Journal of the Korean Geographical Society, Vol. 34, No. 5, 1999 (473~491)

An Exploratory Spatial Analysis of Shigellosis

Key-Ho Park*

세균성 이질의 탐색적 공간분석

박 기 호*

Abstract : The incidence of shigellosis in Korea, being one of the first-class infectious diseases that are under the surveillance program of the health authority, has grown drastically since 1998. This paper attempts at geographical interpretations of spatial and temporal distribution of the shigellosis outbreak of Sasang-gu, Pusan in March 1999. Our study on shigellosis is mostly centered on an exploratory description of the spatial patterns and their expansion over time. The shigellosis intensities and relative risk measures are mapped and assessed based on both aggregated and individual geo-references of cases. A series of statistical tests for spatial dependence are performed to confirm spatial cluster of cases. We also address some of the issues involving the applicability of GIS, mapping schemes adapted to background population on the probability scale, data visualization, and techniques for the analysis of geographical variability at multiple points in time.

Key Words : Shigellosis, Medical Geography, Spatial Point Pattern Analysis, GIS

요약 : 세균성 이질은 국내 제1종 법정 전염병으로 분류되어 관리되고 있는 질환으로서 1998년 이후 그 발병 사례가 급속히 증가하고 있다. 본 연구는 1999년 3월 부산시 사상구에서 집단 발병한 세균성 이질을 대상으로 하여, 각 환자들의 발병 시점과 장소의 분포패턴에 대한 지리학적 고찰을 목적으로 한다. 환자분포의 특징적 공간패턴과 그들의 시계열적 확산 양상 등을 탐색하기 위한 방법론은 보건지리학과 지도학 및 공간통계학에 기반을 둔 공간분석기법을 중심으로 설정하였다. 분석자료는 해당 지역의 수치지형도, 지적도, 인구 센서스 자료를 포함한 GIS 데이터베이스로 구축되었다. 인구분포를 감안한 밀도구분도를 바탕으로 개별환자의 위치자료와 동 단위로 집계된 자료를 자료의 형태에 따라 분석기법을 달리하였으며, 환자 발생 밀도, 상대적 위험지수 등을 지도화하여 역학자료의 시각적 통계적 분석을 수행하였다. 환자분포의 공간적 중심위치와 분산의 변화 등 기술적 통계분석과 함께 제1차 공간속성을 커널 추정법으로 찾아 보았다. 이와 더불어 '공간적 의존성'과 관련된 제2차 공간속성은 K-함수와 시뮬레이션을 통해 분석하여 군집성 등이 통계적으로 확인되었다. 본 연구를 통해 역학조사시 GIS의 활용사례가 제시되었으며, 모집단 인구를 고려한 확률지도 작성 기법과 다양한 데이터 가시화 방법, 그리고 시계열별 발생 환자들의 지리적 변이를 분석하는데 따르는 문제들이 논의되었다.

주요어 : 세균성 이질, 보건지리학, 점 패턴의 공간적 분석, GIS

1. Introduction

Since the landmark research of cholera by John Snow(Snow, 1855), epidemiologic studies on disease have accommodated the spatial analytic perspective to such an extent that "spatial epidemiology" or "geo-pathology" became an established field in epidemiology. The public health is one of the application domains for which geographical analysis involving cartographic methodology and spatial statistics has had significant contributions (McGlashan and Blunden, 1983), and as such "medical geography" attracts many geographers today. The distinctive role of spatial analysis is most prominent in the exploration of non-random disease patterns, the identification of areas of elevated relative risk, or associations between disease incidence and social

* Assistant Professor, Department of Geography, Seoul National University

and environmental factors. Exploratory spatial analysis of disease incidences should be routine task in disease surveillance program (Teutsch and Churchill, 1994).

With the advent of Geographic Information System (GIS) in recent years, health-related geographic research has grown significantly in USA and Europe (Gatrell and Loytonen, 1999). The research on medical geography, however, has not been very active in the academic society of geography in Korea. Thus, the invaluable health statistics from diverse source are accumulated on a daily basis, but spatial interpretation of them are hard to find in literature of both geography and medicine.

This paper attempts at geographical interpretations of spatial and temporal distribution of the shigellosis outbreak of Sasang-gu, Pusan in March 1999. Our study on shigellosis is mostly centered on an exploratory description of spatial pattern and expansion, and a series of test for spatial cluster. Issues on mapping, data visualization, and techniques for the analysis of geographical variability at multiple points in time are also addressed.

1) Shigellosis in Sasang-gu, Pusan

In early March 1999, the health agency of Pusan confirmed 3 cases of *shigella sonnei* infections reported from the microbiology laboratory (Dong-A Ilbo, 1999). The infections occurred in children age 4-6 years in Sasang-gu of the city of Pusan. An epidemiologic investigation traced that the initial cases acquired shigellosis by eating contaminated food, and that the point source of exposure was a day care center named "Milal", located in Dugpo1-dong. Another point of exposure was also a day care center, "Changgu", and the date of onset was March 7. From February 26 through March 31, a large outbreak of diarrhea with 166 cases affected the community of Sasang-gu. 66 persons among them were actually identified with culture-confirmed infection. Cases occurred primarily among children; and were evenly distributed by sex. The spread of disease and the reported number of secondary cases declined and the outbreak ceases in March 31.

Shigellosis is an bacterial infectious disease, caused by a group of bacteria called *shigella*, involving the large and distal small intestine, characterized by diarrhea accompanied by fever, nausea and sometimes toxemia, vomiting, cramps and tenesmus (CDIIS 1999). In typical cases, the stools contain blood and mucus, many cases present with a watery diarrhea starting a day or two after they are exposed to the bacterium. Shigellosis usually resolves in 5 to 7 days. Some persons who are infected may have no symptoms at all, but may still pass the shigella bacteria to others. Shigella infections may be acquired either from eating contaminated food or by drinking or swimming in contaminated water. Water may become contaminated if sewage runs into it, or if someone with shigellosis swims in it. Family members and playmates of such children are at high risk of becoming infected.

Shigellosis is categorized as the first-class infectious disease, and as such is closely monitored under the surveillance program of the Korean health department. The recent increase in community-wide shigellosis has kept the public health agencies on the control alert. The outbreaks

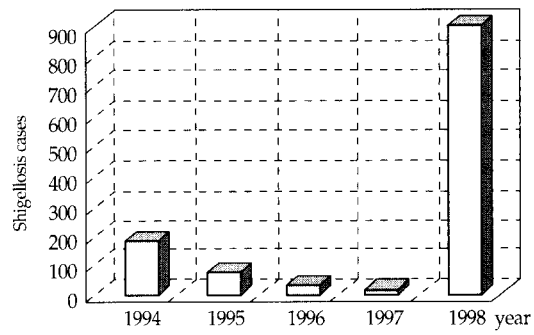

Figure 1. Shigellosis in Korea: 1994-1998

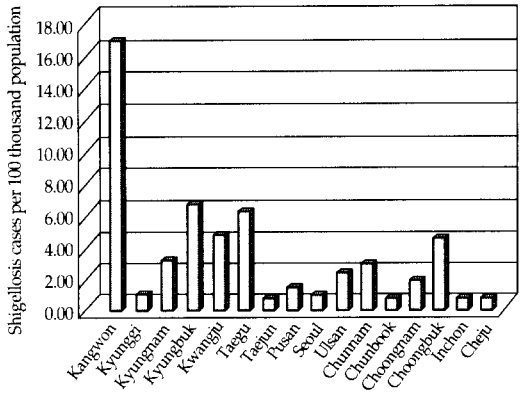

Figure 2. Shigellosis by region(1998)

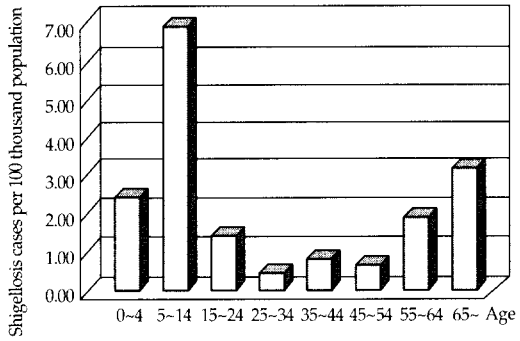

Figure 3. Shigellosis by age(1998)

of shigellosis reported since 1994 are summarized in Figure 1, and note the drastic increase of the occurrence in 1998. The Figure 2 shows that cases in 1998 prevail some of the selected regions of the country such as Kangwon and Kyungsang provinces. It also seems evident in Figure 3 that the age groups of 'over 65' and '5-14', in particular, are most vulnerable to shigellosis.

