

METHODOLOGY

Open Access



Confidence intervals for rate ratios between geographic units

Li Zhu^{1*} , Linda W. Pickle² and James B. Pearson Jr.²

Abstract

Background: Ratios of age-adjusted rates between a set of geographic units and the overall area are of interest to the general public and to policy stakeholders. These ratios are correlated due to two reasons—the first being that each region is a component of the overall area and hence there is an overlap between them; and the second is that there is spatial autocorrelation between the regions. Existing methods in calculating the confidence intervals of rate ratios take into account the first source of correlation. This paper incorporates spatial autocorrelation, along with the correlation due to area overlap, into the rate ratio variance and confidence interval calculations.

Results: The proposed method divides the rate ratio variances into three components, representing no correlation, overlap correlation, and spatial autocorrelation, respectively. Results applied to simulated and real cancer mortality and incidence data show that with increasing strength and scales in spatial autocorrelation, the proposed method leads to substantial improvements over the existing method. If the data do not show spatial autocorrelation, the proposed method performs as well as the existing method.

Conclusions: The calculations are relatively easy to implement, and we recommend using this new method to calculate rate ratio confidence intervals in all cases.

Keywords: Spatial autocorrelation, Rate ratio, Variance, Confidence intervals, Cancer statistics, Linked micromap plot

Background

Ratios of age-adjusted rates are of interest in public health research as a means of comparing rates in a set of geographic units with the rate in the overall area or in an area considered to be a “standard”. They are useful in providing information to the general public on the health condition of the community, and to policy stakeholders on program planning and priority setting. These rates (and ratios) between geographic units are correlated due to two reasons—the first being that each region is a component (sub-region) of the overall area; and the second often referred to as the “First Law of Geography” [1], i.e., everything is related to everything else, but near things are more related than distant things. The second source of correlation is called spatial autocorrelation.

Earlier work developed methods that estimated confidence intervals (CI) on age-adjusted rates and ratios of age-adjusted rates between non-overlapping regions [2] as well as between a sub-region and its parent region [3]. The latter took into account the correlation due to overlapping between the two regions, and showed that the F and the normal approximation were more efficient than the gamma approximation and the F interval [4]. Spatial autocorrelation is well known to be present in the distribution of diseases [5, 6] but has not yet been incorporated into rate ratio interval estimates published by the National Cancer Institute.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute [7] is an authoritative source of information on cancer incidence and survival in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 30% of the US population. Cancer statistics, including rate ratios, are released annually at the national, registry, state and/or county level via the

*Correspondence: li.zhu@nih.gov

¹ Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, 9609 Medical Center Dr., Suite 4E346, Rockville, MD 20850, USA
Full list of author information is available at the end of the article

statistical software SEER*Stat [8]. Rate ratios are provided but their interval estimates are computed assuming non-overlapping regions and no spatial autocorrelation.

In this paper, we present revised confidence intervals that take spatial autocorrelation into account and show that the interval coverage is more accurate and provides a higher statistical power. The proposed method considers both overlapping (non-spatial) and spatial autocorrelation. The three methods referred to throughout the paper are (1) the Tiwari method [3] that considers overlapping between a sub-region and its parent region; (2) the non-spatial method which is a special case in our proposal in that spatial autocorrelation is ignored and only overlap is accounted for; and (3) the spatial method that includes both overlap and spatial autocorrelation. Methods (1) and (2) are equivalent, only based on different probability distribution assumptions.

Methods

Age-adjusted rates

To study the health or disease status of a region, it is common to calculate the disease rate and compare it across geographic regions. A straightforward measure is the crude disease mortality (or incidence) rate which is calculated by dividing the total number of deaths (or new cases) in a given time period by the total number of people at risk during the same time period. A major problem with the crude rate is that it is an overall average rate that does not take into account possible confounding factors. The most common confounding factor for public health data is age, because many health conditions are age-related. Instead of the crude disease rate, the age-adjusted rate is usually calculated and reported to adjust for different age profiles between regions. Assuming that there are I geographic units and J age groups in the study area, and the data available are D_{ij} , the number of deaths (or new cases), and n_{ij} , the count of the population size from region i and age group j , then the age-specific rate, R_{ij} , often expressed as number of cases per 100,000 people at risk, is calculated as

$$R_{ij} = \frac{D_{ij}}{n_{ij}} \times 100,000 \tag{1}$$

Age-adjustment could be done internally, where the age-specific rate of the standard (or reference) population is weighted by the proportions of each age group in the study population [9]. A more common age-adjustment is called the direct method where the age-adjusted rate R_i of region i is calculated as

$$\begin{aligned} R_i &= \sum_{j=1}^J w_j \frac{D_{ij}}{n_{ij}} \times 100,000 \\ &= \sum_{j=1}^J w_j R_{ij} \end{aligned} \tag{2}$$

where w_j is the proportion of population size for age group j in the standard population and $\sum_{j=1}^J w_j = 1$. Hence, the age-adjusted rate is the weighted average of age-specific rates, weighted by the standard population in order to minimize the effect of a difference in age distributions between regions. Let Ω denote the total region of interest, e.g., the entire U.S. Then the overall rate for Ω is computed by age adjusting after summing the number of deaths (numerator) and population (denominator) over all of the geographic regions, i.e.,

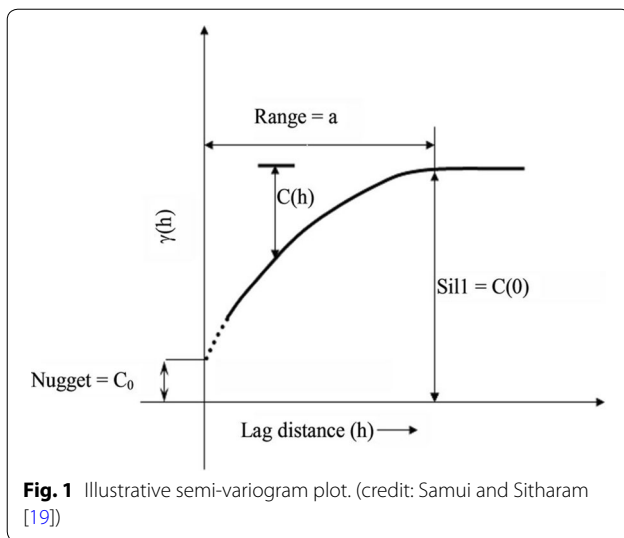
$$\begin{aligned} R_\Omega &= \sum_{j=1}^J w_j \frac{\sum_{i=1}^I D_{ij}}{\sum_{i=1}^I n_{ij}} \times 100,000 \\ &= \sum_{j=1}^J w_j R_j \end{aligned} \tag{3}$$

For the rest of this paper, we use R_i , R_Ω and D_i , D_Ω to denote the random variables for the sub-regional and overall area age-adjusted rate and count respectively. The random variable of age-and region-specific rate, r_{ij} , has expectation λ_{ij} .

Spatial autocorrelation

Suppose $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{S}\}$ represents a random process on surface $\mathbf{S} \in \mathbf{R}^2$, and $Z(s_1), Z(s_2), \dots, Z(s_I)$ represents a partial realization of the random process. $Z(\cdot)$ is said to be second-order stationary if $E[Z(\mathbf{s})] = \mu$ for all $\mathbf{s} \in \mathbf{S}$ (i.e., the mean of the process does not depend on location) and $Cov[Z(s_i), Z(s_{i'})] = C(s_i - s_{i'})$ for all $s_i, s_{i'} \in \mathbf{S}$. That is, the covariance function $C(\cdot)$, a measure of spatial correlation, depends only on the difference between locations s_i and $s_{i'}$, not on the locations themselves. If $Var[Z(s_i) - Z(s_{i'})] = 2\gamma(s_i - s_{i'})$, then $Z(\cdot)$ is said to be intrinsically stationary and the function $2\gamma(\cdot)$ is called a variogram. The semivariogram, $\gamma(\cdot)$, has a value $r(s_i - s_{i'})$ which is a function of the difference between locations s_i and $s_{i'}$ [10]. If, in addition, the covariance function $C(\cdot)$ does not depend on the direction between locations s_i and $s_{i'}$, the process is called isotropic.

