographic variation and ecological risk of diseases. Analyses that use small neighborhood sizes yield individualistic patterns and large sizes reveal the global structure of data where local variation is obscured. Therefore, choosing an optimal neighborhood size is important for understanding ecological associations with diseases. This paper uses Hartley’s test of homogeneity of variance (Fmax) as a methodological solution for selecting optimal neighborhood sizes. The data from a study area in Vietnam are used to test the suitability of this method.

Results: The Hartley’s Fmax test was applied to spatial variables for two enteric diseases and two socioeconomic determinants. Various neighbourhood sizes were tested by using a two step process to implement the Fmax test. First the variance of each neighborhood was compared to the highest neighborhood variance (upper, Fmax1) and then they were compared with the lowest neighborhood variance (lower, Fmax2). A significant value of Fmax1 indicates that the neighborhood does not reveal the global structure of data, and in contrast, a significant value in Fmax2 implies that the neighborhood data are not individualistic. The neighborhoods that are between the lower and the upper limits are the optimal neighborhood sizes.

Conclusion: The results of tests provide different neighborhood sizes for different variables suggesting that optimal neighborhood size is data dependent. In ecology, it is well known that observation scales may influence ecological inference. Therefore, selecting optimal neighborhood size is essential for understanding disease ecologies. The optimal neighbourhood selection method that is tested in this paper can be useful in health and ecological studies.

Introduction

Spatial filtering can be used to create smoothed maps of health and ecological patterns [1-4]. Since population distributions are highly heterogeneous in space, an ordinary
point plot of all cases is not useful. Smoothing data by adjusting for the population at risk is necessary to identify areas with higher disease rates [5]. Smoothing disease data can provide the true relative risk of a disease across a study area [6]. There are other reasons to filter health and ecological data. Field survey data gathering systems usually generate errors. Filtering removes random noise caused by inaccurate records or mislocated cases [1, 7, 8]. There are many intervening factors at the individual level that may influence spatial processes of disease phenomena. For instance, an individual’s biological or socioeconomic status may influence their health status. Neighbors usually have similar risk, particularly for environmentally related diseases, unless the spatial process of the disease is exclusively affected by individual-level characteristics. Also, some risk factors of diseases genuinely operate at the population level [9].

People do not live in isolation; they live in groups (neighborhood) that may influence their life style, health, and health seeking behavior. Thus, a neighborhood level study is sometimes essential to identify important public health problems and to generate hypotheses about their potential causes [9]. Twigg et al. showed that the behavioral practices of an individual are influenced by neighbors [10]. Some variables do not make sense at the individual scale and should be modeled as ecological variables. For instance, a household with a good sanitation system can be exposed to bad sanitation from neighbors. Ecological factors are more meaningful if the data are measured by

Figure 1
The geographic features of the study area in Vietnam. Map showing the geographic characteristics of the study area along with the geographic positing of the study area in Vietnam.
neighborhood. Spatial filtering can be used to model such neighborhood level phenomena.

Ecological variables can be measured at different geographic scales from local to global. In ecology, it is well known that observation scales influence ecological inference [11-13]. Determining the neighborhood size (or the area) over which densities of the phenomena are estimated is important. A large neighborhood makes the data flat over the entire study area whereby important local level variation is obscured that could point to ecological associations. In contrast, a small neighborhood may reveal individualistic patterns [1], and that may not be useful for identifying ecological relationships with health outcomes. Defining an optimal neighborhood size is difficult [14,15]. Bailey and Gatrell [16] suggested exploring different sizes and looking at the variation at those scales to come up with an optimal neighborhood size. However, literature that describes methodologies for selecting the optimal size of neighborhood is scarce. Thus, one often chooses the scale arbitrarily, and the use of an arbitrary scale may yield spurious outcomes. This paper introduces a methodological approach for selecting the optimal neighborhood size that can be used to measure ecological variables and to investigate ecological links with local variation of diseases.