2) Study area and data compilation

As is referenced in the index map of Figure 4, the study area, Sasang-gu, is one of the 16 sub-districts in the city of Pusan. The labor-intensive light industries, located in the western and southern part of the area, dominate the economic profile of Sasang-gu. The total area of Sasang-gu is 35.84 Km^2, which covers about 4.8% of the entire city of Pusan. The residential area is about 5.76Km^2 (16% of the whole area), in which 63,845 housing units are supplied. The northeastern part of the area is mostly forest and shows little inhabitation. The population of Sasang-gu is 294,711 according to

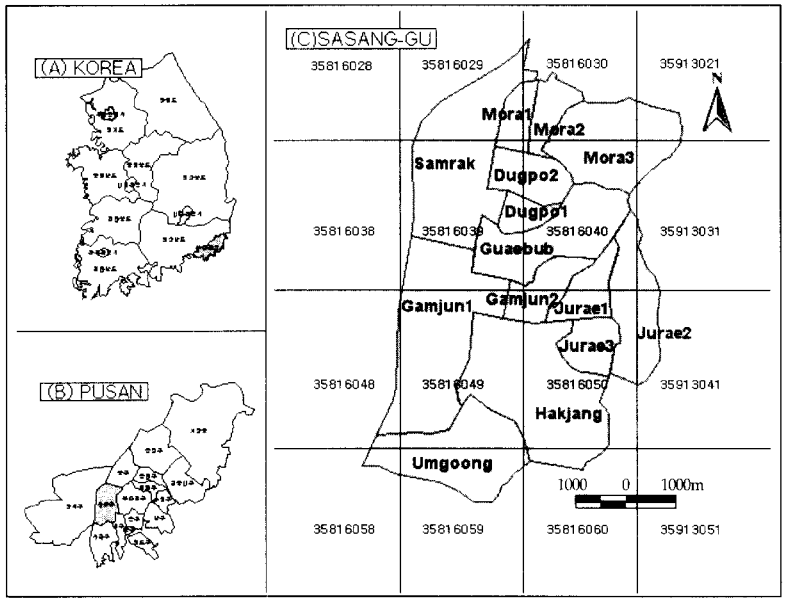

Figure 4. Index map of Sasang-gu

Table 1. Population characteristics of Sasang-gu, Pusan

Dong	Total Population	Sex Ratio	Population (0~14)	Population (65+)	Average Age
Samrak-dong (삼락동)	14,153	107.18	3,409(24.09%)	392(2.77%)	28.19
Mora1-dong (모라1동)	13,722	107.34	3,006(21.91%)	445(3.24%)	29.55
Mora2-dong (모라2동)	29,497	97.62	8,000(27.12%)	808(2.74%)	28.01
Mora3-dong (모라3동)	24,097	90.62	5,646(23.43%)	1,022(4.24%)	29.10
Dugpo1-dong (덕포1동)	18,423	105.49	4,366(23.70%)	601(3.26%)	28.72
Dugpo2-dong (덕포2동)	18,639	101.24	4,990(26.77%)	507(2.72%)	27.75
Guaebub-dong (괘법동)	23,599	105.15	5,119(21.69%)	710(3.01%)	29.30
Gamjun1-dong (감전1동)	14,282	107.55	3,451(24.16%)	376(2.63%)	28.22
Gamjun2-dong (감전2동)	10,428	101.89	2,231(21.39%)	300(2.88%)	29.59
Jurae1-dong (주례1동)	19,521	102.71	5,116(26.21%)	547(2.80%)	28.09
Jurae2-dong (주례2동)	22,487	102.31	5,154(22.92%)	706(3.14%)	29.09
Jurae3-dong (주례3동)	22,879	100.97	6,031(26.36%)	700(3.06%)	28.59
Hakjang-dong (학장동)	38,265	105.27	9,631(25.17%)	1,177(3.06%)	29.15
Umgoong-dong (엄궁동)	24,719	104.93	6,652(26.91%)	692(2.80%)	27.60
Total	294,711	102.37	72,794(24.70%)	8,989(3.05%)	28.62

Source : Census of Population and Housing, 1995.

1995 Census, and 39,555 out of 87,079 total family units are living in the apartment complexes. The people of ages between 0 to 14 accounts for about one fourth of the whole population, and 3% of the population are of ages 65 and over. A detailed statistics of the population and its composition are listed at the level of *dong* in Table 1.

As for the base maps of our study area, we

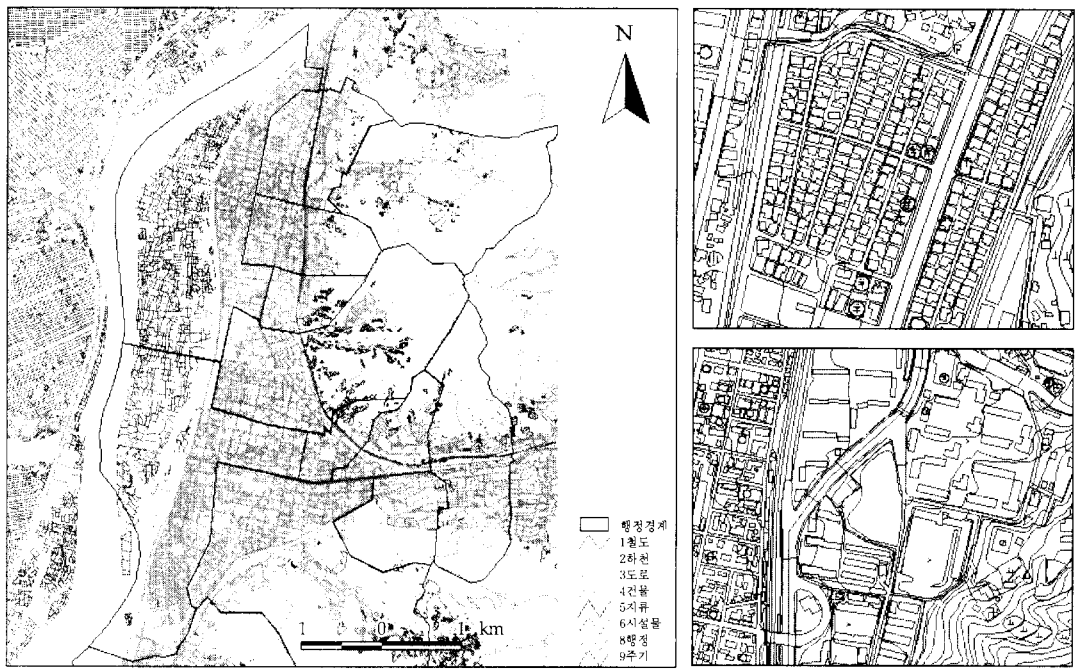

Figure 5. Geographic layers extracted from NGIS digital maps

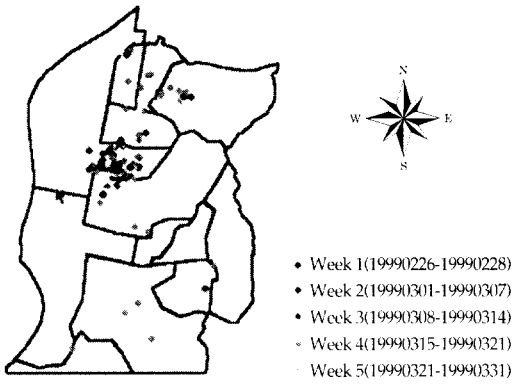

Figure 6. Pin map of shigellosis cases

acquired and ingested into a GIS database a set of 12 contiguous coverages of Korean National GIS Digital Topographic Map Series in 1:5,000 scale. The number within each rectangular division in Figure 4 represents coverage ID. Each of the coverages was converted into the 'shape file' format to be usable in the ArcView GIS environment (ESRI, 1998), and the mosaic of coverages was subsequently masked to fit the study area. The digital map contains over hundred feature categories organized in 3 tiers as is specified in the National Standard (MIT, 1996). We identified 8 feature classes as relevant to our study, and they were extracted as the thematic layers of GIS. These themes include administrative boundaries, 2 categories of residential area, road networks, and public places such as school, public bath, and swimming pool. Figure 5 is a snapshot of GIS database showing several selected layers; the right hand side of Figure 5 contains examples of zoomed-in features in the residential area of single-family housing and apartment complexes.

The epidemiologic dataset on the March outbreak of shigellosis was obtained from the public health authorities and research scientists of medicine in Pusan. The dataset contains 166 records of individuals in total who were diagnosed for shigellosis/diarrhea, and the descriptive fields for each patient include the date of onset, age, sex, and residency address, among others.