For a stationary and isotropic spatial process, the semivariogram is a function of distance alone, i.e., $\gamma(\mathbf{h}) \equiv \gamma(\|\mathbf{h}\|)$ where $\|\mathbf{h}\|$ denotes the pairwise inter-point distance of vector \mathbf{h} . A plot of this function against separation distance, $\|\mathbf{h}\|$, conveys the spatial variability of the process (see Fig. 1). For a process with positive spatial autocorrelation, i.e., observations closer together are more alike than those further apart, the semivariogram value is non-decreasing with distance, indicating increasing variation with longer distance between two locations. Usually the semivariogram will approach a constant value (called the *sill*) at a large separation distance (called the *range*), beyond which the observations are considered



spatially uncorrelated. The value of semivariogram at $||\mathbf{h}|| = 0$ is referred to as *nugget effect*, and it represents the variation between two observations that are fairly close together. If the *nugget effect* is positive (larger than 0), it may be due to measurement error or a spatially discontinuous process.

For a second-order stationary spatial process $Z(\cdot)$, semivariogram is related to the covariance function as

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

If $C(\mathbf{h}) \rightarrow 0$ as $||\mathbf{h}|| \rightarrow \infty$, then $r(\mathbf{h}) \rightarrow C(\mathbf{0})$. So $C(\mathbf{0})$ is the variance of $Z(\mathbf{s})$ and the *sill* of the semivariogram. A *partial sill* is defined as the difference between the *sill* $C(\mathbf{0})$ and the *nugget effect*, or $C(\mathbf{0}) - r(\mathbf{0})$. When we are comparing two spatial processes, it is useful to measure the correlation (spatial autocorrelation) instead of covariance. By definition, the spatial *correlogram* is

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{C(\mathbf{0})} = \frac{C(\mathbf{0}) - \gamma(\mathbf{h})}{C(\mathbf{0})} \tag{4}$$

$\rho(\mathbf{h})$ is analogous to a typical correlation in that $|\rho(\mathbf{h})| \leq 1$. When the distance between two locations exceeds the *range*, $r(\mathbf{h}) = C(\mathbf{0})$ and $\rho(\mathbf{h}) = 0$, i.e., there is no spatial autocorrelation. There are a few commonly used parametric semivariogram models, including spherical, exponential, Gaussian, and power models. A plot of the observed semivariogram of our lung cancer data suggested an exponential model, which is expressed as

$$r(\mathbf{h}) = \begin{cases} 0, & h = 0 \\ c_0 + c_e[1 - \exp(-h/a_e)] & h > 0 \end{cases} \tag{5}$$

where c_0 is the nugget effect, c_e is the partial sill, and $c_0 + c_e$ is the sill. In this model, $r(\mathbf{h})$ approaches the sill

asymptotically and an effective range is defined as the distance at which the autocorrelogram is 0.05. Here, the effective range is $3a_e$ (see Fig. 1). Replacing $r(\mathbf{h})$ in (4) with (5), the spatial correlogram becomes

$$\rho(\mathbf{h}) = \begin{cases} 1 & h = 0 \\ \frac{c_e}{c_0+c_e} \exp(-h/a_e) & h > 0 \end{cases} \tag{6}$$

A larger proportion of partial sill to sill, $\frac{c_e}{c_0+c_e}$, and longer range, a_e , mean stronger spatial autocorrelation.

Variance calculation

Using notation similar to Tiwari et al. [2] where variances and confidence intervals of age-adjusted rates were derived, let $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_I)'$ denote the vector of age-specific rates for the regions 1 through I , and each component $\mathbf{r}_i = (r_{i1}, \dots, r_{ij})'$ represent the rates for the J age groups in region i . Also let $\bar{\mathbf{R}} = (R_1, \dots, R_I, R_\Omega)'$ denote the vector of age-adjusted rates for regions 1 through I and the overall age-adjusted rate R_Ω . Tiwari et al. [3] derived confidence intervals (and therefore the relevant variances and covariances) for an age-adjusted rate and of the difference and ratio of two age-adjusted rates, specifically R_i and R_Ω . as above. The calculation took into account the correlation due to the overlap between the sub-regions and the parent region. The derived 95% CI were shown to be more efficient than previously proposed methods [4]. However, both of these derivations ignored potential spatial autocorrelation among the area-specific rates. In this report, we will follow the development of Tiwari et al. [3] to derive the variance/covariance matrix for $\ln(R_i/R_\Omega)$, the logarithm of the rate ratio for region i relative to the overall rate, adding spatial autocorrelation as necessary.

The age-adjusted rate vector is a linear combination of the age-specific rate vector,

$$\bar{\mathbf{R}} = \mathbf{A}\mathbf{r}$$

where

$$\mathbf{A} = \begin{bmatrix} w' & 0 & \dots & 0 \\ 0 & w' & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ & & & \ddots & 0 \\ 0 & \dots & \dots & 0 & w' \\ g'_1 & g'_2 & \dots & \dots & g'_I \end{bmatrix}, \quad g'_i = \left(\frac{n_{i1}w_1}{\xi_{1n}}, \frac{n_{i2}w_2}{\xi_{2n}}, \dots, \frac{n_{ij}w_j}{\xi_{jn}} \right)$$

and $\xi_j = n_j/n = \sum_i n_{ij} / \sum_j \sum_i n_{ij}$. So $\text{Var}(\bar{\mathbf{R}}) = \mathbf{A}\text{Var}(\mathbf{r})\mathbf{A}'$; the dimensions of \mathbf{A} are $(I + 1) \times (IJ)$ and the dimensions of the $\text{Var}(\mathbf{r})$ variance/covariance matrix are $(IJ) \times (IJ)$, so that the dimensions of the $\text{Var}(\bar{\mathbf{R}})$ matrix are $(I + 1) \times (I + 1)$. Using the Delta Method [11]:

$$\begin{aligned} \text{Var}(\ln \bar{\mathbf{R}}) &\approx \text{diag} \left(\frac{1}{R_1^2}, \dots, \frac{1}{R_I^2}, \frac{1}{R_\Omega^2} \right) \text{Var}(\bar{\mathbf{R}}) \\ &= \text{diag} \left(\frac{1}{R_1^2}, \dots, \frac{1}{R_I^2}, \frac{1}{R_\Omega^2} \right) A \text{Var}(\mathbf{r}) A' \end{aligned}$$

Our goal is to find the variance of $\ln(R_i/R_\Omega)$, the logarithm of the ratio of the rate for area i to the overall rate:

$$\begin{aligned} \text{Var}(\ln(R_i/R_\Omega)) &= (1-1) \text{Var} \begin{pmatrix} \ln R_i \\ \ln R_\Omega \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \text{Var}(\ln R_i) + \text{Var}(\ln R_\Omega) - 2\text{Cov}(\ln R_i, \ln R_\Omega) \end{aligned}$$

Therefore, we need to compute $\text{Var}(\mathbf{r})$, multiply by the factors in the above equation to estimate $\text{Var}(R_\Omega)$, and then use the components of that result to compute the variance of the logarithm of the rate ratio.