**Methodology**

**The Study area**

Health and socioeconomic data of a study area in Khanh Hoa Province, on the coast of central Vietnam, were used to test the proposed method. The size of the study area is 740 square kilometers consisting of 33 communes in two districts: Nha Trang (151 square kilometres) and Ninh Hoa (589 square kilometres). A dynamic population database is maintained for the study area which is updated on yearly basis. In 2002, the population of the study area was 329,596, of which 54% of the population is from Nha Trang. We created a household-level geographic information system (GIS) database that includes a point for each active household (described below) in 2002. The household settlements are clustered, leaving a large portion non-inhabited land within the administrative boundaries, which led us to define household-based working study area by creating 500 meters radius buffer around each household point and dissolving boundaries between buffers (Figure 1). This resulted in a 394 square kilometres working study area for the entire population and 79 square kilometres for Nha Trang specifically.

**The GIS data**

In 2003, we conducted a global positioning system (GPS) survey using handheld receivers to identify the geographic locations of every household in the study area. A base map of commune boundaries and other geographic features (e.g., rivers, roads, railways, lakes) was first acquired in digital format. The household GPS survey data were projected in the same geographic referencing system (i.e., Transverse Mercator) so that the household points could be accurately integrated with the study area base map. When several households shared a single structure or closely connected structures a single point was plotted. We received a list of 72,152 households from the census 2002. A total of 3,587 households could not be included in the GIS for a variety of reasons (e.g., migration out, living on a military base with no access, and household confirmation was not possible due to the absence of household members). Thus, the GIS was created for the 68,565 households, which were referenced spatially by 32,542 points. Several checks were made for missing households, misallocations, and wrong identification numbers, and the erroneous data were corrected. Finally, the household data were mapped in groups (the smallest geographic entity), and their positions were verified by ground-truthing.

**The attribute data**

We used two social variables, religion and literacy (i.e., years of schooling), and two enteric disease incidence variables (i.e., shigellosis and *Vibrio parahaemolyticus*). The disease incidence data were obtained from a population-based passive surveillance system that was begun in January 1997 [17]. The socioeconomic data were obtained from a 2002 population survey. The survey data shows approximately 85% of the population is secular, 10%
Buddhist, and 5% Christian. Only 11% had not attended school, 37% had received a primary education, and the rest (52%) completed secondary or higher education.

The *Vibrio parahaemolyticus* data were derived from the disease surveillance of the study area from 1997 to 1999. *V. parahaemolyticus*, a gram-negative, halophilic bacterium, inhabits marine and estuarine environments. The microorganism was first identified as a cause of food borne illness in Japan in 1950 [18]. A polymerase chain reaction (PCR) based method to detect the *toxR* sequence specific to *Vibrio parahaemolyticus* was used to identify cases as reported elsewhere [19]. The shigellosis study was carried out for three years (2001–2003) in Nha Trang in collaboration with the Diseases of the Most Impoverished Program [20]. All shigellosis cases were detected through microbiological test of stool samples.

**Data categorization and manipulation**

We categorized the two variables, literacy and religion, to define the social status of the study population. A person having six years of schooling or above was considered to be literate, and the religion was classified by Buddhist and non-Buddhist. Since the data were obtained at the individual level, the data were aggregated by spatially referenced household points. Neighborhood level data were then obtained for each of the spatially referenced points of household (32,542 points for the 329,596 individuals in 68,565 households) by aggregating household level data of surrounding neighbors using circular windows of various sizes. The neighborhood level social variables were estimated by the percentage of people living within neighborhoods, and the disease incidence variables were expressed in rates per 10,000 people. Our aim was to create a local-level neighborhood variable for these phenomena. Therefore, based on the working size of the study area and the spatial distribution of the population we set the minimum size to a 100-meter radius neighborhood and increased the size stepwise by 100 meters until a 2000 meter size was reached. This resulted in 20 different neighborhood sizes from which to select an optimal neighborhood size.

### Table 1: Descriptive statistics and results of variance ratio (Fmax) test for the *Vibrio parahaemolyticus* incidence under various neighborhoods, Khanh Hoa Province, Vietnam, 1997–99. (n = 29,211)