With the ancillary data and maps such as topographic map series and land registration cadastre for geo-referencing, we pinned down onto our base map the locations of 166 individual cases of confirmed shigellosis and severe diarrhea. The result is shown together with the administrative boundaries in Figure 6. The cases are divided into 5 groups based on the date of onset. The spans of time periods, which will be used consistently throughout the study, are as follows: the week 1 as from Feb.26 to Feb. 28, the week 2 as from Mar. 1 to Mar. 7, the week 3 as from Mar. 8 to Mar. 14, the week 4 as from Mar. 15 to Mar. 21, and the week 5 as from Mar 21 to Mar. 31, respectively.

As for the epidemiologic background information about population at risk, we took the 1995 Census for housing and population (Table 1) as the baseline data. The attributes available in the Census data are then integrated into the GIS base map via the standard numeric code for the smallest administrative district as the join-key.

The computing environment for subsequent analyses includes S-Plus v. 4.5 (MathSoft, 1999) with extended modules for spatial statistics in addition to ArcView GIS v. 3.1. Since the capabilities of these tools are limited in the area of spatial statistical analysis, we had written the program scripts and extension modules for our analyses.

2. Spatial analysis

The spatial and spatio-temporal distribution of disease remains one of the oldest puzzles and yet awaits new theoretical framework and methodologies (Cliff and Haggett, 1988). The proper use of cartographic and statistical methods together with spatial data manipulations available in GIS can greatly advance our understanding of the epidemiology of diseases (de Lepper *et al.*, 1988). The spatial analyses in this section are

organized according to the data types involved, and are presented in steps ranging from visualization, exploration, and to confirmatory analysis.

1) Visualization and exploration of shigellosis incidence

(1) Analysis of geo-demographic characteristics of cases

As an initial step towards understanding the nature of shigellosis occurrences, the demographic characteristics of cases need to be examined. A set of visual plots accompanied by descriptive statistics is generated to verify if there were significant difference of incidences among the groups of different age and/or sex. Likewise the difference among the periods of time is investigated. The usual descriptive statistics are tabulated in Table 2 and Table 3. The two districts, Gamjun1 and Jurae1 dongs, had no incidence. Overall, it is clear from the tables that the incidences were somewhat evenly distributed over the time, but were confined to 5 or 6 districts, i.e., clustered over the space.

Figure 7 contains a visual summary of age composition of cases, in which each of the time periods may be looked into separately. Similarly, Figure 8 plots the age compositions in different parts of the area. We use box-plots for visualization since box-plots have proven to be quite an effective exploratory tool, especially when several box-plots are placed side by side for comparison (Cleveland, 1993). The most striking visual feature is the box which shows the limits of the middle half of the data (the line inside the box represents the median). Extreme points are also highlighted. Box-plots not only show the location and spread of data but indicate skewness, as well. The whiskers are drawn to the nearest value not beyond a standard span

Table 2. Age distribution of cases by area

Dong	Cases	Min	1st Quartile	Median	Mean	3d Quartile	Max
Gamjun2	13	2.0	5.5	9.0	14.0	20.0	31.0
Guaebub	6	1.0	6.0	8.0	15.5	12.0	76.0
Dugpo1	10	2.0	5.0	6.0	20.7	33.0	87.0
Dugpo2	13	2.0	4.0	7.0	14.1	26.0	39.0
Mora1	55	2.0	2.5	5.0	8.2	6.0	29.0
Mora 2	25	3.0	4.8	7.0	11.4	7.8	38.0
Mora3	29	2.0	5.0	8.0	9.5	11.0	34.0
Samrak	3	2.0	5.0	7.0	17.5	29.0	76.0
Umgoong	2	2.0	14.0	26.0	26.0	38.0	50.0
Jurae2	2	3.0	4.0	5.0	5.0	6.0	7.0
Jurae3	6	15.0	18.0	21.0	21.0	24.0	27.0
Hakjang	2	5.0	9.0	10.0	16.5	26.0	34.0
Total	166	1.0	5.0	7.0	16.2	28.8	87.0

Table 3. Age distribution of cases by time

Dong	Cases	Min	1st Quartile	Median	Mean	3d Quartile	Max
week 1	3	3.0	4.0	5.0	13.0	18.0	31.0
week 2	33	2.0	4.0	5.0	9.5	6.0	69.0
week 3	38	2.0	5.0	12.5	23.8	33.8	87.0
week 4	52	2.0	5.0	9.0	16.2	31.0	50.0
week 5	40	1.0	6.0	9.0	14.9	12.8	75.0

from the quartiles; the standard span is 1.5 times of the inter-quartile range (IRQ), and points beyond (outliers) are drawn individually.

The median age of cases is within the bounds of 5 and 12, and does not show much variation across time. The median ages by districts, however, seem to have spatial variations. Note, for example, that the median ages for Umgoong and Jurae3 dong are relatively high. There could be two explanations for this. First, the two districts have such a few total incidences that the statistic could be unstable. Second, the incidences of these districts were from the secondary infection involving the grown-ups of the family. Overall, the most vulnerable group to the disease seemed those of around 9 years of age. Although the cases range from 1 to 87 years old, the inter-quartile range of age distribution also seems to be stable except that of the last week. The week 2 and week 4 contain a number of outliers in elderly, and this seems to be due to within-family contagion. Interestingly enough, the cases were evenly distributed among male and female regardless of time and place. Thus, we do now show any tables or plots here, and the factor of sex may well be excluded in the remaining analysis.

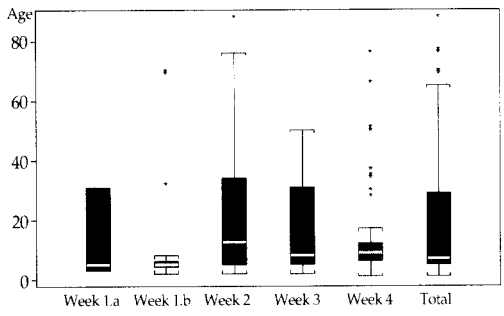

Figure 7. Age structure of cases by time

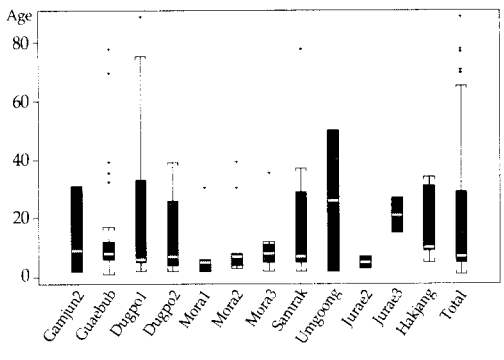

Figure 8. Age structure of cases by area

(2) Analysis of spatially aggregated incidence rates

① Standardized Morbidity Ratio (SMR) and Probability Mapping

The usual was of depicting spatial distribution of disease incidence rates is by maps. One of the popular indices conveying epidemiologic measure of disease incidence anomalies is so-called *standardized mortality/morbidity ratio* (SMR). SMR is a relative measure of health risk involving a normalization of raw count data by the background population and its cohorts (Sahai and Khurshid, 1996). Put another way, SMR is the ratio of "observed" count of cases occurred in an area over the count of cases that is "expected" considering the overall rate and the population at risk in the area. For a value of each unit in the map of SMR, we first estimate an overall expected rate, $\hat{\mu}_i$, for each area i by:

$$\hat{\mu}_i = n_i \left(\frac{\sum y_i}{\sum n_i} \right)$$

where y_i is the count of observations of the area i as independent Poisson random variable, and n_i is the population of the corresponding area. The final mapped value will then be y_i over $\hat{\mu}_i$.