Recall that age-place-specific rates $\mathbf{r} = (r_{11}, r_{12}, \dots, r_{1J}, r_{21}, \dots, r_{2J}, \dots, r_{I1}, \dots, r_{IJ})'$. Assuming that D_{ij} are spatially dependent Poisson random variables with means $n_{ij}\lambda_{ij}$, we can write the variance of the age-place-specific rates as a matrix where the diagonal represents the variance of independent rates plus a matrix of off-diagonal terms representing the spatial autocorrelation. That is,

$$\text{var}(r_{ij}) = \text{var} \left(\frac{D_{ij}}{n_{ij}} \right) = \frac{\text{var}(D_{ij})}{n_{ij}^2} = \frac{\lambda_{ij}}{n_{ij}}, \text{ and so}$$

$$\begin{aligned} \text{Var}(\mathbf{r}) &= \text{diag} \left(\frac{\lambda_{11}}{n_{11}}, \dots, \frac{\lambda_{1J}}{n_{1J}}, \dots, \frac{\lambda_{I1}}{n_{I1}}, \dots, \frac{\lambda_{IJ}}{n_{IJ}} \right) \\ &\quad + \text{cov} \left(\frac{D_{ij}}{n_{ij}}, \frac{D_{i'j'}}{n_{i'j'}} \right) \end{aligned}$$

where $(i, j) \neq (i', j')$. The Tiwari method assumed independence of D_{ij} 's across both age and place, so that the 2nd (covariance) term in $\text{Var}(\mathbf{r})$ above was a matrix of all zeroes. We will assume that the risk of disease is independent across age groups but the risk can be correlated among nearby places because of shared risk factors. That is, we assume independence across age but allow the possibility of spatial autocorrelation across places.

We will assume that the structure of spatial autocorrelation in our data follows the exponential form. Therefore

$$\text{cov} \left(\frac{D_{ij}}{n_{ij}}, \frac{D_{i'j'}}{n_{i'j'}} \right) = \rho_{ii'} \sqrt{\text{var} \left(\frac{D_{ij}}{n_{ij}} \right) \text{var} \left(\frac{D_{i'j'}}{n_{i'j'}} \right)} = \rho_{ii'} \sqrt{\frac{\lambda_{ij}}{n_{ij}} \frac{\lambda_{i'j'}}{n_{i'j'}}$$

where $\rho_{ii'}$, according to formula (6), represents the spatial autocorrelation between areas i and i' with a form $\rho_{ii'} = \frac{c_e}{c_0+c_e} \exp(-h_{ii'}/a_e)$. Here $h_{ii'}$ is the distance between the centroids of areas i and i' . In the case of

positive spatial autocorrelation between two rates, $\text{Var}(\mathbf{r})$ as calculated above will be the sum of two positive components. Therefore $\text{Var}(\mathbf{r})$ will be a larger value than when assuming geographic independence, i.e., no spatial autocorrelation, as in the Tiwari method. Define a $(IJ) \times (IJ)$ matrix \mathbf{C} as

$$\mathbf{C} = \begin{bmatrix} 0 & C_{12} & \dots & C_{1I} \\ C'_{12} & 0 & \dots & C'_{2I} \\ \dots & \dots & \dots & \dots \\ C'_{I1} & C'_{2I} & \dots & 0 \end{bmatrix}$$

where $C_{ii'} = \frac{c_e}{c_0+c_e} \exp(-h/a_e) \sqrt{\frac{r_{ij}}{n_{ij}} \frac{r_{i'j'}}{n_{i'j'}}$ for $i \neq i'$, and $C_{ii} = 0$, both are $J \times J$ matrix where the rows are indexed by j and the columns by j' . So

$$\begin{aligned} \text{Var}(\ln \bar{\mathbf{R}}) &= \text{diag} \left(\frac{1}{R_1^2}, \dots, \frac{1}{R_I^2}, \frac{1}{R_\Omega^2} \right) \\ &\quad \left\{ A \left(\text{diag} \left(\frac{\lambda_{11}}{n_{11}}, \dots, \frac{\lambda_{1J}}{n_{1J}}, \dots, \frac{\lambda_{I1}}{n_{I1}}, \dots, \frac{\lambda_{IJ}}{n_{IJ}} \right) \right) A' + A \mathbf{C} A' \right\} \end{aligned}$$

where \mathbf{C} is the block matrix defined above. Multiplying the first term components yields

$$\begin{aligned} \text{diag} \left(\frac{1}{R_1^2} \sum_j \frac{\lambda_{1j} w_j^2}{n_{1j}}, \frac{1}{R_2^2} \sum_j \frac{\lambda_{2j} w_j^2}{n_{2j}}, \dots, \frac{1}{R_I^2} \sum_j \frac{\lambda_{Ij} w_j^2}{n_{Ij}}, \right. \\ \left. \frac{1}{R_\Omega^2} \sum_i \sum_j \frac{n_{ij} \lambda_{ij} w_j^2}{n^2 \xi_j^2} \right) \end{aligned}$$

This matches the result of Tiwari et al. [2] Appendix A, for spatially independent places. Substituting for ξ_j , the final term of this result can be rewritten as

$$\frac{1}{R_\Omega^2} \sum_i \sum_j \frac{n_{ij} \lambda_{ij} w_j^2}{n^2 \xi_j^2} = \frac{1}{R_\Omega^2} \sum_i \sum_j \frac{n_{ij} \lambda_{ij} w_j^2}{n^2} (I+1)(I+1)$$

Multiplying the components of the second term $A \mathbf{C} A'$ results in a square matrix of dimension $I+I$ that provides the additional variation and correlation due to spatial autocorrelation. Components of this matrix are

$$\begin{cases} 0 & \text{for } (i, i), i = 1, 2, \dots, I \\ \frac{1}{R_1^2} \sum_{j'} \sum_j w_j w_{j'} \rho_{ii'} \sqrt{\frac{\lambda_{ij}}{n_{ij}} \frac{\lambda_{i'j'}}{n_{i'j'}}} & \text{for } (i, i'), i = 1, 2, \dots, I \\ \frac{1}{R_\Omega^2} \sum_{i' \neq i} \sum_{j'} \sum_j \frac{n_{ij} w_{j'} w_j}{n_j} \rho_{ii'} \sqrt{\frac{\lambda_{i'j'}}{n_{i'j'}} \frac{\lambda_{ij}}{n_{ij}}} & \text{for } (I+1, i'), i' = 1, 2, \dots, I \\ \frac{1}{R_\Omega^2} \sum_{i'} \sum_{i \neq i'} \sum_{j'} \sum_j \frac{n_{ij} n_{i'j'} w_j w_{j'}}{n_j n_{j'}} \rho_{ii'} \sqrt{\frac{\lambda_{ij}}{n_{ij}} \frac{\lambda_{i'j'}}{n_{i'j'}}} & \text{for } (I+1, I+1) \end{cases}$$

Combining relevant terms from the above results, we see that

$$\begin{aligned}
 \text{Var}(\ln(R_i/R_\Omega)) &= \text{Var}(\ln R_i) + \text{Var}(\ln R_\Omega) - 2\text{cov}(\ln R_i, \ln R_\Omega) \\
 &= \frac{\text{Var}(R_i)}{R_i^2} + \frac{\text{Var}(R_\Omega)}{R_\Omega^2} - 2\frac{\text{cov}(R_i, R_\Omega)}{R_i R_\Omega} \\
 &= \frac{1}{R_i^2} \sum_i \frac{\lambda_{ij} w_j^2}{n_{ij}} + \frac{1}{R_\Omega^2} \left(\sum_i \sum_j \frac{n_{ij} \lambda_{ij} w_j^2}{n_j^2} + \sum_{i' \neq i} \sum_{j'} \sum_j \frac{n_{i'j'} w_{j'} w_j}{n_{i'} n_{j'}} \rho_{ii'} \sqrt{\frac{\lambda_{ij} \lambda_{i'j'}}{n_{ij} n_{i'j'}}} \right) \\
 &\quad - 2\frac{1}{R_\Omega R_i} \left(\sum_j \frac{w_j^2 \lambda_{ij}}{n_j} + \sum_{i' \neq i} \sum_{j'} \sum_j \frac{n_{i'j} w_{j'} w_j}{n_j} \rho_{ii'} \sqrt{\frac{\lambda_{i'j} \lambda_{ij}}{n_{i'j} n_{ij}}} \right)
 \end{aligned} \tag{7}$$

