The analysis of geographically weighted regression pertaining to gastric cancer and Taiwanese ethnic communities

Pui-Jen Tsai

Center for General Education, Aletheia University, Republic of China (Taiwan)

Abstract. The aim of this study was to examine the local versions of spatial regression associated with gastric cancer and the 4 major Taiwanese ethnic communities. In this study, we utilize the improvement of geographically weighted regression, in which the significance of parameter estimates was determined by the Benjamini-Hochberg procedure for controlling the false discovery rate, to interpret a correlation between the response and the explanatory variables. Local versions of spatial regression in correlation with gastric cancer and 4 major Taiwanese ethnic communities are mapped, including the sign and magnitude of parameter estimates, the significance determination referred to as the false discovery rate, and local R². The improvement of the GWR model is more powerful in interpreting the outcomes of the GWR version. Based on the improved GWR method, the conclusions indicate that gastric cancer is prevalent in the clusters of the central southern region in which the Hakka community resides and in the area of the major aboriginal townships in which the aboriginal group lives.

Keywords: geographically weighted regression, false discovery rate, Taiwanese ethnic communities, gastric cancer.

1. Introduction

Geographically weighted regression (GWR) is a local version of spatial regression that generates parameters disaggregated by the spatial units of analysis. This allows the assessment of spatial heterogeneity in the estimated relationships between the response and explanatory variables. In the global model, it is usual to test whether the parameter (referred to as the coefficient) estimates are significantly different from zero. This can be accomplished with a t-test, in which the t statistics and their associated p-values are usually provided in the computer output. A parameter may have a value estimated as little more than zero, that is to say, associated with a variable whose variation does not contribute to the model. Variables with non-significant parameter estimates can be dropped from the model. As one set of parameters is associated with each regression point, as well as one set of standard errors, there are potentially hundreds or thousands of tests that would be required to determine whether parameters are locally significant. Parameter estimates for variables that are close to zero often tend to be spatially clustered, indicating that in these parts of the study area, changes in this variable do not influence changes in the response variable. This is the specter that is raised by multiple testing. Using a Bonferroni correction (which downweights the significance level by the number of tests being made) is inappropriate when the tests being carried out are highly correlated, as it is highly conservative and therefore likely to miss many real differences [1,2]. A solution for GWR exists in the Benjamini-Hochberg (1995) False Discovery Rate (FDR) procedure, which modifies the significance level for each separate test in a consistent fashion [2,3].

In this study, the GWR method and a significant determination of local parameter estimates tested by FDR procedure are employed to ascertain the spatial features as they relate to gastric cancer and the 4 major Taiwanese communities. The goal is focused on the interpretation of an explanatory variable (for example: a

1 Corresponding author. Tel.: +886-2-26212121; fax: +886-2-26292229.
E-mail address: pujentsai@gmail.com.
Taiwanese ethnic communities) that contributes to these health care events under the improvement of the GWR method.

2. Methods

2.1. Study area

The study area includes the main island of Taiwan only (excluding all islets), which in 2000 comprised more than 22 million inhabitants living in an area of 36,000 km². A total of 349 local administrative government areas included 5 main urban areas, 2 secondary urban areas, 162 rural townships, and 54 plain and mountain aboriginal townships (Figure 1). According to a bulletin from the Ministry of the Interior issued in 2002, urban areas are regions having at least one metropolitan center and can include neighboring cities and townships that share socioeconomic activities. Main urban areas are defined as those with a population larger than one million, specifically, Taipei-Keelung, Kaohsiung, Taichung-Changhua, Jhongli-Taoyuan and Tainan. Secondary urban areas are defined as those with a residential population ranging from 0.3 to 1 million (for example Hsinchu and Chiayi).

2.2. Data collection and management

The Taiwan National Health Insurance (NHI) program was implemented in 1995. The coverage rate of the program has increased from 92.41% in 1995 to more than 96.16% in 2000. Coverage further increased to 98% after the inclusion of active military forces in 2001. At the beginning of 2004, NHI data related to medical care, such as the leading causes of death, were reclassified and reprocessed in relation to smaller units or areas (e.g., precincts or townships rather than the country as a whole). Regional data from the statistical analysis system (SAS) program are announced publicly by the NHI in regular annual reports (for example: NHI, 2005–2008). Disease codes were classified by gender and age. Cases with the same ID numbers but different diseases were counted as different instances.