We produced SMRs for each specific period of time, and the spatial variation of shigellosis incident rates are explored by a series of choropleth maps as in Figure 9. The Dugpo1-dong, which was the origin of initial outburst, is consistently highlighted as being the area of elevated risk. The disease seemed to be controlled to a moderate level in Mora1-dong and Mora2-dong, even though the total counts of incidence in these districts are very

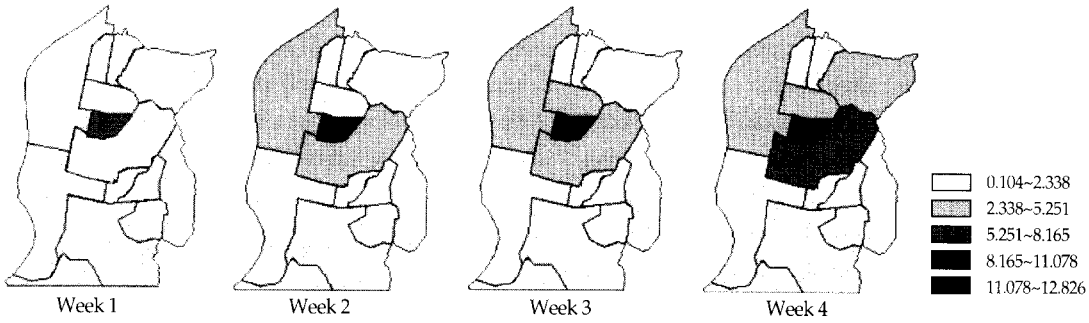

Figure 9. Choropleth map of SMR

Table 4. SMR of Sasang-gu by time

Dong	Week 1	Week 2	Week 3	Week 4	Mean
Samrak	1.5	2.9(4.4)	3.5(7.9)	4.1(12.0)	3.0
Mora1	0.7	0.7(1.4)	2.0(3.3)	2.3(5.6)	1.4
Mora2	0.1	0.5(0.6)	1.1(1.8)	1.4(3.2)	0.8
Mora3	0.5	1.1(1.6)	2.0(3.6)	2.5(6.1)	1.5
Dugpo1	5.3	8.7(14.0)	11.0(25.0)	12.8(37.8)	9.5
Dugpo2	0.6	1.2(1.8)	3.0(4.8)	5.2(10.0)	2.5
Guaebub	1.2	2.5(3.7)	4.1(7.8)	5.9(13.7)	3.4
Gamjun1	0.3	0.3(0.6)	0.3(0.9)	0.3(1.2)	0.3
Gamjun2	0.5	0.9(1.4)	1.8(3.1)	1.8(4.9)	1.2
Jurae1	0.2	0.2(0.4)	0.2(0.6)	0.2(0.8)	0.2
Jurae2	0.2	0.2(0.4)	0.2(0.6)	0.6(1.2)	0.3
Jurae3	0.2	0.3(0.5)	0.5(1.0)	0.5(1.5)	0.4
Hakjang	0.1	0.1(0.2)	0.5(0.7)	0.7(1.4)	0.4
Umgoong	0.2	0.2(0.4)	0.5(0.9)	0.5(1.4)	0.3

note: cumulative rate in parenthesis.

high. The relative incidence rate in Guaebub-dong showed a monotonic increase.

The exploratory mapping of SMR favored in official health statistics may be refined further if the rates are projected on a *probability* scale. The problem associated with using the rate is that it may become extremely variable when rare disease is being considered or when the population of an area under consideration is very small. For example, when we deal with a rare disease incidence, an addition of a case or two in an area may cause the relative ratio to increase dramatically. The probability map has been suggested to improve the reliability and comparability of disease rates measures (Norcliffe, 1980). The basic idea is to map the probability that the observed count in an area deviates significantly from the expected count, rather than to map the ratio of observed to expected counts. The final value of mapping for each area i will be p_i which is computed using the expected value $\hat{\mu}_i$ we derived before such that

$$p_i = \sum_{x \geq y_i} \frac{\hat{\mu}_i^x e^{-\hat{\mu}_i}}{x!} \quad (y_i \geq \hat{\mu}_i)$$

$$p_i = \sum_{x \leq y_i} \frac{\hat{\mu}_i^x e^{-\hat{\mu}_i}}{x!} \quad (y_i \leq \hat{\mu}_i)$$

The values of p_i in either tails of the probability density function indicate that the area's disease rate

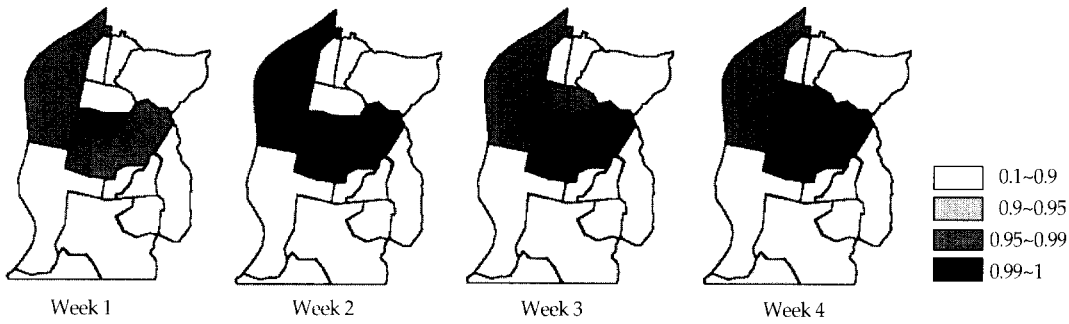

Figure 10. Choropleth map of poisson probability

is unusually high or low. A couple of features are uncovered in the probability map when compared with SMR maps. For example, Samrak-dong stands out as having such an unusual incidence rates that such numbers or higher can only be obtained with the probability of less than 5%. Likewise, the rates in the area around Mora and the southern part of the region do not differ much from that are expected.

② Dasymetric Mapping

It is often the case that the simple choropleth representation of SMR is quite misleading. The implicit assumption underlying a choropleth map of population-related attributes is that the population is distributed evenly over the areal unit (Robinson *et al.*, 1984). The administrative boundaries may form a convenient areal subdivision for data collection and report, and yet they are arbitrary with respect to geographical variation of population distribution.

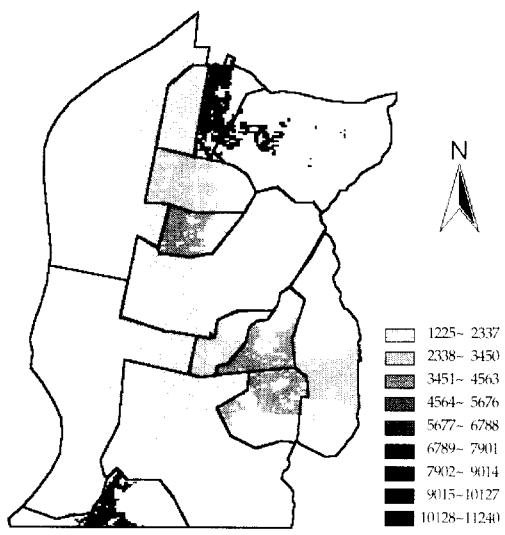

Figure 11. Dasymetric base map of population density

A dasymetric base map for population is called for in the current study since the size and the relative location of residential zones within each areal unit varies greatly. The limiting variable we employed in producing a dasymetric map is the 2 categories of residential zones, i.e., single-family housing, complex housing including duplex housing and apartment complex. Each of the residential categories was assigned different weight in dis-aggregating population considering the proportion of the family-units living in each housing type. With the dasymetric population density map such as Figure 11 in hand, the 'cases per unit area' can be appropriately converted into 'cases per unit population' in deriving intensity measures.

Figure 12 shows a series of SMR (and those of probability scale) portrayed onto the dasymetric base map, which reflects the underlying geo-demographic distribution. The area of excess cases and local variation of relative risk are conveyed much more effectively with greater detail from visual inspection of these maps. The illness

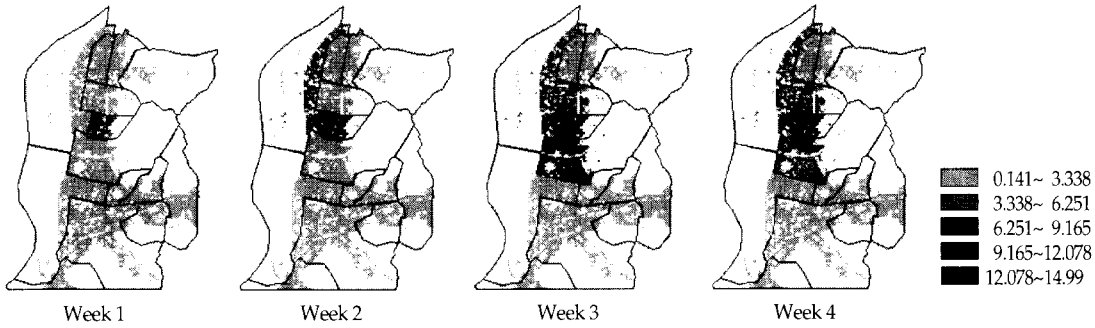

Figure 12. SMR onto dasymetric map

disproportionately affected the people in the communities of Dugpo-dong and Guaebub-dong. Some of the misleading impression of elevated risk in Samrak-dong, for example, is now cleared. The compilation of dasymetric maps is one of the prime examples where GIS plays a significant