Formula (7) shows that the variance of a rate ratio is comprised of three components. If there were no correlation at all between the area i and standard rates, i.e., area i is not a sub-region of the standard and there is no spatial autocorrelation among the area rates, then the variance would simply be the first two terms in Eq. (7), $\frac{1}{R_i^2} \sum_i \frac{\lambda_{ij} w_j^2}{n_{ij}} + \frac{1}{R_\Omega^2} \sum_i \sum_j \frac{n_{ij} \lambda_{ij} w_j^2}{n_j^2}$, representing the case of no correlation. The variance due to overlap between region i and the overall region will be reduced by $2\frac{1}{R_\Omega R_i} \sum_j \frac{w_j^2 \lambda_{ij}}{n_j}$, the fourth term of Eq. (7), representing the case of area overlap. The additional variance due to spatial autocorrelation is

$$\begin{aligned}
 &\frac{1}{R_\Omega^2} \sum_{i'} \sum_{i' \neq i} \sum_{j'} \sum_j \frac{n_{i'j'} w_{j'} w_j}{n_{i'} n_{j'}} \rho_{ii'} \sqrt{\frac{\lambda_{ij} \lambda_{i'j'}}{n_{ij} n_{i'j'}}} \\
 &\quad - 2\frac{1}{R_\Omega R_i} \sum_{i' \neq i} \sum_{j'} \sum_j \frac{n_{i'j} w_{j'} w_j}{n_j} \rho_{ii'} \sqrt{\frac{\lambda_{i'j} \lambda_{ij}}{n_{i'j} n_{ij}}}
 \end{aligned} \tag{8}$$

the third and fifth terms of Eq. (7). The first component in formula (8) is the additional variance in $\text{Var}(\ln R_\Omega)$ and the second component is the additional covariance in $\text{Cov}(\ln R_i, \ln R_\Omega)$ due to spatial autocorrelation. In most cases, adding spatial autocorrelation makes the additional variance (8) a negative value since the term representing -2 times the covariance dominates the sum of the two terms. When a sub-region accounts for a large proportion of the population in the parent region, the correlation between the sub-region and parent region is large because of a large overlap in the population, and so the additional covariance due to spatial correlation is relatively small. In that case, the first term in formula (8) dominates the overall value. Thus, although adding spatial autocorrelation will usually reduce the total variance (7), a high correlation between some sub-regions and the overall region can result in a larger variance for the rate ratio.

The full term in formula (7) is referred to as the spatial method, and subtracting the extra variance (8) due to spatial autocorrelation, we get the non-spatial version of

the variance of the logarithm of rate ratio. To transform back to rate ratios, we apply the delta method again to get

$$\text{Var}(R_i/R_\Omega) = (R_i/R_\Omega)^2 \text{Var}[\ln(R_i/R_\Omega)] \tag{9}$$

All age-place-specific rate (λ_{ij}) terms in the variance calculation will be approximated by the observed rates R_{ij} .

Results

To explore the impact of spatial autocorrelation in the confidence intervals of rate ratios, we will apply the methods with and without spatial autocorrelation to cancer incidence data from the SEER Program and cancer mortality data from the National Centers for Health Statistics [12]; both incidence and mortality data are available via the SEER*Stat software. The current SEER*Stat software provides age-adjusted rates with associated standard errors, confidence intervals, and between-geographic-area rate ratios with associated intervals and the p value (to test the rate ratio equals to 1). The rate ratio intervals are calculated using the Tiwari et al. [2] method (non-overlapping) but when rate ratio between two overlapping regions are requested, an alert message pops up that reads “The algorithms for the confidence intervals and the significance testing assume non-overlapping groups. Please use caution when interpreting these results.”

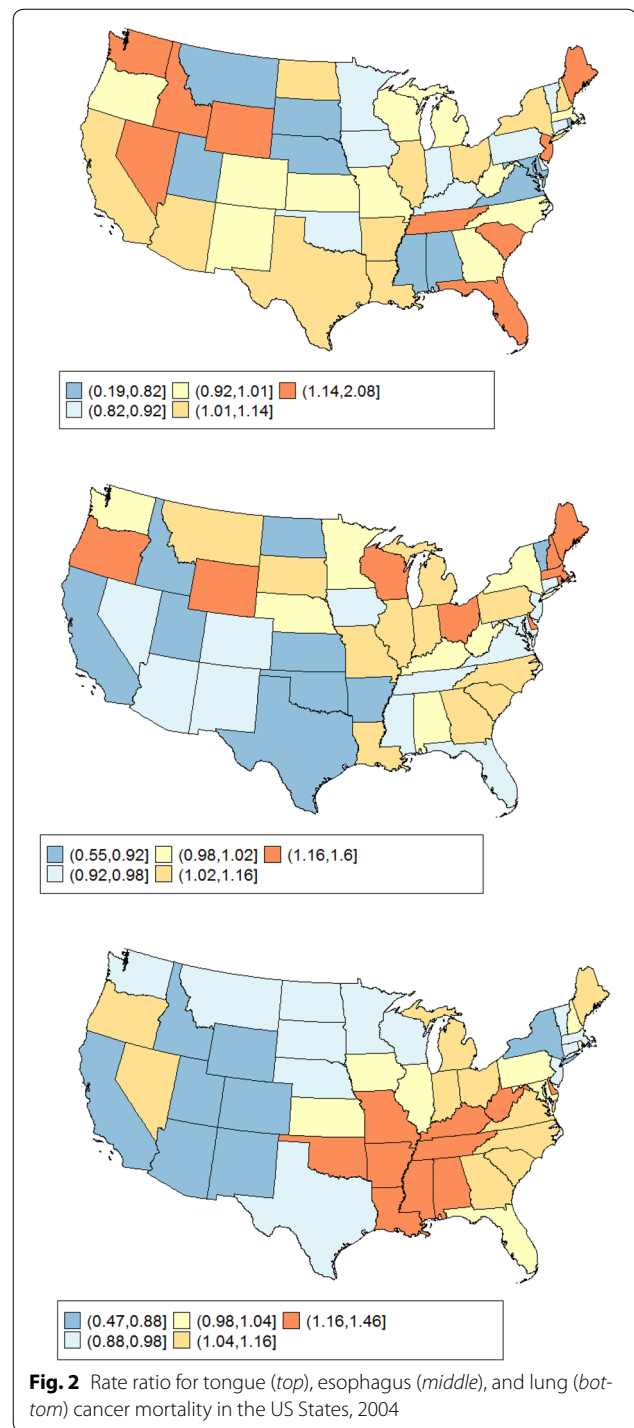
All age-adjusted rates in this paper are calculated using the 2000 US Standard Population [13] using the direct method, and the unit of the age-adjusted rates is per 100,000 people at risk. In addition to the real datasets, corresponding simulated rate ratios were created under the assumption that the logarithm of rate ratios comes from a spatial Gaussian process. The variation of the mean vector and variance-covariance structure will be described below. The simulation was implemented using the *spam* package [14] in R [15] with 10,000 realizations were created for each simulation.

