Health & Place 16 (2010) 321–330

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

# Visualizing and testing the impact of place on late-stage breast cancer incidence: A non-parametric geostatistical approach

Pierre Goovaerts *

*BioMedware Inc., 516 North State Street, Ann Arbor, MI 48104, USA*

ARTICLE INFO

*Article history:*
Received 6 August 2009
Received in revised form
21 October 2009
Accepted 28 October 2009

*Keywords:*
Breast cancer
Multiple testing
Semivariogram
Scan statistic
Poverty
Screening

ABSTRACT

This paper describes the combination of three-way contingency tables and geostatistics to visualize the non-linear impact of two putative covariates on individual-level health outcomes and test the significance of this impact, accounting for the pattern of spatial correlation and correcting for multiple testing. The methodology is used to explore the influence of distance to mammography clinics and census-tract poverty level on the rate of late-stage breast cancer diagnosis in three Michigan counties. Incidence rates are significantly lower than the area-wide mean (18.04%) mainly in affluent neighbourhoods [0–5% poverty], while higher incidences are mainly controlled by distance to clinics. The new simulation-based multiple testing correction is very flexible and less conservative than the traditional false discovery rate approach that results in a majority of tests becoming non-significant. Classes with significantly higher frequency of late-stage diagnosis often translate into geographic clusters that are not detected by the spatial scan statistic.



## 1. Introduction

Many studies in the literature report association between late-stage breast cancer diagnosis and covariates, such as socio-economic status, access to health care, marital status, ethnicity and neighbourhood of residence (Farley and Flannery, 1989; Barry and Breen, 2005; MacKinnon et al., 2007). Such relationships are explored using a variety of techniques that depend primarily on the spatial support of the data (aggregated versus individual-level data) and are typically based on a linear or log-linear model. For individual-level data, logistic regression is fitted to indicators of early/late-stage diagnosis; see examples in Barry and Breen (2005) or Hahn et al. (2007). On the other hand, Poisson regression is the method of choice for modeling count data, such as the number of late-stage breast cancer cases aggregated at the level of ZIP codes (Wang et al., 2008) or counties (Thomas and Carlin, 2003; Lin and Zhang, 2007). Multilevel models are also increasingly used to include information about the individual woman along with characteristics of her neighbourhood of residence (e.g., see Gumpertz et al., 2006). In all cases, little attention is paid to the visual description of the relationships among variables. Yet, visualization is a vital tool for the analyst, often providing a more intuitive view of the association or interaction than numerical summaries alone.

The use of multiway contingency tables is here proposed to help visualize the impact of putative factors on categorical health outcomes, such as cancer stage at diagnosis. Meyer et al. (2008) review the main tools available to visualize multiway tables, such as mosaic, association and sieve plots, and an example of the application of mosaic plots to visualize three-way log-linear models is given in Theus and Lauer (1999). For the simple case of two covariates and one binary health outcome (i.e. early versus late-stage diagnosis) a simple table of the frequency of occurrence of either outcome, say late stage, provides a much more intuitive and interpretable graphical display. This tool has been used, for example, to display the likelihood of observing a particular geological facies as a function of the recorded value of two seismic attributes (Hong et al., 2008).

Besides the exploratory visualization of the impact of covariates, the health scientist is usually interested in flagging any statistically significant behavior, which could confirm or invalidate a particular hypothesis. At first glance, a simple randomization approach (Edgington and Onghena, 2007) could be used to test whether the observed frequency of late-stage diagnosis is significantly lower or higher than what is expected under the assumption of no impact of putative factors. However, such a procedure would overlook the presence of spatial autocorrelation in the data (Fortin and Jacquez, 2000): residences of late-stage cancer diagnosis might not be distributed randomly in space. In addition, one would like to conduct such test for different levels of the covariates (e.g. different poverty levels) to account for non-linear relationships, thereby increasing the number of tests and

* Tel.: +1 734 913 1098; fax: +1 734 913 2201.
*E-mail address:* goovaerts@biomedware.com

1353-8292/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.healthplace.2009.10.017

the risk that some tests will turn out significant by chance alone. In this paper, a geostatistical simulation-based approach (Goovaerts, 2009a) is developed to incorporate spatial dependence and multiple testing correction in the testing procedure. This innovative approach is used to explore the influence of distance to mammography clinics and census-tract poverty level on the rate of late-stage breast cancer diagnosis in three Michigan counties.

## 2. Data and methods

Invasive breast cancer cases, diagnosed during the calendar years 1985–2002 in Michigan, were used to illustrate the methodology. Approximately 92% of these records, which were compiled by the Michigan Cancer Surveillance Program (MCSP), were successfully geocoded at residence at time of diagnosis. The present study focused on cases diagnosed for white women in 83 census tracts of three counties in Southwestern Michigan: Berrien, Cass and Van Buren; see Fig. 1A and B (data are aggregated for confidentiality reasons). Out of the 2118 women diagnosed with breast cancer during that time period, 18% of cases were defined as late-stage (i.e. regional and distant metastatic cancer) according to the SEER General Summary Stage classification (Young et al., 2001).