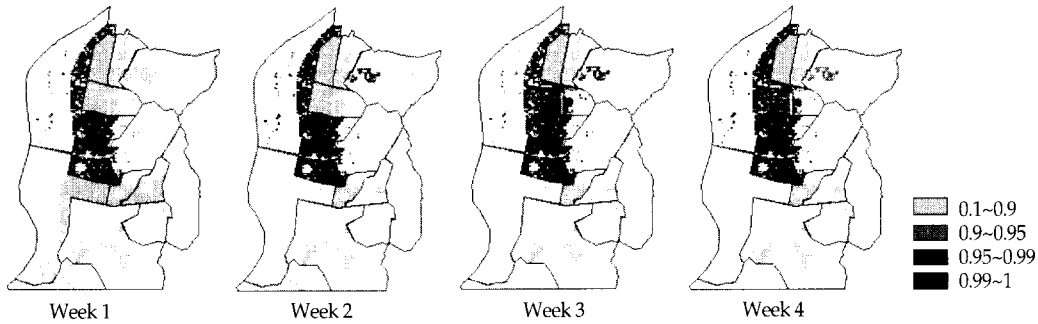

Figure 13. Poisson probability onto dasymetric map

Table 5. SMR of residential area by time

Dong	Single-family housing					Complex housing				
	week 1	week 2	week 3	week 4	Mean	week 1	week 2	week 3	week 4	Mean
Samrak	1.7	3.5(5.2)	4.2(9.4)	4.8(14.2)	3.5	1.9	1.9(3.8)	1.9(5.7)	1.9(7.6)	1.9
Mora1	0.8	0.8(1.6)	2.1(3.7)	2.5(6.2)	1.5	1.7	1.7(3.4)	3.5(6.9)	3.5(10.4)	2.6
Mora2	0.2	0.5(0.7)	0.5(1.2)	1.0(2.2)	0.6	0.3	0.8(1.1)	2.0(3.1)	2.0(5.1)	1.3
Mora3	0.8	0.8(1.6)	0.8(2.4)	0.8(3.2)	0.8	0.7	1.4(2.1)	2.5(3.6)	3.2(6.8)	1.9
Dugpo1	6.1	10.5(16.6)	12.8(29.4)	15.0(44.4)	11.1	4.1	4.9(6.0)	7.3(13.3)	8.1(21.4)	6.1
Dugpo2	0.7	1.5(2.2)	3.5(5.7)	6.2(11.9)	3.0	1.0	1.0(2.0)	2.1(4.1)	2.1(6.2)	1.6
Guaebub	1.4	3.0(4.4)	4.9(9.3)	7.0(16.3)	4.1	1.2	1.2(2.4)	1.2(3.6)	1.2(4.8)	1.2
Gamjun1	0.3	0.3(0.6)	0.3(0.9)	0.3(1.2)	0.3	1.9	1.9(3.8)	1.9(5.7)	1.9(7.6)	1.9
Gamjun2	0.6	1.1(1.7)	2.2(3.9)	2.2(6.1)	1.5	2.2	2.2(4.4)	2.2(6.6)	2.2(8.8)	2.2
Jurae1	0.3	0.3(0.6)	0.3(0.9)	0.3(1.2)	0.3	0.8	0.8(1.6)	0.8(2.4)	0.8(1.4)	0.8
Jurae2	0.2	0.2(0.4)	0.2(0.6)	0.7(1.3)	0.3	1.3	1.3(2.6)	1.3(3.9)	1.3(5.2)	1.3
Jurae3	0.2	0.4(0.6)	0.6(1.2)	0.6(1.8)	0.5	0.9	0.9(1.8)	0.9(2.7)	0.9(3.6)	0.9
Hakjang	0.1	0.1(0.2)	0.3(0.5)	0.4(0.9)	0.2	0.4	0.4(0.8)	0.4(1.2)	0.4(1.6)	0.4
Umgoong	0.2	0.2(0.4)	0.7(1.1)	0.7(1.8)	0.4	0.5	0.5(1.0)	0.5(1.5)	0.5(2.0)	0.5

role in epidemiologic studies

(3) Analysis of shigellosis incidence point pattern

Although disease mapping is a very powerful tool in understanding the epidemiology of a problem, they are by themselves are mute. As Hammond and McCullagh(1978) noted, spatial statistics are usually essential in the process of making explicit what is implicit in maps. They add precision to qualitative verbal description. By offering objective, quantitative criteria, they facilitate the making of comparisons between distributions. They may draw attention to characteristics unlikely to be noticed by intuitive inspection. In this section, we provide a preliminary description and comparison of the spatial patterns embedded in the shigellosis outbreak in terms of their central location, dispersion, and areal expansion.

① Central location

The central location may be of interest when comparing multiple distributions at a single point in time or temporal movements of a single distribution over multiple periods of time. Central location may be described in several ways which include median center, arithmetic mean, and geographic centroid. The median center is the point of intersection of two orthogonal median lines. The second method of identifying the center of a point

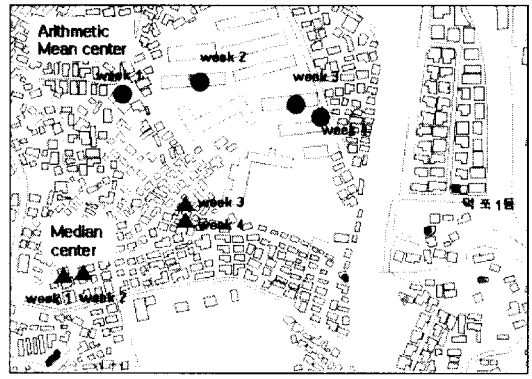

Figure 14. Weekly shift of central location of shigellosis

set is by calculating arithmetic mean of the x and y coordinates. The arithmetic mean is, in general, more sensitive to outlying points. Geographic centroid is the point about which the areal extent of the point set would balance, if density were uniform. The central locations for each period of time are depicted in Figure 14. The displacement of median centers is less than that of mean centers as expected. The central locations in general moved gradually to the east, where we may find the central streets of urban life and thus interactions among people.

② Spatial extent

Tracing the spatial extent of a disease is of value in understanding the nature of disease spread, i.e., whereto and how fast the disease progresses over space and time. Even though a detailed spatio-temporal interaction analysis is beyond the scope

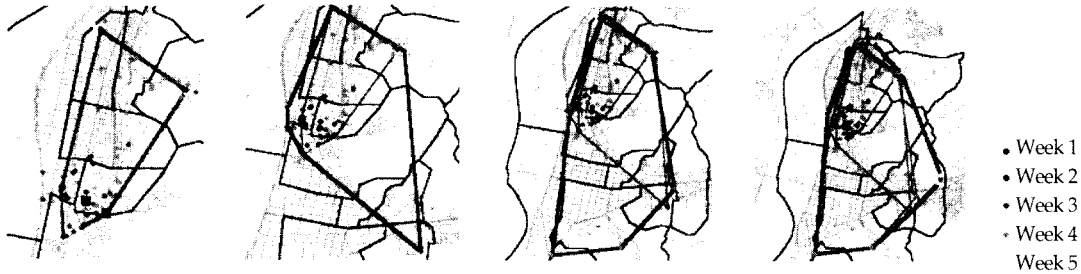

Figure 15. Spread of shigellosis

of this paper, we provide a generic and simple explorative analysis here. The minimum bounding rectangle is a simple way of determining the areal extent of a given incidence point distribution. An alternative way to set the areal extent is to use convex hull, which means the minimum bounding polygon constructed by tracing outermost points.

The preliminary exploration of spread of shigellosis over the time is made possible based on areal extent alone. As can be seen in Figure 15, the dynamic growth of areal extent, delineated in terms of convex hull, seems most evident from week 3 to week 4 in southwest direction.

③ Spatial dispersion

Given that the areal dispersion of an infectious disease will grow over time, our focus in this section will be on identifying both the spatial concentration and the temporal profile of the dispersion: i.e., when and how the "core" region grows over time. A simple measure of dispersion around the arithmetic mean or centroid may be the *standard distance deviation*, which is equivalent to the standard deviation. The standard deviational distance SD is calculated as follows where N represents the total number of observed cases:

$$SD = \sqrt{\left(\frac{\sum x_i^2}{N} - \bar{x}^2\right) + \left(\frac{\sum y_i^2}{N} - \bar{y}^2\right)}$$

The spatial dispersion of cases about the central location we computed previously may be explored in terms of the concentric circles as illustrated in Figure 16-a. The interval of the radius of circles is half the standard distance. The dispersion of the cases is to be compared with that of the underlying population (approximated by the underlying residential area). To facilitate comparison, a couple of cumulative frequency distributions of population and cases are plotted in Figure 18. Note that the x-axis represents radius of the standard deviational

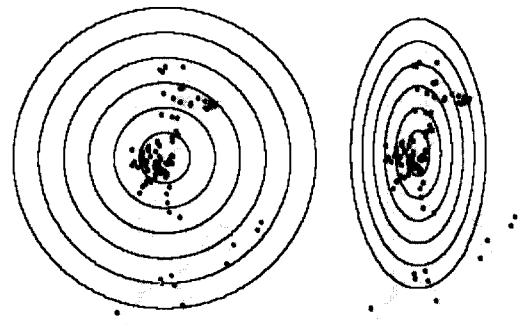

Figure 16. Circular and ellipsoidal representation of Standard Distances

circles in the unit of standard distance and the frequencies in y-axis are scaled onto 0.0 and 1.0. It is apparent from these Figures that while the population is relatively uniformly distributed within circular progression, the cases are clustered within 1.5 standard distance. The pattern is what is expected of shigellosis dispersion considering its infectious nature.