State-level data on different cancers

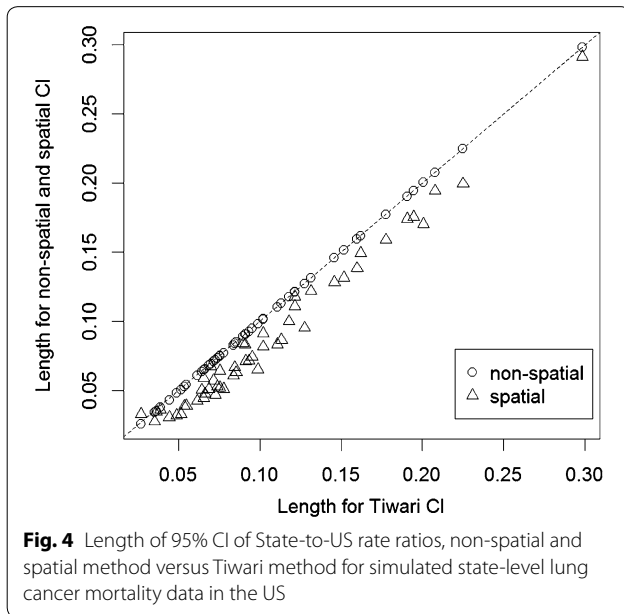
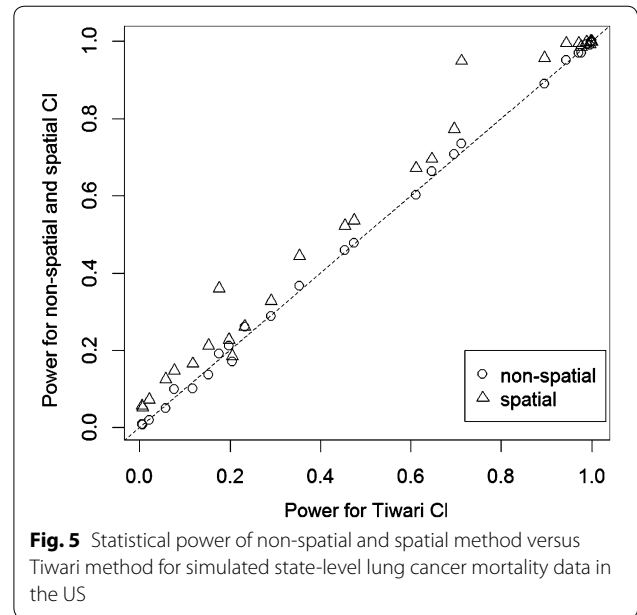
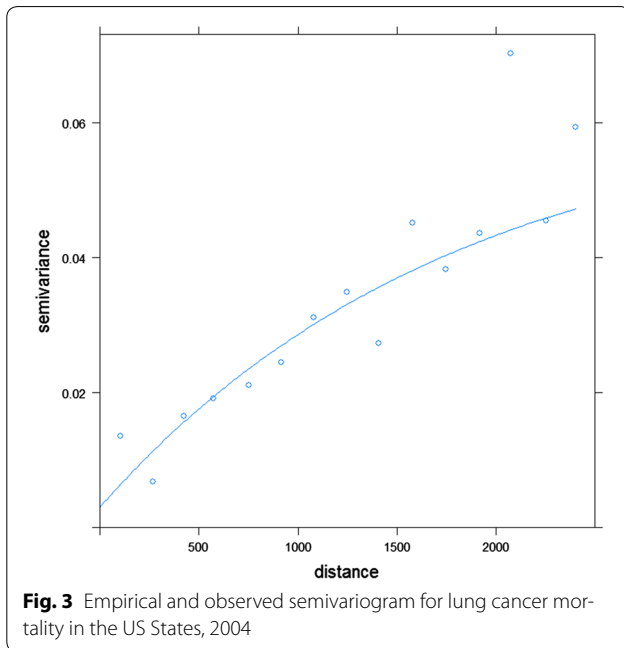
For rate ratios between a state and the US, the same data as in Tiwari et al. [3] were chosen to test the method

developed above. Specifically, age-adjusted mortality rates for tongue, esophagus, and lung cancer of the 50 states and the District of Columbia in year 2004 [16] were obtained and the ratios between the state rates and the US rate were calculated in SEER*Stat software. These three cancer sites are selected to represent the spectrum of cancer incidence, from rare cancer (age-adjusted mortality rate of 0.62 for tongue cancer in the US), to moderate (age-adjusted mortality rate of 4.35 for esophagus), and to common cancer (age-adjusted mortality rate of 53.30 for lung). The data also represent different levels of spatial autocorrelation in the state-level rate ratios (Fig. 2). Tongue cancer and esophagus cancer do not show a spatial pattern, but lung cancer has a clear spatial autocorrelation pattern. To estimate the variance–covariance structure between states, the observed semivariogram values (points in Fig. 3) for lung cancer mortality were modeled and plotted against the estimated values (curve in Fig. 3) using an exponential model. The exponential model produced an estimate of 0.06 for the *partial sill* and 0.065 for the *sill*, which implied that about 92% of the variation in the data can be attributed to spatial autocorrelation. The empirical range estimate was 1792 km, about one-third of the maximum state-to-state distance in the US. The semivariogram plot was cut off at half of the maximum distance, about 2500 km, since values beyond the range are uninformative. The parameter estimates from these models were used for the calculation of covariance matrix. Simulated data were created using the observed rate ratios as the mean and the estimated covariance matrix for each cancer site. Lengths of 95% CI and statistical power are compared between the Tiwari normal-based method, non-spatial and spatial method developed according to formulas (7)–(9). In the esophagus cancer data, the *partial sill* to *sill* ratio was 97% and the empirical range was only 98 km. The tongue cancer data had nugget effect = 0, and the empirical range was 104 km. Both situations indicate that states further than 100 km apart are basically considered spatially unrelated. We varied the strength and range in simulation studies, and the resulting data did not show any detectable spatial correlation for either cancer site, most likely due to the small number of cases.

Since lung cancer data show a clear spatial pattern, we focus on the results from lung cancer mortality State-to-US rate ratios in this sub-section. Figures 4 and 5 compare the Tiwari method, non-spatial and spatial methods in terms of the length of 95% CI for rate ratios and statistical power for the simulated state-level lung cancer data. It can be seen that the non-spatial method is very close to the Tiwari method in both plots. The spatial method is better than the Tiwari or non-spatial methods in that it provides shorter 95% CI and hence more accurate estimates, as well as higher statistical power. California is an



anomaly in that adding correlation from both sources increases its variance estimate and resulting confidence interval length (see Fig. 6). The population in California accounts for about 12% of the US population, so the correlation between the state and US rate due to the overlap is much larger than the spatial autocorrelation. However,



due to the large population size, the interval estimates for the rates (and rate ratios) were already very accurate and the statistical power was high, so the revised interval estimates and power were virtually unchanged.

A linked micromap plot [17] reveals the spatial pattern and associations between multiple variables simultaneously. To explore factors that contribute to the higher precision (and shorter confidence intervals) associated

with the spatial method, we used a linked micromap plot (Fig. 6) to show the percent of length reduction in the 95% CI of lung cancer rate ratios from the Tiwari method to the spatial method in the US, expressed as percentages of the length of the Tiwari 95% CI for each state. The panels in the plot (from left to the right) are the maps of states with the highest to lowest % Reduction (from top to bottom row), the state names, the values of the % Reduction, the population size in each state and the estimated rate ratio with its 95% CI using the spatial method. The region with the highest % Reduction is in the mid-west, and expands to the east, west, and south, and finally to New England, Florida, and the Pacific West. Except for the few states with really large populations (e.g., New York, Texas, Florida, and California), there is a positive correlation between the population size and the % Reduction. In other words, the higher a state's population size, the larger benefit it gains from the spatial method. Many of the larger states have such precise rate ratio estimates that the CI line is completely covered by the estimate's dot. The only state that has a larger confidence interval using the spatial method is California, due to a relatively large proportion of the population of the state of the US total and hence a large overlap, as explained above. By adding the spatial correlation, the variance of the logarithm of the US rate, $\text{var}(\ln(R))$, increases from $6.4E-6$ to $1.0E-4$, a 15-fold change.

For esophagus and tongue cancer data, spatial autocorrelation is weak and the advantage of the spatial method is minimal. The results of the three measures are very close across the Tiwari, non-spatial, and spatial methods for esophagus and tongue cancer (data not shown).