| r    | Population size | Incidence Rate/10000 Population | Upper Fmax Test | Lower Fmax Test |
|------|-----------------|---------------------------------|-----------------|-----------------|
|      | Min  | Max  | Mean | Min  | Max  | Mean | Variance | Fmax1 | DF1 | CV1 | Fmax2 | DF2 | CV2 |
| 100  | 1    | 1779 | 158  | .00  | 1429 | 00  | 4.612 | 646.023 | 49.116 | 20  | 1.571 | 1000 | 1   | 3.842 |
| 200  | 1    | 5143 | 492  | .00  | 370  | 40  | 4.451 | 195.453 | 14.860 | 19  | 1.587 | 3305 | 2   | 2.996 |
| 300  | 1    | 7372 | 914  | .00  | 208  | 30  | 4.538 | 108.599 | 8.257  | 18  | 1.604 | 5949 | 3   | 2.605 |
| 400  | 1    | 9252 | 1416 | .00  | 161  | 30  | 4.533 | 73.509  | 5.589  | 17  | 1.623 | 8788 | 4   | 2.372 |
| 500  | 3    | 12265| 1971 | .00  | 144  | 90  | 4.494 | 52.889  | 4.021  | 16  | 1.644 | 1225 | 5   | 2.214 |
| 600  | 4    | 15784| 2571 | .00  | 227  | 30  | 4.486 | 42.481  | 3.230  | 15  | 1.667 | 1520 | 6   | 2.099 |
| 700  | 4    | 19178| 3236 | .00  | 92   | 59  | 4.441 | 33.666  | 2.560  | 14  | 1.692 | 1918 | 7   | 2.010 |
| 800  | 4    | 21949| 3953 | .00  | 52   | 91  | 4.434 | 28.671  | 2.180  | 13  | 1.720 | 2253 | 8   | 1.939 |
| 900  | 4    | 24982| 4711 | .00  | 38   | 46  | 4.446 | 25.298  | 1.923  | 12  | 1.753 | 2553 | 9   | 1.880 |
| 1000 | 4   | 26772| 5508 | .00  | 35   | 71  | 4.463 | 22.733  | 1.728  | 11  | 1.789 | 2841 | 10  | 1.813 |
| 1100 | 6   | 28821| 6324 | .00  | 36   | 50  | 4.493 | 20.647  | 1.570  | 10  | 1.813 | 3289 | 11  | 1.789 |
| 1200 | 25  | 31691| 7160 | .00  | 36   | 50  | 4.510 | 18.831  | 1.432  | 9   | 1.880 | 3430 | 12  | 1.753 |
| 1300 | 50  | 34877| 8029 | .00  | 46   | 51  | 4.509 | 17.453  | 1.327  | 8   | 1.939 | 3705 | 13  | 1.720 |
| 1400 | 63  | 36311| 8921 | .00  | 37   | 17  | 4.518 | 16.444  | 1.250  | 7   | 2.010 | 3928 | 14  | 1.692 |
| 1500 | 63  | 37334| 9832 | .00  | 31   | 65  | 4.526 | 15.655  | 1.190  | 6   | 2.099 | 4126 | 15  | 1.667 |
| 1600 | 63  | 38471| 10760| .00  | 28   | 33  | 4.542 | 15.116  | 1.149  | 5   | 2.214 | 4273 | 16  | 1.644 |
| 1700 | 63  | 39259| 11693| .00  | 28   | 17  | 4.551 | 14.593  | 1.109  | 4   | 2.372 | 4426 | 17  | 1.623 |
| 1800 | 63  | 40278| 12651| .00  | 27   | 70  | 4.572 | 14.114  | 1.073  | 3   | 2.605 | 4577 | 18  | 1.604 |
| 1900 | 63  | 41457| 13657| .00  | 26   | 32  | 4.599 | 13.639  | 1.037  | 2   | 2.996 | 4736 | 19  | 1.587 |
| 2000 | 63  | 42492| 14703| .00  | 26   | 32  | 4.615 | 13.153  | 1.000  | 1   | 3.842 | 4911 | 20  | 1.571 |

† r = size of neighborhood in meter radius
* DF = degrees of freedom
** CV1 and CV2 = critical values at 95% confidence level for Upper Fmax and Lower Fmax respectively Bold figures in Fmax1 and Fmax2 are the upper and lower limit of optimal neighborhoods, and the bold figure in "r" column is the choice of optimal neighborhood size.
Statistical analysis

Since the data from smaller neighborhoods are individualistic in nature, a high variance value is expected. In contrast, a low variance value is expected for larger neighborhoods. A high variance value means that data are local and low variance means that they are global. To select an optimal size of neighborhood that can ensure that the ecological data are neither local nor global, we used Hartley's test of homogeneity of variance \[21\] that evaluates variation in variances across neighborhoods. The Hartley's test statistic, \(F_{\text{MAX}}\), is calculated by

\[
F_{\text{MAX}} = \frac{S_{\text{MAX}}^2}{S_{\text{MIN}}^2}
\]

where

- \(S_{\text{MAX}}^2\) = maximum value of the variances among groups
- \(S_{\text{MIN}}^2\) = minimum value of the variances among groups

Under the null hypothesis, the test assumes that the variances are equal. The critical value (CV) was calculated under the F-distribution with \((k, n_{\text{MAX}} - 1)\) degrees of freedom at \(\alpha = 0.05\). Here, \(k\) is the number of groups and \(n_{\text{MAX}}\) is the maximum sample size among groups.