In examining the dispersal pattern of cases, the speed of disease spread is another feature of interest. To trace the dynamic growth of disease-plagued area, we proceed to break down the cases by time and then compute the standard distances of dispersion for each of the time period. The temporal profile of Figure 17 illustrates how the standard distances grew over time. A rapid expansion of shigellosis area can be observed

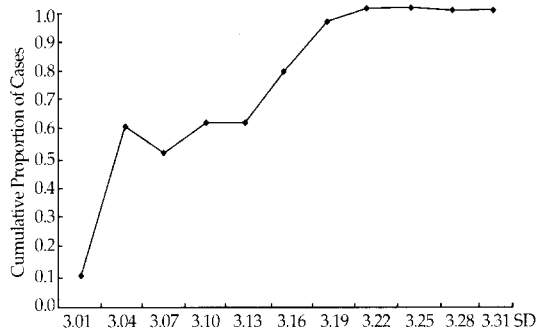

Figure 17. Temporal profile of Standard Distances

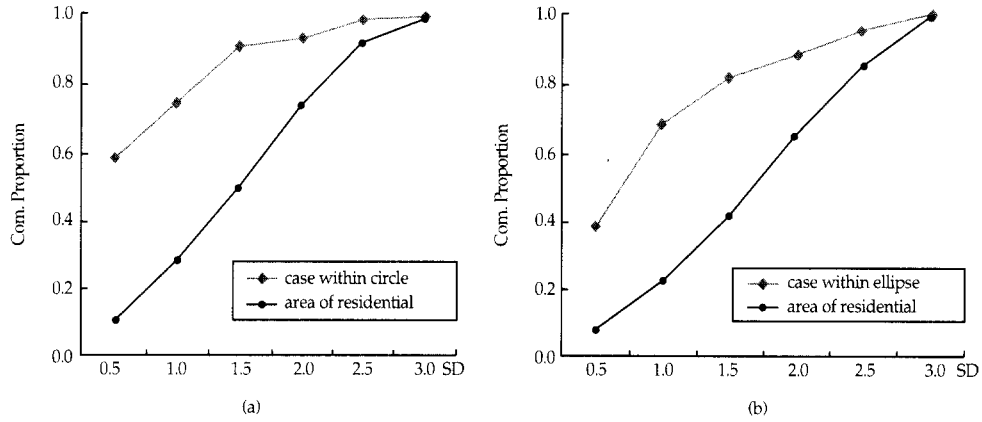

Figure 18. Comparison of cumulative distribution of shigellosis cases and population across distances from center

between Mar. 15 and Mar. 20 after the initial burst of cases in the first week of March.

The circular standard deviational distances we adopted to measure spatial dispersion, however, may be questionable since the actual dispersion is greater in the north-south direction than east-west. The underlying residential area that confines locations of cases is not circular, and thus the spatial process of shigellosis occurrence is obviously not isotropic. To remedy the situation, we employ standard deviational ellipse that has angular orientation and the standard distances are computed in both directions of long and short axis. The angle of rotation of the transposed axes is calculated using coordinates whose values are determined relative to the center, i.e.

$$x' = x_i - \bar{x} \quad and \quad y' = y_i - \bar{y}$$

The orientation of the ellipse from vertical, or 90o, is calculated as

$$\tan\theta = \frac{\sum x'^2 - \sum y'^2 \sqrt{(\sum x'^2 - \sum y'^2)^2 + 4(\sum x'y')^2}}{2\sum x'y'}$$

and the standard distances on each axis are:

$$\sigma_x = \sqrt{\frac{(\sum x'^2)\cos^2\theta - 2(\sum x'y')\sin\theta\cos\theta + (\sum y'^2)\sin^2\theta}{n}},$$

$$\sigma_y = \sqrt{\frac{(\sum x'^2)\sin^2\theta + 2(\sum x'y')\sin\theta\cos\theta + (\sum y'^2)\cos^2\theta}{n}}$$

The set of ellipses are plotted on top of the corresponding case distribution in the increment of half the standard distances in Figure 16-b. Although the general impression about dispersal is similar, the cluster within 1 standard distance is much more prominent in ellipsoidal analysis.

2) Spatial point pattern analysis of shigellosis

The aggregated data, by definition, loses locality information of spatial processes of disease occurrences such as relative distance and distribution within each district. If we have geographic locations of individual cases, however, GIS allows us to explore disease risk around a point source of health hazard exposures. Spatial point pattern analysis is one of statistical methods useful in spatial epidemiology to conduct analysis that use the individual locations instead of a set of counts for aggregated zones. There are a number of methods to evaluate the spatial pattern deduced from observed data against the theoretical distribution known for spatial point processes. The simplest of which may be the Poisson model of complete spatial randomness. With the spatial point pattern analyses in this section, we would describe and interpret the shigellosis phenomena in

terms of the so-called 'first order' and 'second order' properties of spatial process. More specifically, we measure both the intensity of observed cases and the spatial dependence between pairs of cases.

(1) First order component of spatial distribution

The properties distinguished as the first order component in the spatial phenomena are related with a global scale variation in the mean value of the spatial process (Bailey and Gartel, 1995). We take two different approaches to estimate and visualize how the intensity of cases vary over the study area: hexagonal binning and kernel estimation. Access to GIS coupled with spatial statistical modules is essential in such analysi.

① Intensity via hexagonal binning

A simple way to measure how intensity varies over the study area is by collecting counts of the number of events in subsets of study area. This is the idea behind the popular quadrat-based methods involving rectangular grids. The hexagonal tessellation is often preferred to other shapes of grid because of its enhanced visual appearance and representational accuracy (Carr *et al*, 1992). A visual representation of the global pattern in the shigellosis distribution is provided in Figure 19. As is the case with the quadrat-based analysis, the hexagonal binning results in different levels of generalization depending on the size of the unit hexagon.

② Intensity surface via kernel estimation

A viable alternative in estimating the intensity of spatial point pattern is *Kernel Estimation* method which incorporates the concept of band width (Silverman, 1986). It may be regarded as a two-dimensional extension of the univariate moving average estimator involving probability density function. Intensity estimation by kernel method produces a smooth surface map of density values in which the density at each location reflects the concentration of points in the neighboring areas defined by a given band width. Changing the parameters such as band width and distance-decaying function may generate a number of different intensity surfaces (Williamson and McLafferty, 1998). Figure 20 visualizes an example of the intensity maps in the form of 3 dimensional surface we computed via the following formulae:

$$\hat{\lambda}_\tau(s) = \frac{1}{\delta_\tau(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right), \text{ where}$$

$$\delta_\tau(s) = \int_R \frac{1}{\tau^2} k\left(\frac{(s-u)}{\tau}\right) du, \text{ and}$$

$$k(u) = \frac{3}{\pi}(1 - u^T u)^2 \quad for \ (u^T u \le 1)$$
$$0 \qquad\qquad otherwise$$

(2) Second order component of spatial distribution

Second order component of spatial phenomena results from spatial dependence in the process. It could be investigated through distances and directions between observed events under the

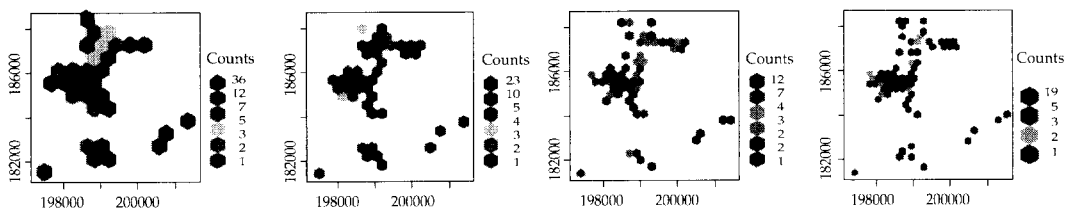

Figure 19. Hexagonal binning of shigellosis intensity

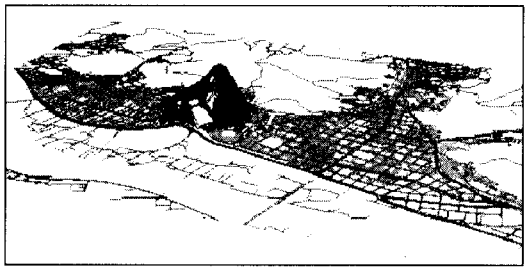

Figure 20. 3D Intensity Surface for Shigellosis, estimated using Kernel method

stationary spatial process. We would assume the stationary and isotropic process for the simplicity of modeling. Under these assumptions, statistical properties are independent of absolute location in the area, and its covariance between values at two sites s_i and s_j depends only on their relative locations, ie., distances.