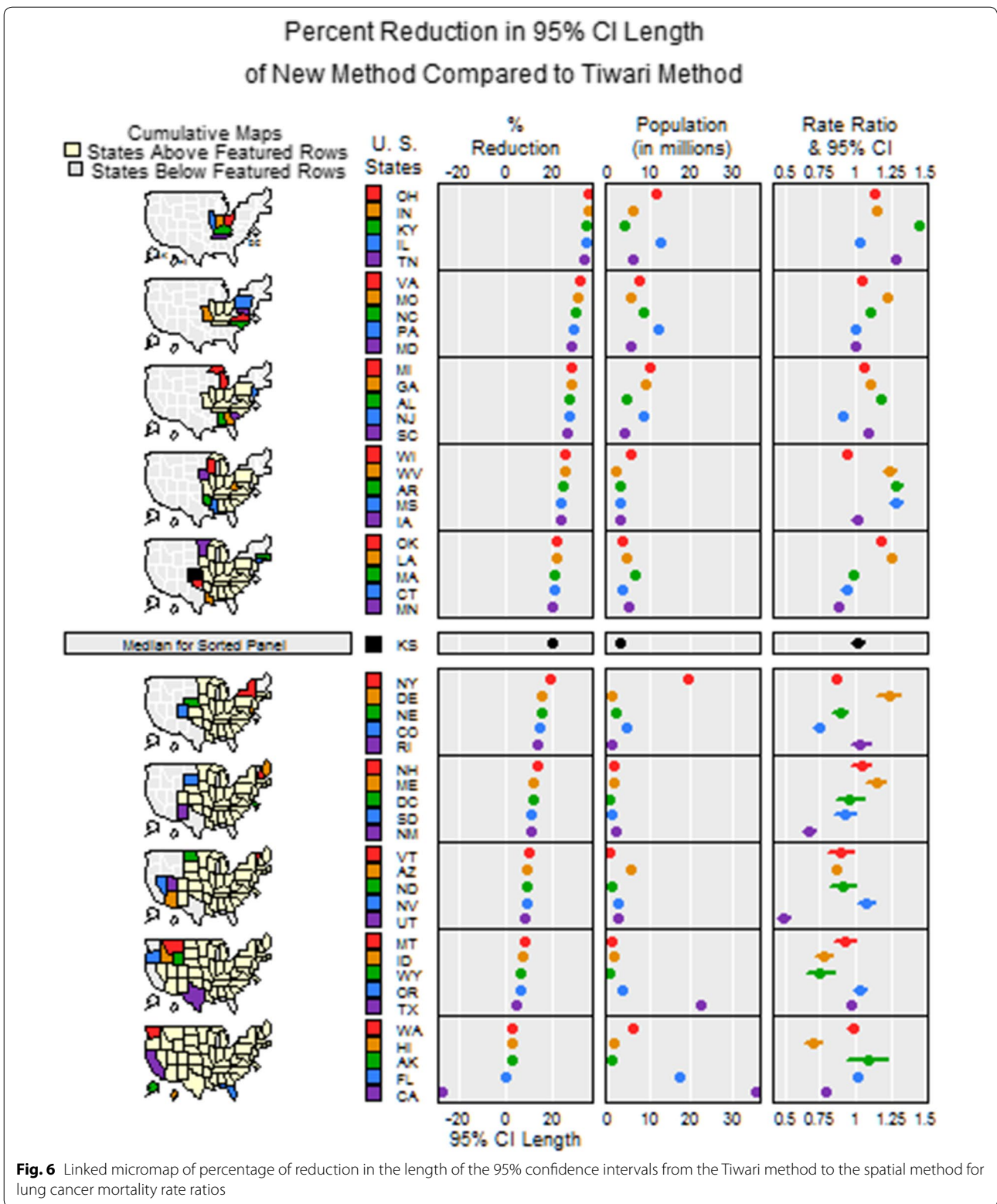


Fig. 6 Linked micromap of percentage of reduction in the length of the 95% confidence intervals from the Tiwari method to the spatial method for lung cancer mortality rate ratios

State-level data with varying autocorrelation strength and scale

To further explore the impact of spatial autocorrelation on the confidence intervals of rate ratios, we applied the competing methods to a set of simulated state-level data. The observed rate ratios of 2004 lung cancer mortality were taken as the mean, and the variance–covariance matrix was set with a variety of autocorrelation strength and scale. As shown in formula (6), for an exponential semivariogram model, spatial autocorrelation depends on the proportion of partial sill to sill, $\frac{c_e}{c_0+c_e}$, as well as the range, a_e . The partial sill to sill proportion ranges between 0 and 1, with a higher value representing stronger spatial autocorrelation. The range is the distance between two regions beyond which the observations are considered spatially uncorrelated. Of the 50 US states and District of Columbia, the pairwise distance ranges from the minimum at 25.7 km (between District of Columbia and Maryland) to the median at 1608.0 km (between Florida and Oklahoma), and the maximum at 4984.0 km (between Maine and Hawaii). We set the range a_e in formula (6) at three different values, 500, 1700, and 4000 km, to represent local, regional, and global scale of spatial autocorrelation. The partial sill to sill proportion, $\frac{c_e}{c_0+c_e}$, was set at 0.1, 0.5, and 0.9, to represent weak, moderate, and strong spatial autocorrelation. A total of nine simulated datasets were created, and percent of length reduction from non-spatial method to spatial method was calculated for each state and averaged across the whole US.

Since the non-spatial method turns out to be very close to the Tiwari method, this section only compares the spatial method to the non-spatial method. Table 1 presents the percent of length reduction for the 95% CI from the non-spatial method to the spatial method, averaged

Table 1 Percent of length reduction for 95% CI for rate ratios, from non-spatial method to spatial method with varying spatial autocorrelation scale (measured by range in semivariogram) and strength (measured by partial sill to sill proportion in semivariogram)

	Scale of spatial autocorrelation measured by range (km)		
	500 (local)	1700 (regional)	4000 (global)
Strength of spatial autocorrelation			
0.1 (weak)	(0, 0.004, 0.095) 0.48	(0.005, 0.039, 0.098) 1.7	(0.029, 0.067, 0.099) 2.7
0.5 (moderate)	(0, 0.02, 0.47) 2.4	(0.027, 0.19, 0.49) 9.2	(0.14, 0.33, 0.50) 14.6
0.9 (strong)	(0, 0.036, 0.85) 4.5	(0.048, 0.35, 0.89) 17.7	(0.26, 0.60, 0.89) 29.3

Numbers in parentheses are the minimum (between Maine and Hawaii), median (between Florida and Oklahoma), and maximum (between Maryland and District of Columbia) pairwise correlation coefficients between regions

across the 50 US states and District of Columbia, for varying scale and strength of spatial autocorrelation. At the local scale, when spatial autocorrelation strength increases from weak to strong, on average, the spatial method reduces the length of 95% CI by 0.48–4.5%, which is minimal. At the regional scale, the range of the semivariogram is 1700 km, which is about the average pairwise distance between US states. In other words, at the regional spatial autocorrelation scale, every state is related to about half of all other states. When the strength of the correlation increases from weak to strong, the spatial methods reduces the length of 95% CI by 1.7–17.7%. If the autocorrelation is on a global scale when every state is related to all other state, spatial methods reduces the length of 95% CI by 2.7–29.3%.

County-level data

The state of Kentucky has the highest cancer rates for both incidence and mortality. Cigarette smoking prevalence is high and cancer screening rate is low, especially in the southeast area of the state which is part of the Central Appalachia region [18]. We calculated the age-adjusted incidence rates of lung cancer for males for the 5-year period between 2006 and 2010. The state rate is 129.94 per 100,000 and the rates vary considerably among the 120 counties, from 57.01 to 207.21, resulting in the county-to-state rate ratios between 0.45 and 1.63 (Fig. 7). There is also a spatial pattern, with higher rates in the southeast mountain area, and lower rates in the north and central areas.