The \(F_{\text{MAX}}\) test involved two steps. First the variance of each neighborhood was compared to the highest neighborhood variance (upper, \(F_{\text{MAX}1}\)) and then they were compared with the lowest neighborhood variance (lower, \(F_{\text{MAX}2}\)). A significant value (means the value does not fall within the CV) of \(F_{\text{MAX}1}\) indicates that the neighborhood does not have a global structure of data, and in contrast, a significant value in \(F_{\text{MAX}2}\) implies that the neighborhood data are not individualistic. The neighborhoods that are between the lower and the upper limits are the optimal neighborhood sizes.

Results

There were 131 cases of Vibrio parahaemolyticus in 127 household points in the entire study area for the three years of study (1997–1999), and 308 cases of shigellosis were observed in 295 household points for the year 2001 through 2003 in Nha Trang. Out of the total 329,596 population, 31,924 (9.7%) were Buddhists who were identified in 3,681 household points of the total study area. And, a total of 168,699 (51.2%) literate persons were observed in 30,069 household points.
The data variances for the *Vibro parahaemolyticus* incidence rates under various neighborhood sizes show a declining trend with an increase in neighborhood size (Figure 2). The rapid decline observed at smaller scales virtually disappears with larger neighborhood sizes. The pattern is similar for shigellosis as well as for both socioeconomic variables (figures not shown).

The test results for homogeneity variance of *Vibro parahaemolyticus* incidence rates under various neighborhood sizes are listed in Table 1. The $F_{max}$ test statistic at the level $\alpha = 0.05$ shows a neighborhood size above 900 meters would reveal the global structure of the data, and the $F_{max2}$ statistic shows that any neighborhoods below 200 meters would make the data too individualistic. Thus, the choice of optimal neighborhood lies between 200 and 900 meters. Considering the values of several parameters such as minimum population size, skewness and kurtosis of the incidence rate, we argue that a 500-meter neighborhood is optimal size for modeling the local variation of the disease incidence.

When looking at literacy, the $F_{max1}$ test statistic shows a neighborhood above 600 meters would reveal the global pattern, and the $F_{max2}$ test statistic demonstrates any neighborhoods below 700 meters would make the data individualist (Table 2). In this case, we believe that 700 meters is the optimal size because the difference between $F_{max1}$ and $CV_1$ is smaller than the difference between $F_{max2}$ and $CV_2$ of a 600-meter neighborhood. The summary statistics and test results of religious status under various neighborhood sizes are shown in Table 3. For religion, a 700-meter size neighborhood is also appropriate.

The test results for the choice of optimal neighborhood size for shigellosis obtained from the Nha Trang subpopulation are shown in Table 4. The $F_{max1}$ test statistic reveals that a neighborhood size over 800 meters would produce a global pattern. On the other hand, the $F_{max2}$ test statistic illustrates that a neighborhood below 300 meter would yield an individualistic pattern. Out of the choices between 400 and 800 meters, we suggest a 400 meter neighborhood size based on the criteria mentioned above for *Vibro parahaemolyticus*.