① detection of spatial cluster via nearest neighbor statistics

We can investigate the degree of spatial dependence in a point pattern by examining the observed distribution of nearest distances between points. The methods based on nearest neighbor overcome the problems associated with quadrat method such as the loss of within-quadrat information and sensitivity of results to the scale. Our first attempt at identifying spatial dependence among shigellosis cases adopts the empirical cumulative distribution function $G(w)$ of the following form where w represents a case-to-case nearest distance, the numerator of the equation represents the number of cases which have nearest neighbors within the distance w, and the denominator n represents the total number of cases:

$$\hat{G}(w) = \frac{\#(w_i \le w)}{n}$$

The null hypothesis of *complete spatial randomness* (CSR) regarding point pattern is based on the notion that each location of cases constitutes a realization of a homogeneous planar Poisson process (See, for example, Cliff and Ord, 1981; Ripley 1981; and Upton and Fingleton, 1985).

Figure 21 is a comparison of empirical distribution function of nearest neighbor distances h to the theoretical distribution under CSR. It tells us the spatial pattern of shigellosis cases exhibits apparent clustering at small distance scale. In Figure 22, the dotted lines form the upper and lower bound envelope we computed for 1000 simulations of CSR, and the solid line represents the $G(w)$ profile observed for the actual disease points. It is clear that the observed spatial pattern of shigellosis is well beyond the bound that is expected of CSR, and more specifically is clustered.

② detection of spatial cluster via K-function Statistics

K-function (Ripley, 1976) is a powerful statistics for summarizing spatial dependence over a large

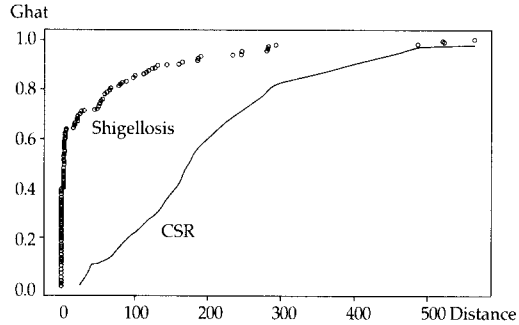

Figure 21. $\hat{G}(w)$ for observed data and CSR

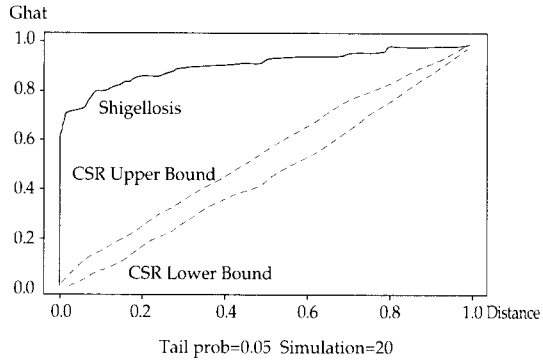

Figure 22. Simulation envelop of $\hat{G}(w)$

range of scales, i.e., consideration is not limited to the nearest neighbor alone. It doesn' t depend on the shape of the study region; It utilizes precise locations of cases in its estimation; And more importantly, it presents spatial information at all scales of pattern (Cressie, 1993). It is defined as follows:

$$\hat{K}(h) = \frac{1}{\lambda^2 R} \sum_{i \neq j} \sum \frac{I_h(d_{ij})}{w_{ij}}$$

where the lambda is a constant mean intensity, R the total area of the region, d_{ij} the distance between case i and case j, and the function I an indicator function to see if the given distance is less than h. Another advantage of the K-function is that the theoretical distributions for some spatial process models are known. For example, under the assumption of CSR, $\hat{K}(h) = \pi h^2$. So we computed $\hat{K}(h)$ and its transformation $\hat{L}(h)$ to explore spatial dependence across various distances h. $\hat{L}(h)$is defined as follows:

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{\pi}} - h$$

If the value of $\hat{L}(h)$ be close to zero, CSR is implied. The positive values imply that the pattern under consideration is clustered, and the negative values imply regular patterns. In the Figure 23 of K-functions, the lower curve represents CSR, and the upper shigellosis. In an alternative representation via L-function in Figure 24, the horizontal line of 0 represents the CSR reference line, and the upper curve represents L-function profile over distance.

The positive deviation from the baseline CSR implies rather strong evidence of clustering of shigellosis cases within the distance range of 3000. The spatial dependence among cases appears to be strongest at the scale of about 850 distance units.

(3) Correcting for geo-demographic variation in spatial cluster detection

The uniformity of population density is the

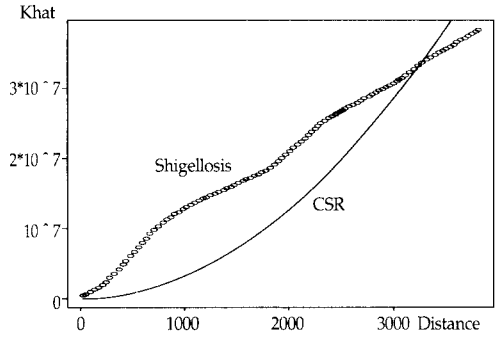

Figure 23. K-function estimates against CSR

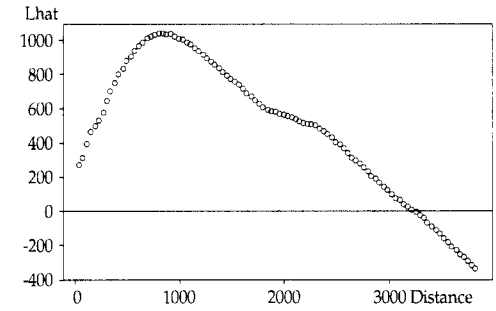

Figure 24. L(h) estimates for shigellosis cases

exception rather than the rule in practice. If the intensity of shigellosis cases be expected to change from location to location along with the population density, then the base hypothesis should involve a heterogeneous Poisson process of varying intensity rather than CSR. This implies that the variation of at-risk population (i.e., control) must be accounted for if we were to assert spatial clustering of cases *over and above* the natural spatial clustering of controls.

This initial analysis of geo-demographic distribution in our study area has led us to adopt an approach to compensate the apparent spatial variations in the background population. The idea is that if there is no clustering of cases relative to controls then the process is essentially that of random labeling of cases and controls. It follows that under the null hypothesis of no spatial clustering, the following should hold true:

$$K_{case}(h) = K_{control}(h), \text{ for all } h$$

The practical implication of the above relationship is that we could investigate the validity of null hypothesis by assessing the significance of differences among estimates of the two K-functions. This is because the K-function is invariant under random thinning. In particular, when the function $D(h) = K_{case}(h) - K_{contol}(h)$ is considered, positive values of $D(h)$ indicate spatial clustering of case over and above the degree of spatial clustering of controls. Figure 25(a) depicts the random sample of background population that was used as controls in our test, whereas Figure 25(b) represents the actual cases. The three functions mentioned above are computed and plotted simultaneously in Figure 26 to facilitate visual inspection. It is clear from the figure that a spatial cluster exists in the pattern of shigellosis incidence even after we control for the underlying population.

3. Conclusion

We provided the visual and statistical description, comparison, and interpretation of the spatial patterns embedded in the shigellosis cases of Sasang-gu, Pusan. The mapping and analytic capabilities of GIS are tapped into in generating database of a variety of base maps and geo-demographic attributes as well as the disease incidence geo-references.

The maps of age-sex corrected SMR unveiled the elevated shigellosis rate in *Guaebub*-dong which is likely to be unnoticed if raw rate or counts of occurrence is mapped. Also even though week 3 has largest number of cases, the largest spatial expansion happened in week 2. We mapped shigellosis incidence rate on the scale of Poisson probability. The motivation of using probability scale instead of rate scale is to achieve robustness and comparability of the measure. The two approaches to mapping relative measure of risk are juxtaposed and compared.