A simulation study was performed based on the Kentucky county-level male lung cancer incidence rates. The simulation study serves two purposes. First, we would like to establish the relationship among the three interwoven factors—county to state rate ratios, county population

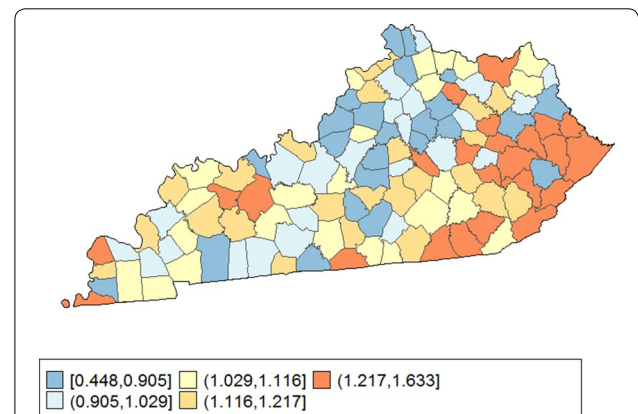


Fig. 7 County-to-State ratio of age-adjusted rates for male lung cancer incidence in Kentucky, 2006–2010

size, and statistical power in testing the hypothesis of rate ratios equal to 1. Then we would compare the non-spatial and spatial methods. To serve the first purpose, we simulated a set of county rates (according to the population size of each county in Kentucky) with the county-to-state rate ratios ranging between 0.2 and 2.0, with an increment of 0.05. This range is broader than that in a realistic situation and should provide a complete picture of the multi-dimensional relationship between rates, rate ratios, population size, and power. However, this is a hypothetical scenario since it is impossible to observe that all counties have the same rate ratio at a certain fixed level, so the analysis in this scenario only helps with understanding the impact of rate ratio and population size on the statistical power. It does not estimate how much improvement the spatial method will provide in a real world situation. To serve the second purpose, we created the simulated data assuming the hypothesized mean rate ratios and the variance-covariance structure were the same as the observed values. Then we compare the statistical power and coverage probability of the 95% CI between the non-spatial and spatial method.

Table 2 lists the statistical power and coverage probability of the non-spatial and spatial methods for the simulated Kentucky incidence data at a few hypothetical fixed rate ratios. The power of both non-spatial and spatial methods decreases from both small and large rate ratios toward 1.0, and power of the spatial method is consistently higher than that of the non-spatial method. The power increase is the largest when the county to state rate ratio is close to 1.0, and diminishes when the rate ratio turns further away from 1.0. For cancer data with a similar level of rates, the county to state rate ratio needs to be smaller than 0.7 or greater than 1.4 to reach a statistical power of 80% (data not shown). The coverage probability of the spatial method is very close to the nominal value of 95%, consistently better than of non-spatial method across all rate ratio values, which is around 97%.

Table 2 Power and coverage probability of the non-spatial and spatial methods in the simulation study for Kentucky county-level male lung cancer incidence data

	Power		Coverage probability	
	Non-spatial (%)	Spatial (%)	Non-spatial (%)	Spatial (%)
Simulation under hypothetical fixed rate ratios				
0.50	98.7	98.9	96.7	95.0
0.75	65.5	71.0	96.7	95.0
1.0	3.23	4.98	96.8	95.0
1.25	48.7	55.3	96.7	94.9
1.50	85.7	88.5	96.8	95.0
1.75	96.3	97.0	96.8	95.0

Figure 8 compares the statistical power of the non-spatial and spatial methods for the same simulated data assuming the hypothesized mean rate ratios and the variance-covariance structure is the same as the observed values. It is clear that the spatial method has a higher power than the non-spatial method in all but Jefferson County, which accounts for about one-fifth of the population in Kentucky. The coverage probability of the spatial method is consistently better than that of the non-spatial method (not shown). Figure 9 reveals the relationship between statistical power vs rate ratio and population size in the simulated situation for the spatial method. Power is smallest when the rate ratio is close to 1.0, and increases as the rate ratio is further from 1.0. The counties with high power (greater than 0.8) have large populations (>70,000) and the rate ratios are either <0.8 or >1.2. Jefferson County's population is the highest in the state, with a 5-year total of 1.76 million, which supports a high statistical power to detect its rate ratio of 0.87 as significantly different from 1.0. Two smaller counties (Monroe County with a 5-year population size of 27,000 and Breathitt County with a 5-year population size of 35,000) have rate ratios close to 1.4; their high rate ratios, rather than population size, drive their statistical power higher.

Discussion

Presenting confidence intervals along with point estimates provides a measure of uncertainty of the estimate. Failure to include spatial autocorrelation when it is present can lead to errors in the estimates and subsequent inferential errors. We have extended a previous method of calculating confidence intervals for the rate ratio of a

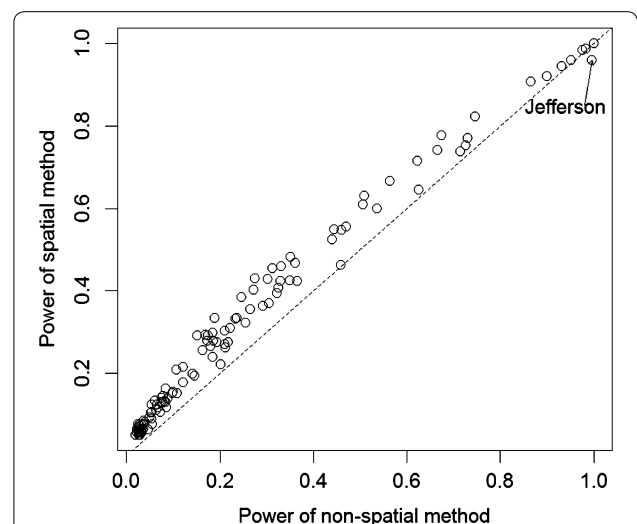
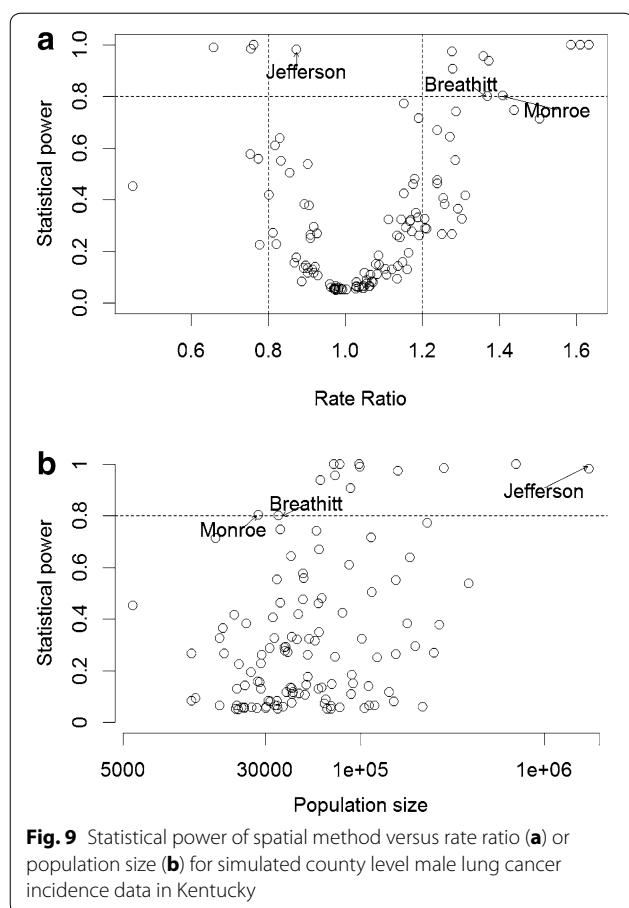


Fig. 8 Statistical power of spatial method versus non-spatial method for simulated county level male lung cancer incidence data in Kentucky



sub-region to the full region to include spatial autocorrelation. Tiwari et al. [3] proposed a method to compute a 95% confidence interval around the ratio of an area-specific rate to the rate for a larger area which includes that area. This method accounts for correlation between the single area rate and that of the larger area, but does not account for likely spatial autocorrelation between rates of neighboring areas. Our method includes components of total variance due to the standard distributional assumption, the correlation between the single area rate and the larger area rate, and spatial autocorrelation among the areas.