To get an understanding of local geographic variation of the disease and social variables, we created isopleth maps with the spatially smoothed data by using the optimal

---

Table 3: Descriptive statistics and results of variance ratio ($F_{max}$) test for the ethnicity status under various neighborhoods, Khanh Hoa Province, Vietnam, 2002. (n = 32,542)

| r | Population size | Incidence Rate/10000 Population | Upper $F_{max}$ Test | Lower $F_{max}$ Test |
|---|----------------|--------------------------------|----------------------|----------------------|
|   | Min | Max  | Mean  | Min | Max | Mean  | Variance | $F_{max1}$ | DF | $CV_1$ | $F_{max2}$ | DF | $CV_2$ |
|---|-----|------|-------|-----|-----|-------|----------|-----------|-----|--------|-----------|-----|--------|
| 100 | 1   | 1859 | 195   | .00 | 100.00 | 7.283 | 243.304 | 4.026 | 20 | 1.571 | 1.000 | 1 | 3.842 |
| 200 | 1   | 5681 | 611   | .00 | 100.00 | 7.344 | 183.209 | 3.032 | 19 | 1.587 | 1.328 | 2 | 2.996 |
| 300 | 1   | 8362 | 1143  | .00 | 100.00 | 7.383 | 153.024 | 2.532 | 18 | 1.604 | 1.590 | 3 | 2.605 |
| 400 | 2   | 11276| 1785  | .00 | 100.00 | 7.402 | 133.603 | 2.211 | 17 | 1.623 | 1.821 | 4 | 2.372 |
| 500 | 2   | 15425| 2504  | .00 | 100.00 | 7.467 | 121.327 | 2.008 | 16 | 1.644 | 2.005 | 5 | 2.214 |
| 600 | 2   | 20047| 3282  | .00 | 88.57 | 7.534 | 112.520 | 1.862 | 15 | 1.667 | 2.162 | 6 | 2.099 |
| 700 | 2   | 23875| 4146  | .00 | 88.57 | 7.609 | 105.353 | 1.744 | 14 | 1.692 | 2.309 | 7 | 2.010 |
| 800 | 2   | 28601| 5075  | .00 | 82.50 | 7.682 | 99.320 | 1.644 | 13 | 1.720 | 2.450 | 8 | 1.939 |
| 900 | 2   | 33159| 6055  | .00 | 75.00 | 7.719 | 93.729 | 1.551 | 12 | 1.752 | 2.596 | 9 | 1.880 |
| 1000| 4   | 36182| 7081  | .00 | 69.49 | 7.745 | 88.828 | 1.470 | 11 | 1.789 | 2.739 | 10 | 1.831 |
| 1100| 9   | 39576| 8129  | .00 | 63.73 | 7.766 | 84.867 | 1.404 | 10 | 1.831 | 2.867 | 11 | 1.789 |
| 1200| 25  | 43628| 9194  | .00 | 62.40 | 7.776 | 81.459 | 1.348 | 9 | 1.880 | 2.987 | 12 | 1.752 |
| 1300| 25  | 47769| 10297 | .00 | 62.40 | 7.793 | 78.140 | 1.293 | 8 | 1.939 | 3.113 | 13 | 1.720 |
| 1400| 25  | 49747| 11427 | .00 | 61.91 | 7.801 | 74.610 | 1.235 | 7 | 2.010 | 3.261 | 14 | 1.692 |
| 1500| 25  | 51108| 12580 | .00 | 60.84 | 7.820 | 71.456 | 1.183 | 6 | 2.099 | 3.405 | 15 | 1.667 |
| 1600| 66  | 52454| 13750 | .00 | 58.06 | 7.851 | 69.027 | 1.142 | 5 | 2.214 | 3.525 | 16 | 1.644 |
| 1700| 128 | 53725| 14925 | .00 | 55.55 | 7.873 | 66.823 | 1.106 | 4 | 2.372 | 3.641 | 17 | 1.623 |
| 1800| 138 | 55422| 16133 | .00 | 46.79 | 7.891 | 64.531 | 1.068 | 3 | 2.605 | 3.770 | 18 | 1.604 |
| 1900| 138 | 57913| 17398 | .00 | 44.54 | 7.907 | 62.339 | 1.032 | 2 | 2.996 | 3.903 | 19 | 1.587 |
| 2000| 148 | 59428| 18708 | .00 | 39.71 | 7.915 | 60.426 | 1.000 | 1 | 3.842 | 4.026 | 20 | 1.571 |

† r = size of neighborhood in meter radius
* DF = degrees of freedom
**CV1 and CV2 = critical values at 95% confidence level for Upper $F_{max}$ and Lower $F_{max}$ respectively Bold figures in $F_{max1}$ and $F_{max2}$ are the upper and lower limit of optimal neighborhoods, and the bold figure in "r" column is the choice of optimal neighborhood size.
neighborhood sizes. Spatially smoothed data are more appropriate for disease and ecological mapping than the raw data [22]. A widely used geostatistical interpolation method called kriging [23,24] was used to create those maps. The maps ware produced as a quintile distribution for the respective phenomenon. Figure 3 shows the local geographic pattern of the Vibrio parahaemolyticus incidence rate in Khanh Hoa province, Figure 4 shows the geographic pattern of literacy status in Khanh Hoa province, Figure 5 shows the pattern of religion in Khanh Hoa province, and Figure 6 shows the pattern of shigellosis incidence in Nha Trang. All of the maps show clear local geographic variation of the phenomena.