In exploring the spatial dispersal of disease relative to background population, the usage of standard deviation distance is examined. The direction and speed of spread of disease is also

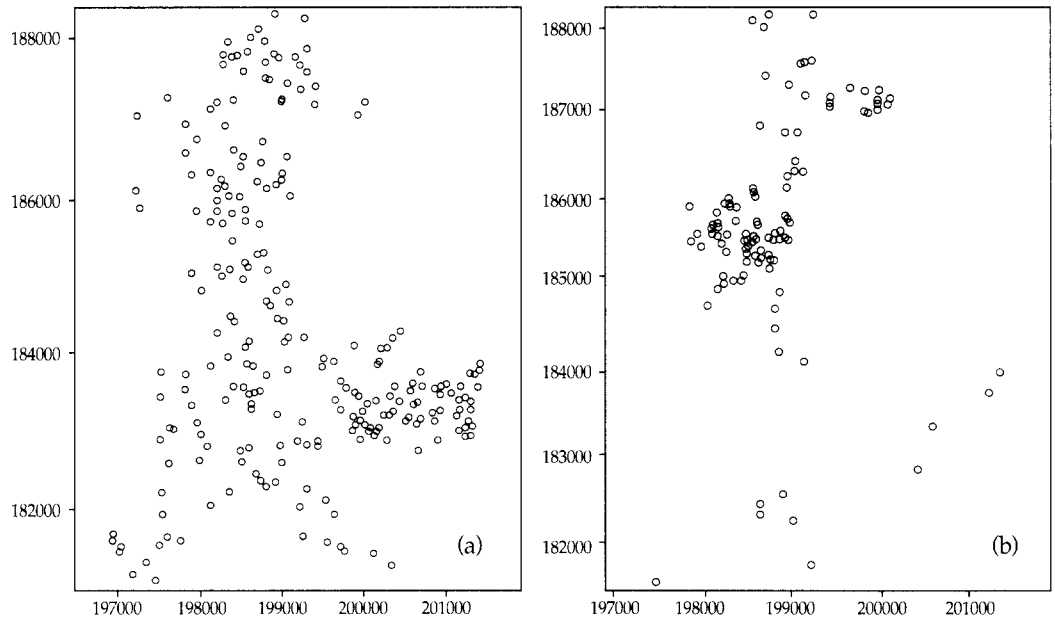

Figure 25. Distribution of (a) control samples and (b) shigellosis cases

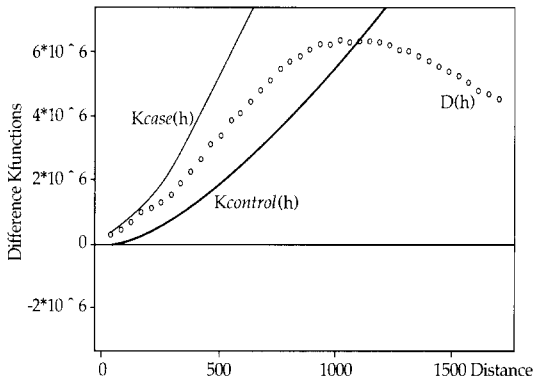

Figure 26. Difference between K functions of cases and controls

described in terms of spatial extent at multiple points in time. The spatial expansion was by and large confined to the upper half of Sasang-gu. The spatial distribution of cases remained clustered throughout the outbreak, while the central location gradually moved in the easterly direction. The 3 dimensional surface representation maps of disease intensities are produced to identify the first order trend in the shigellosis pattern.

The rationale of using dasymetric base maps for disease record is provided, and how such map may be compiled is demonstrated. The problem associated with spatially aggregated dataset and the within-unit variability of control population, in particular, seem be overcome to some extent by utilizing dasymetric maps. One of the findings that are made possible by dasymetric maps is that the people living in the single-family housing were more susceptible to the disease than those in the apartment complexes. This may be partly because of the difference of hygiene among the residential types.

A suite of statistical exploration and test based on Monte-Carlo simulation was performed to examine the spatial dependence in the data. The test results involving both G-function and K-function confirmed that the shigellosis cases showed clustered pattern over and above the control population. Further studies to assess the spatio-temporal interaction in the shigellosis expansion process are planned in the form of Mantel and space-time bivariate K-function test, and the results will be presented elsewhere.

Acknowledgements

We thank Prof. Hae-Rim Shin at School of Medicine, Dong-A University for providing us with the raw data of shigellosis cases. We are also grateful for the assistance in data compilation and programming from Eun-Hae Yu, Yang-Won Lee, Sung-Min Park, and other staffs in the GIS Research Laboratory, Department of Geography at Seoul National University.

References

Bailey, T. and Gatrel, A., 1995, *Interactive Spatial Data Analysis*, Longman, Harlow.

Carr, D. and Olsen, A. (eds.), 1992, Hexagon mosaic maps for display of univariate and bivariate geographical data, *Cartography and Geographical Information Systems*, 19, 228-236.

CDIIS, 1999, *Shigellosis*, Communicable Disease Internet Information System(CDIIS), http:// 210.122.166.108/stat/shigella.html.

Cleveland, W., 1993, *Visualizing Data*, Murray Hill: New Jersey.

Cliff, A. and Haggett, P., 1988, *Atlas of Disease Disease Distribution: Analytic Approaches to Epidemiological Data*, Blackwell.

Cliff, A. and Ord, J., 1981, *Spatial Processes: Models and Applications*, London: Pion.

Cressie, N. 1993, *Statistics for Spatial Data*, Jhon Wiley, Chichester, part III.

Cuzick J. and Edwards, R., 1990, Spatial Clustering for Inhomogeneous Population, *Journal of the Royal Statistical Society, Series* B, 52, 73-104.

De Leppper, M., Scholten, H. and Stern, R. (eds.),

1995, *The Added Value of Geographic Information Systems in Public and Environmental Health*, Kluwer Academic Publishers, Dordrecht.

DisWeb, 1999, *Shigellosis: General Information*, http://dis.mohw.go.kr/ temp_main_2.html.

Dong-A Ilbo, 1999, *Shigellosis Alert in Pusan*, 1999, Mar. 21, News scrap of Dong-A Ilbo.

ESRI , 1999, *ArcView GIS v. 3.1 Refernce Manual*, Redland.

Gartrel, A., Bailey, T., Diggle, P. and Rowlingson, B., 1996, Spatial point pattern analysis and its application in geographical epidemiology, *Transactions*, Institute of British Geographers, 21, 256-274.

Gartel, A. and Rowlingson, B., 1994, Spatial point process modeling in a GIS environment, *Spatial Analysis and GIS*, 147-163.

Hammond, R. and McCullagh, P., 1978, *Quantitative Techniques in geography: An Introduction*, Clarendon Press, Oxford.

Kulldorff, M., 1998, Statistical methods for spatial epidemiology: test for randomness, *GIS and Health*,Taylor & Francis, 49-62

MathSoft, 1999, *S-Plus v. 4.5 User's Manual*, MathSoft, Inc., Seattle: WA.

McGlashan, N. and Blunden, J., (eds.), 1983, *Geographical Aspects of Health*, Academic Press, Longon.

Ministry of Information and Telecommunication (MIT), 1966, *The Standard for Korean National Digital Base Maps*.

Norcliff, G., 1980, *Statistics in Geography: an Inferential Approach*, Hutchinson, London.

Ripley, B., 1976, The second-order analysis of stationary point processes, *Journal of Applied Probability*, 13, 255-266.

Ripley, B., 1981, *Spatial statistics*, New York: Wiley.

Robinson, A., Sale, R., Morrison, J. and Muehrcke, P., 1984, *Elements of Cartography*, fifth edition, John Wiley & Sons.

Sahai, H. and Khurshid, A., 1996, *Statistics in Epidemiology*, CRC Press.

Silverman, B., 1986, *Density Estimation*, Chapman and Hill, London.

Snow, J., 1855, *On the Mode of Communication of Cholera*, Churchill, London, Reproduced in 1965 *Snow on Cholera*, Hafner, New York.

Teutsch, S. and Churchill, R., 1994, *Principles and Practice of Public Health Surveillance*, Oxford University Press, New York.

Thomas, R., 1992, *Geomedical Systems: Intervention and Control*, Routledge, London, 200-215.

Upton, G. and Fingleton, B., 1985, *Spatial Data Analysis by Example, Vol. 1*, Wiley.

Williamson, D. and McLafferty, S. (eds.), 1998, Smoothing Crime Incident Data: New Methods for Determining the Bandwidth in Kernel Estimation.