Two covariates that according to the literature (e.g. Barry and Breen, 2005; MacKinnon et al., 2007; Wang et al., 2008) could potentially explain the spatial pattern in late-stage diagnosis were considered: percentage of habitants living below the federally defined poverty line in 1990, and distance to mammography clinics located in these three counties and adjacent counties in Michigan (Fig. 1C and F). Poverty data, which were available at the census-tract level, were disaggregated using the Area-to-Point (ATP) kriging method introduced in Goovaerts (2008) to map the within-tract variation (Fig. 1D). In this illustrative example, access to screening facilities was quantified using two simple metrics: Euclidian distance between each residence and the nearest clinic based on 2006 location (Fig. 1E), and a population-based Euclidian distance to account for lower travel speeds expected in urban versus rural census tracts. In the second case, the following heuristic procedure was developed: (1) census-tract population data were disaggregated to the nodes of a 300-m grid using the same method as for poverty data in Figs. 1E, (2) each patient residence and clinic was relocated to the closest node on that 300-m grid, (3) each residence was then linked to each of the 22 clinics by a suite of linear segments joining grid nodes to form an approximately straight travel path, (4) the population data $P_d$ at each node along the path were then combined using the heuristic formula: $\Sigma 300 \times \log(1+10P_d)$, and (5) the smallest of the 22 distances was rescaled by the average population density and used as the population-based Euclidian distance. The rank correlation between the two metrics was 0.90. Since the population-based distance yielded a slightly larger $R^2$ and odd ratio in a logistic regression using cancer stage as dependent variable, it was adopted in the present study. Beware that this paper does not pretend to conduct a thorough analysis and interpretation of the spatial pattern of breast cancer incidence for this small area in Michigan, but rather, the main objective is to showcase some of the features of the proposed methodology.

### 2.1. Characterization of spatial patterns

The information about each cancer case, referenced geographically by its residence's spatial coordinates $\mathbf{u}_\alpha=(x_\alpha,y_\alpha)$, takes the form of an indicator of early/late-stage diagnosis:

$$i(\mathbf{u}_\alpha)=\begin{cases}1 & \text{if late stage diagnosis}\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

The spatial pattern of these indicator data can be characterized using the experimental semivariogram computed as

$$\hat{\gamma}_I(\mathbf{h})=\frac{1}{2N(\mathbf{h})}\sum_{\alpha=1}^{N(\mathbf{h})}[i(\mathbf{u}_\alpha)-i(\mathbf{u}_\alpha+\mathbf{h})]^2 \quad (2)$$

where $N(\mathbf{h})$ is the number of pairs of cases within a given class of distance and direction, known as spatial lag and denoted $\mathbf{h}$. The spatial increment $[i(\mathbf{u}_\alpha)-i(\mathbf{u}_\alpha+\mathbf{h})]^2$ is non-zero only if cases at $\mathbf{u}_\alpha$ and $\mathbf{u}_\alpha+\mathbf{h}$ are diagnosed at different stages. The indicator variogram $2\hat{\gamma}_I(\mathbf{h})$ thus measures how often the stage of diagnosis of two cases a vector $\mathbf{h}$ apart is different. In other words, it quantifies the transition frequency between early- and late-stage diagnosis as a function of $\mathbf{h}$. In presence of spatial clusters of early or late-stage diagnosis, the variogram value is expected to increase with the lag $\mathbf{h}$ and reaches a plateau at a distance, called range, which corresponds to the average size of these clusters. If these clusters are non-circular, different ranges will be observed along different directions, a situation referred to as spatial anisotropy. The study of indicator semivariograms can thus provide important information about the nature and scale of the process responsible for the spatial distribution of cancer stages at diagnosis.

### 2.2. Quantifying the impact of covariates

The impact of covariates, such as proximity to screening facilities or area-based measures of economic deprivation, can be assessed by computing the frequency of occurrence of the event of interest (i.e. late-stage diagnosis) for a given combination of these factors. For example, the frequency for late-stage diagnosis at residence $\mathbf{u}$ with poverty level $v(\mathbf{u})$ within the class $[v_{l-1};v_l]$ (e.g. 0–5%) and distant from the closest screening facility by $s(\mathbf{u})$ miles within the class $[s_{l'-1};s_{l'}]$ (e.g. 0–5 miles) is simply computed as the proportion of late-stage diagnosis for all $n_{ll'}$ cases residing in this poverty × distance class:

$$f_{ll'}=\text{Prob}\{\text{Late stage}|\text{v}_l,\text{s}_{l'}\}=\frac{1}{n_{ll'}}\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)i(\mathbf{u}_\alpha;v_l)i(\mathbf{u}_\alpha;s_{l'}) \quad (3)$$

where $n$ is the total number of cases, $n_{ll'}=\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha;v_l)i(\mathbf{u}_\alpha;s_{l'})$, with $i(\mathbf{u}_\alpha;v_l)=1$ if $v_{l-1}<v(\mathbf{u}_\alpha)\le v_l$ and zero otherwise, while $i(\mathbf{u}_\alpha;s_{l'})=1$ if $s_{l'-1}<s(\mathbf{u}_\alpha)\le s_{l'}$ and zero otherwise. As the number $L$ and $L'$ of classes increases, fewer cases might fall within each class, resulting in less reliable frequencies. In this paper, joint frequencies $f_{ll'}$ were smoothed by application of a $3\times3$ moving window: the frequency $f_{ll'}$ is estimated using data from classes $]v_{l-2};v_{l+1}]$ and $]s_{l'-2};s_{l'+1}]$. The marginal frequency, that is the frequency within the class of a single attribute, is simply computed as

$$f_l=\text{Prob}\{\text{Late stage}|\text{v}_l\}=\frac{1}{n_l}\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)i(\mathbf{u}_\alpha;v_l)=\frac{1}{n_l}\sum_{l'=1}^{L'} f_{ll'}n_{ll'} \quad (4)$$

with $n_l=\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha;v_l)$.

Both joint and marginal frequencies can be assembled into a $L\times L'$ frequency table and marginal frequency plots to visualize the joint and individual impacts of covariates on health outcomes. For example, Fig. 2 shows the table and plots obtained for 13 classes of poverty level and distance to clinics ($L=L'=13$). The frequency table can be viewed as a particular case of a three-way contingency table where the third variable takes only two possible values: late or early stage of diagnosis. The use of marginal and joint frequencies offers a flexible alternative to parametric models such as logistic regression since no