Discussion and conclusion
The Hartley’s $F_{max}$ test of homogeneity provides a solution for determining the optimal neighborhood size for modeling the local variation of health and social determinants. The methodological approach illustrates that the choice of optimal neighborhood is data dependent. Vibrio parahaemolyticus incidence requires a scale from 200 and 900 meters, and we argued that a 500 meter neighborhood is most appropriate based on the values of other parameters. Similarily, out of the options between 300 and 800 meters for shigellosis incidence in Nha Trang, we suggest a 400 meter neighborhood. The choice of neighborhood size for social variables (i.e., literacy and religion) ranged from 600 to 700 meters, and we suggested 700 meters for both. Similarly, out of the options between 300 and 800 meters for shigellosis incidence in Nha Trang, we suggest a 400 meter neighborhood. The maps produced using optimal neighborhood sizes show clear local geographic variation of the respective phenomenon suggesting the suitability of the approach. Since the ecological process may differ from one variable to another [25], different optimal neighborhood sizes are expected. The results of our analyses confirm this notion.

Measuring ecological data at a neighborhood scale to understand the spatial variability requires considerable knowledge of the phenomenon being measured [26]. For example, dissemination of an innovation may diffuse to close neighbors through literate persons. However, the media through which it diffuses is assumed spatially heterogeneous. For instance, a friendly neighborhood may accelerate the innovation, but disputes among neighbors may impede diffusion of the innovation. It would be ideal to assign weight for these social factors while measuring ecological variables, but that requires considerable knowledge about the spatial process of the phenomenon.

| r | Population size | Incidence Rate/10000 Population | Upper $F_{max}$ Test | Lower $F_{max}$ Test |
|---|----------------|-------------------------------|----------------------|---------------------|
|   | Min | Max | Mean | Min | Max | Mean | Variance | $F_{max1}$ | $F_{max2}$ | $F_{max1}$ | $F_{max2}$ |
| 100 | 3  | 5692 | 1015 | .00 | 333.30 | 6.041 | 28.304 | 5.641 | 1.571 | 3.842 |
| 200 | 3  | 17440 | 3223 | .00 | 333.30 | 6.155 | 74.927 | 9.638 | 1.587 | 2.635 |
| 300 | 3  | 25689 | 6026 | .00 | 333.30 | 6.112 | 38.808 | 4.992 | 1.605 | 2.586 |
| 400 | 10 | 34531 | 9388 | .00 | 41.67 | 6.116 | 21.386 | 3.134 | 1.624 | 2.537 |
| 500 | 30 | 47539 | 13112 | .00 | 57.47 | 6.102 | 22.808 | 3.641 | 1.644 | 2.675 |
| 600 | 33 | 61692 | 17105 | .00 | 333.30 | 6.129 | 18.871 | 2.427 | 1.667 | 2.609 |
| 700 | 33 | 73490 | 21480 | .00 | 35.46 | 6.121 | 16.566 | 2.131 | 1.692 | 2.541 |
| 800 | 87 | 88015 | 26126 | .00 | 34.36 | 6.140 | 14.514 | 1.667 | 1.721 | 2.666 |
| 900 | 90 | 102106 | 30938 | .00 | 34.36 | 6.140 | 13.734 | 1.667 | 1.753 | 2.708 |
| 1000 | 87 | 111244 | 35834 | .00 | 31.15 | 6.158 | 12.651 | 1.667 | 1.789 | 2.743 |
| 1100 | 87 | 122187 | 40711 | .00 | 19.67 | 6.137 | 11.774 | 1.515 | 1.813 | 2.778 |
| 1200 | 150 | 134222 | 45547 | .00 | 18.28 | 6.116 | 11.057 | 1.422 | 1.881 | 2.813 |
| 1300 | 425 | 146883 | 50410 | .00 | 17.64 | 6.119 | 10.536 | 1.355 | 1.939 | 2.848 |
| 1400 | 628 | 152901 | 55270 | .00 | 16.40 | 6.135 | 10.079 | 1.297 | 2.010 | 2.883 |
| 1500 | 628 | 156963 | 60900 | 1.21 | 15.92 | 6.129 | 9.588 | 1.233 | 2.099 | 2.918 |
| 1600 | 628 | 161121 | 64829 | 1.77 | 15.92 | 6.123 | 9.104 | 1.171 | 2.155 | 2.953 |
| 1700 | 628 | 165020 | 69457 | 1.21 | 15.92 | 6.129 | 8.751 | 1.124 | 2.373 | 3.085 |
| 1800 | 628 | 170190 | 74078 | 2.73 | 15.92 | 6.135 | 8.411 | 1.082 | 2.606 | 3.216 |
| 1900 | 628 | 177965 | 78890 | 2.68 | 15.92 | 6.153 | 8.099 | 1.042 | 2.996 | 3.359 |
| 2000 | 628 | 182551 | 83821 | 2.98 | 15.92 | 6.171 | 7.774 | 1.000 | 3.842 | 3.537 |