Demographic information was purchased by the Ministry of the Interior. The smallest administrative units coded for examination of gastric cancer (ICD 151) were precincts and townships. Age-adjusted standard prevalence rates, a direct adjustment using the world population in 2000 as the standard population, was then calculated. The percentage of 4 major Taiwanese ethnic groups in each township was obtained from an official report of the Council for Hakka Affairs (2004). According to self-reports in official governmental statistics, 98% of Taiwan's population is made up of Han Chinese, while 2% are Taiwanese Aborigines. The composite category of "Taiwanese people" is often reputed by many Taiwanese to include a significant population of at least 4 constituent ethnic groups: the Hoklo (73.3%), the Hakka (13.5%), the Mainlander (8%), and the Taiwanese Aborigines (1.9%).

2.3. Geographically weighted regression

GWR is an extension of the traditional standard regression framework that allows local rather than global parameters to be estimated. It is a type of local statistics that can produce a set of local parameter estimates showing how a relationship varies over space, and then allows examination of the spatial pattern of the local estimates to get some understanding of possible hidden causes for this pattern. In contrast, a traditional regression method such as ordinary least squares (OLS) is a type of global statistics that assumes the relationship under study is constant over space, so the parameter is estimated to be the same for the entire study area. The models are fitted and mapped by GWR in the performance of ArcMap 9.3.

2.4. The significant determination of local parameter estimates

The Benjamini-Hochberg (B-H) procedure is manipulated to control the false discovery rate, which modifies the significance level for each separate test consistently. Here it is used as a solution to determine the significance of parameter estimates raised from the GWR model. Thissen et al. (2002) reported a quick and easy implementation for calculating the B-H procedure false discovery rate in Microsoft Excel. The B-H approach accomplishes control of the FDR by sequentially comparing the observed p value for each of a family of multiple test statistics, in order from largest to smallest, to a list of computed B-H critical values. The critical value on the list is computed for each test statistic, indexed by i, by linear interpolation between α/2 (for the largest observed p value) to (α/2)/m, where m is the family size, for the
smallest of the \( p \) values. The last value is the Bonferroni critical value, so the reason for the gain in power of B-H relative to Bonferroni is clear: In the B-H approach, only the smallest of the \( m \) observed \( p \) value is compared to the Bonferroni critical value; all of the other \( p \) values are computed to less stringent criteria \cite{9}. The local parameter is estimated to be significant as if the \( p \) value is less than the B-H critical value; otherwise it has only non-significance.

### 3. Results and discussion

Maps are presented of parameter estimates, the significant determination of the false discovery rate, and local \( R^2 \), in which gastric cancer figures fit the GWR models with the explanatory variables of the Hoklo, the Hakka, the Mainlander, and the Aboriginal communities, respectively (Figure 2). In the GWR models, the significant and positive signs of parameter estimates are observed in some clusters, where the northern Taiwan are found with the explanatory variable of the Hoklo community, the central southern region with the Hakka community and the majority of the aboriginal townships area with the Aborigines.

Several meta-analyses have shown a strong and consistent association between *Helicobacter pylori* infection and non-cardiac gastric cancer \cite{10-13}. The ecological study in Taiwan suggests an association between *Helicobacter pylori* infection and gastric cancer. *Helicobacter pylori* infection in early childhood may be a key issue, and a long induction time appears to be required for gastric carcinogenesis. High gastric cancer mortality areas were clustered in aboriginal townships where the prevalence of *Helicobacter pylori* was high \cite{14-15}. In this study, the aborigines clustered in the aboriginal townships are associated with gastric cancer but are not related to the aborigines in urban areas and the rural townships. The results are consistent with previous studies. The finding is that gastric cancer is highly prevalent in the Hakka community in the central southern region of Taiwan. More studies are required. However, the explanatory variable of the Hoklo community is considered an invalid interpretation, because the fit in the GWR models is weak; the overall \( R \) square in the significant determinant area is lower than 0.2.

### 4. Conclusion

The improvement of the GWR model is more powerful for interpreting the outcomes than any current GWR version. The false discovery rate helps to clarify issues such as the spatial aspects of both the response and explanatory variables at the location with significant determination. This helps planners to assess spatial risk factors and to ascertain what would be the most advantageous types of health care policies for the planning and implementation of health care services. These issues can greatly affect the performance and effectiveness of health care services and also provide a clear outline for helping us to better understand the results in depth. In Taiwan, gastric cancer is prevalent in the clusters of the central southern region in which the Hakka group resides and in the major aboriginal townships area in which the Aborigines live.
Fig. 1: Map of urban areas and aboriginal townships in the study area. Map of the study area divided into 349 administrative districts including 7 urban areas and an integrated area of 54 plain and mountain aboriginal townships.

Fig. 2: Results of the GWR model for gastric cancer and four ethnic communities in Taiwan. Hoklo community are designated by A; Hakka community, B; Mainlander community, C; Aboriginal community, D. Results of the GWR model are indicated by the numbers: parameter estimates, 1; false discovery rate, 2; local R square, 3.
5. References