The rate ratio is a common measure of the relative ranking of areas, e.g., counties within a state, and is easily computed. The earlier non-overlapping method of Tiwari et al. [2] is now implemented in SEER*Stat, the National Cancer Institute’s popular software to disseminate cancer statistics. Our goals were to improve upon the Tiwari [3] methods by including spatial autocorrelation and to develop a resulting method that could replace the current method in SEER*Stat. Therefore, we used a similar approach as Tiwari to develop our confidence interval

and applied the new method to the same three cancer datasets so as to make the results most comparable.

Several limitations exist in this study. First, isotropy is assumed for spatial autocorrelation. It may be argued that direction as well as distance between regions will have an impact on disease rates. Considering anisotropy will greatly increase the complexity of model and computation. Secondly, the exponential semivariogram model that fit our lung cancer data well may not be the best choice for every outcome variable, although experts believe that the choice of semivariogram model is less important than the inclusion of spatial autocorrelation at all (see p. 379 of [10]).

We were limited in our ability to assess the impact of adding spatial autocorrelation to the rate ratio variance calculations for uncommon and rare cancers like esophagus and tongue cancers. Even when we simulated data with very strong spatial autocorrelation, the resulting data did not show much of a detectable spatial pattern. This is probably due to the small number of cases or the Modifiable Areal Unit Problem (MAUP). Assessing the impact of small numbers or MAUP to spatial autocorrelation is beyond the scope of this paper; our method was developed to account for spatial autocorrelation without regard to how it came about.

Another limitation of our approach is that we have underestimated the uncertainty of the rate ratio variance by assuming no error in the estimation of the covariance parameter estimates. It will be very important to explore approaches that fully account for the uncertainty in the spatial autocorrelation estimation in future work. However, because users are typically not interested in computing a rate ratio for areas with small populations, knowing that the confidence intervals will be large and uninformative in this situation, we suspect that uncertainty in the rate ratio estimate due to the estimation process will be small relative to uncertainty due to unmeasured spatial autocorrelation. Therefore, we believe that while we have shown that our method is superior to the Tiwari et al. method when spatial autocorrelation is present, and equivalent to it when area rates are uncorrelated, any further improvement resulting from a hierarchical (Bayesian) model that can compute the estimation process variation across many replicates will be relatively small. We have initiated a follow-up study to confirm this belief. It is impractical to implement a Bayesian estimation method that typically requires computation of thousands of replicates in a server- or web-based statistical system such as SEER*Stat, and therefore it is important to verify that the method proposed here, one that can easily be implemented, has captured nearly all of the variation in the rate ratio estimate.

Conclusions

We have developed a method that takes into account spatial autocorrelation along with area overlap in calculating variances and confidence intervals of rate ratios between a sub-region and the total region. Our variance is comprised of three components, representing no correlation, area overlap, and spatial autocorrelation, respectively. We have shown that calculating the variance of the rate ratio including the possibility of spatial autocorrelation among the area-specific rates can lead to substantial improvements over the Tiwari method. For U.S. state-level cancer data, confidence intervals were shorter and power was greater than the Tiwari method. The Tiwari method accounted for correlation due to area overlap but not for spatial autocorrelation among the areas.

Application to simulated state-level data showed that the advantage of the proposed method is directly related to the strength and scale of the spatial autocorrelation. Improved results were also seen at the county level using simulated data based on population patterns in Kentucky counties. We did note two instances where results were not as good as in the Tiwari method, both where one region's population constituted a large proportion of the population for the entire aggregated area (California compared to the U.S. and Jefferson County compared to Kentucky). In these cases, the correlation due to the population overlap was much larger than any observed spatial autocorrelation, eliminating any advantage for our method over methods that ignore this source of correlation. Because of the large population size in these regions, though, the variance and interval estimates were already precise, and adding spatial autocorrelation does not practically impact the interval estimates or the statistical power.

One approach to calculating confidence intervals for the rate ratio might be to assess the degree and scale of spatial autocorrelation among the areas and then decide on whether or not to include the spatial autocorrelation. However, these correlations are on a continuum, so it would be difficult to set a cutpoint beyond which the spatial method should be used. We have shown that our method provides the same results as the Tiwari method when there is very little spatial autocorrelation, and it does account for overlap between a sub-region and its parent region. Thus for little additional computational cost, we obtain an estimate equal to that currently used for SEER data if area-specific rates are independent, but in the presence of increasing spatial autocorrelation we can obtain substantial improvements over the existing method. Since the calculations are relatively easy to implement, we recommend using the new method in all cases.

Abbreviation

CI: confidence interval.

Authors' contributions

LZ, and LP developed the method and study design. JBP developed the computing code for the analysis and mapping. All authors read and approved the final manuscript.

Author details

¹ Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, 9609 Medical Center Dr., Suite 4E346, Rockville, MD 20850, USA. ² StatNet Consulting, LLC, Gaithersburg, MD 20882, USA.

Acknowledgements

The authors acknowledge colleagues Eric J. Feuer and Benmei Liu who have provided comments that improved the quality of this manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and material

The software along with datasets supporting the conclusions of this article are included as supporting materials.

Funding

This work was partly supported by the Statistical Methodology and Applications Branch of the National Cancer Institute through a contract HHSN261201400193P.

Received: 1 September 2016 Accepted: 30 November 2016

Published online: 15 December 2016

References

- Tobler W. A computer movie simulating urban growth in the Detroit region. *Econ Geogr.* 1970;46(2):234–40.
- Tiwari RC, Clegg LX, Zou Z. Efficient interval estimation for age-adjusted cancer rates. *Stat Methods Med Res.* 2006;15(6):547–69.
- Tiwari RC, Li Y, Zou Z. Interval estimation for ratios of correlated age-adjusted rates. *J Data Sci.* 2010;8:471–82.
- Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. *Stat Med.* 1997;16(7):791–801.
- Kulldorff M, et al. Breast cancer clusters in the northeast United States: a geographic analysis. *Am J Epidemiol.* 1997;146(2):161–70.
- Pickle L, et al. Atlas of United States mortality. Hyattsville: National Center for Health Statistics; 1996.
- National Cancer Institute. Surveillance Epidemiology and End Results (SEER) Program. 2016. <http://seer.cancer.gov/>.
- National Cancer Institute. SEER*Stat Software. 2016. <http://seer.cancer.gov/seerstat/>.
- Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. New York: Wiley; 2003.
- Waller LA, Gotway CA. Applied spatial statistics for public health data. Hoboken: Wiley; 2004.
- Bishop YM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. Cambridge: MIT Press; 1975.
- Centers for Disease Control and Prevention. Mortality data. <http://www.cdc.gov/nchs/deaths.htm>.
- Zhu L, et al. Predicting US- and state-level cancer counts for the current calendar year. *Cancer.* 2012;118(4):1100–9.
- Furrer R, Sain SR. spam: a sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *J Stat Softw.* 2010;36(10):1–25.
- R Development Core Team. The R Project for statistical computing. 2016. <https://www.r-project.org/>.

16. Surveillance Epidemiology and End Results (SEER) Program, Mortality—All COD. Aggregated with State, Total U.S. (1969–2004). N.C.f.H.S. (NCHS), editor. Bethesda: National Cancer Institute SEER Program; 2007.
17. Pickle LW, Pearson JB, Carr DB. micromapST: exploring and communicating geospatial patterns in U.S. state data. *J Stat Softw.* 2015;63(2):1–25.
18. Christian WJ, et al. Exploring geographic variation in lung cancer incidence in Kentucky using a spatial scan statistic: elevated risk in the Appalachian coal-mining region. *Public Health Rep.* 2011;126(6):789–96.
19. Samui P, Sitharam TG. Application of geostatistical models for estimating spatial variability of rock depth. *Engineering.* 2011;3:886–94.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