† $r$ = size of neighborhood in meter radius  
* DF = degrees of freedom  
**CV1 and CV2 = critical values at 95% confidence level for Upper $F_{max}$ and Lower $F_{max}$ respectively  
Bold figures in $F_{max1}$ and $F_{max2}$ are the upper and lower limit of optimal neighborhoods, and the bold figure in “r” column is the choice of optimal neighborhood size.
sanitation status, a poorly constructed latrine can be an important source of pollution by spreading fecal matter to nearby areas. Therefore, a distance decay weight can be applied here considering there is an inverse relationship from the source of pollution [27].

Since spatial filtering smooths data, average errors may be inherent in the data [28]. Such ecological bias [29] can be more apparent in a predefined geographic area than within the natural boundary created through spatially smoothed data using optimal neighborhood modeling. Ecological bias may also be present when modeling variables with large neighborhood sizes.

One of the biggest problems in spatial epidemiology and ecological exposure assessment is in identifying geographic patterns [29] through spatial interpolation. Selection of an interpolation method has strong implications on the representation of spatial patterns as well as on the accuracy of interpolated data [30]. Interpolating the data based on spatially smoothed data obtained by an optimal neighborhood size could provide more accuracy in the local variation of the phenomena being measured. The optimal neighborhood may help ecological analysis in two ways: aggregating the data (both dependent and independent variables) using optimal neighbourhood scales and performing the analysis at the ecological level, or by
limiting the dependent variable at the individual level, but attaching ecological covariates (obtained through optimal neighbourhood size) to each individual [31].

A scientifically validated method is required to assist geographic research [32], and to properly use GIS technology in health and ecological studies [33]. In our paper, we have outlined a method to choose optimal neighbourhood sizes for addressing local spatial variation of disease and social determinants. The method can be useful in health and ecological studies.

Acknowledgements
This work was supported by the Diseases of the Most Impoverished Program, funded by the Bill and Melinda Gates Foundation and coordinated by the International Vaccine Institute.

References
1. Talbot TO, Kulldroff M, Teven PF, Haley VB: Evaluation of spatial filters to create smoothed maps of health data. Statistics in Medicine 2000, 18:2399-2408.
2. Rushton G, Lolonis P: Exploratory spatial analysis of birth defect rates in an urban population. Statistics in Medicine 1996, 15:717-726.
3. Kafadar K: Smoothing geographical data, particularly rates of disease. Statistics in Medicine 1996, 15:2539-2560.
4. Curtis A: Using a spatial filter and a geographic information system to improve rabies surveillance data. Emerg Infect Dis 1999, 5(5):603-606.
5. Alexander FE, Ricketts TJ, Williams J, Cartwright RA: Methods of mapping and identifying small clusters of rare diseases with applications to geographical epidemiology. Geographical Analysis 1991, 23(2):158-173.

6. Lawson AB: Tutorial in biostatistics: disease map reconstruction. Statistics in Medicine 2001, 20:2183-2204.

7. Ali M, Emch M, Donnay JP: Spatial filtering using a raster geographic information system: methods for scaling health and environmental data. Health & Place 2002, 8(2):85-92.

8. Lawson AB: Statistical Methods in Spatial Epidemiology Chichester, Wiley; 2001.

9. Pearce N: The ecological fallacy strikes back. J Epidemiol Community Health 2000, 54(5):326-7.

10. Twigg L, Moon G, Jones K: Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators. Soc Sci Med 2000, 50:1109-1120.

11. Mercer WB, Hall AD: The experimental error of field trials. J Agric Sci 1991, 118:107-132.

12. Horne JK, Schneider DC: Spatial variance in ecology. Oikos 1995, 74:18-26.

13. Diez Roux AV: Investigating neighborhood and area effects on health. Am J Public Health 2001, 91(11):1783-1789.

14. Scott DW: Multivariate density estimation: theory, practice, and visualization New York, J Wiley; 1992.

15. Diez Roux AV: The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. Epidemiol Rev 2004, 26(1):104-111.

16. Bailey TC, Gatrell AC: Interactive Spatial Data Analysis Longman, Harlow; 1995.

17. Naficy AB, Trach DD, Ke NT, Chuc NT, Sorkin A, Rao MR, Sy TH, Thiem VD, Canh DG, Mahoney RT, Holmgren J, Ivanoff B, Clemens JD: Cost of immunization with a locally produced, oral cholera vaccine in Viet Nam. Vaccine 2001, 19:3720-3725.

18. Tuyet DT, Chien BT, Tung TV, Naficy A, Ali M, Lee H, Sy TH, Nichibuchi M, Clemens J, Trach DD: Clinical, epidemiological, and socioeconomic analysis of an outbreak of Vibrio parahaemolyticus in Khanh Hoa Province, Vietnam. The Journal of Infectious Diseases 2002, 186(15-20).

19. Kim YB, Okuda J, Matsumoto C, Takahashi N, Hashimoto S, Nishibuchi M: Identification of Vibrio parahaemolyticus strains at the species level by PCR targeted to the toxR gene. Journal of Clinical Microbiology 1999, 37:1173-1177.

20. Clemens J, Kotloff KL, Kay B: Generic protocol to estimate the burden of Shigella diarrhea and dysenteric mortality. WHO/V&B/99.26, Geneva: WHO; 1999.

21. Winer BJ: Statistical principles in experimental design New York: McGraw-Hill; 1992.

22. Berke O: Exploratory disease mapping: kriging the spatial risk function from regional count data. Int J Health Geogr 2004, 3(1):18.

23. Cressie NAC: Statistics for spatial data John Wiley & Sons Inc; 1991.

24. Oliver MA, Webster R, Kriging : A Method of Interpolation for Geographical Information Systems. International Journal of Geographical Information Systems 1990, 4313-32.

25. Fortin MJ, Jacquez GM: Randomization tests and spatially autocorrelated data. Bulletin of the Ecological Society of America 2000, 81:201-205.

26. Eastman JR: Guide to GIS and image processing Volume 2. Worcester, MA: Clark Labs, Clark University; 1999.

27. Briggs DJ, Elliott P: The use of geographical information systems in studies on environment and health. World Health Statistics Quarterly 1995, 48:85-94.

28. Kafadar K: Simultaneous smoothing and adjusting mortality rates in U.S. counties: melanoma in white females and white males. Statistics in Medicine 1999, 18:3167-3188.

29. Assner M: The logic in ecological: I:The logic of analysis. Am J Public Health 1994, 84(5):825-829.

30. Goovaerts P, Jacquez GM: Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. Int J Health Geogr 2004, 3:14.

31. Bucher F: Using extended exploratory data analysis for the section of an appropriate interpolation model. In Geographic Information Research: Trans-Atlantic Perspectives Edited by: Craglia M, Onsrud H. London: Taylor and Francis; 1999:391-403.

32. Ten SW, Demisie K, August D, Rhoads GG: Level of aggregation for optimal epidemiological analysis: the case of time to surgery and unnecessary removal of the normal appendix. J Epidemiol Community Health 2001, 55(3):198-203.

33. Scholten HJ, de Lepper MJC: The benefits of the application of geographical information systems in public environment health. World Health Statistics Quarterly 1991, 44:160-170.

34. Onsrud HJ, Pinto JK: Diffusion of geographic information innovations. Int J Geographic Information Systems 1991, 5(4):447-467.