Identification of high-risk regions for schistosomiasis in the Guichi region of China: an adaptive kernel density estimation-based approach

ZHI-JIE ZHANG[1,2,3]*, TILMAN M. DAVIES[4], JIE GAO[1,2,3], ZENGLIANG WANG[1,2,3] *and* QING-WU JIANG[1,2]

[1] *Department of Epidemiology, School of Public Health, Fudan University, Shanghai 200032, People's Republic of China*
[2] *Key Laboratory of Public Health Safety, Ministry of Education, Shanghai 200032, People's Republic of China*
[3] *Laboratory for Spatial Analysis and Modelling, School of Public Health, Fudan University, Shanghai 200032, People's Republic of China*
[4] *Department of Statistics, Institute of Fundamental Sciences, Massey University, Private Bag 11222, Palmerston North, New Zealand*

(*Received 20 August 2012; revised 25 December 2012 and 4 January 2013; accepted 8 January 2013; first published online 7 March 2013*)

SUMMARY

Identification of high-risk regions of schistosomiasis is important for rational resource allocation and effective control strategies. We conducted the first study to apply the newly developed method of adaptive kernel density estimation (KDE)-based spatial relative risk function (sRRF) to detect the high-risk regions of schistosomiasis in the Guichi region of China and compared it with the fixed KDE-based sRRF. We found that the adaptive KDE-based sRRF had a better ability to depict the heterogeneity of risk regions, but was more sensitive to altering the user-defined smoothing parameters. Specifically, the impact of bandwidths on the estimated risk value and risk significance (*P* value) was higher for the adaptive KDE-based sRRF, but lower on the estimated risk variation standard error (S.E.) compared with the fixed KDE-based sRRF. Based on this application the adaptive and fixed KDE-based sRRF have their respective advantages and disadvantages and the joint application of the two approaches can warrant the best possible identification of high-risk subregions of diseases.

Key words: schistosomiasis, kernel density estimation, spatial analysis, spatial epidemiology.

INTRODUCTION

Schistosomiasis japonica is a snail-transmitted parasitic disease that has existed in mainland China for over 2000 years (Zhang *et al*. 2009*b*; Zhou *et al*. 2010). It remains a major public health problem in China (Peng *et al*. 2010; Zhou *et al*. 2010) and the total number of infected people in 2004 was still around 726 000 (Zhou *et al*. 2007). The extensive habitat of *Oncomelania hupensis* (the sole intermediate host of *Schistosoma japonicum*), decreased compliance rate, and the unsustainable effects of chemotherapy, reduced financial supports, frequent floods and climate change have resulted in the rebound of epidemics in many areas of China, even in the places where it met the criteria of transmission interruption (Zhang *et al*. 2008, 2009*a*). Hence, the importance of a sustainable control strategy for schistosomiasis has been frequently emphasized (Utzinger *et al*. 2003, 2009), which is a great challenge for China's current schistosomiasis control. The identification of high-risk regions of schistosomiasis is always an important first step for an effective and sustainable strategy (Zhang *et al*. 2009*a*, *b*) and many spatial statistical methods have been explored with this objective in mind. Among others, the approach of kernel density estimation (KDE) has attracted much attention because of its minimal assumptions on the underlying data structure and flexibility of application (Bithell, 1990, 1991; Kelsall and Diggle, 1995*a*, *b*). As an example, Galvao and coworkers used the technique of KDE to compare the spatial pattern of *Schistosoma mansoni* before and after treatment with different doses of praziquantel (Galvao *et al*. 2010). However, understanding the spatial variation of disease risk per se requires the researcher(s) to not only examine the spatial distribution of disease 'cases', but also the distribution of the at-risk individuals (the 'controls') in order to adjust for any natural non-homogeneity in the underlying population (Zhang *et al*. 2009*a*; Davies and Hazelton, 2010). This motivated the development of the KDE-based spatial relative risk function (sRRF) for spatial case-control designed studies (Bithell, 1990, 1991). Many successful applications in spatial epidemiology

* Corresponding author: Department of Epidemiology, School of Public Health, Fudan University, Shanghai 200032, People's Republic of China. Tel: +86 21 54237410. Fax: +86 21 54237410. E-mail: epistat@gmail.com

Parasitology (2013), **140**, 868–875. © Cambridge University Press 2013
doi:10.1017/S0031182013000048

of KDE-based sRRF have been reported, e.g. motor neurone disease (Sabel *et al*. 2000), biliary cirrhosis (Prince *et al*. 2001), childhood leukaemia (Wheeler, 2007) and Aujeszky's disease (Berke and Grosse Beilage, 2003). In the field of schistosomiasis, Zhang *et al*. were the first to apply the KDE-based sRRF to assess the risk of *S. japonica* in the Guichi region of Anhui province in China (Zhang *et al*. 2009*a*).

Historically, most implementations of the sRRF have made use of a *fixed* bandwidth or smoothing parameter in kernel estimation of the densities, where the amount of smoothing applied to the estimator is constant, regardless of location. It is generally well understood that spatial distributions of human populations tend to be quite heterogeneous due to common geographical features (towns, rivers, etc.). This spatial variation means that it is worth utilizing a potentially beneficial *adaptive* approach. One such adaptive KDE approach, first discussed in depth by Abramson (1982), assigns less smoothing to densely clustered observations in order to preserve spatial detail where there is an abundance of data, and allocates more smoothing to isolated observations in an effort to avoid assigning undue density 'height' to areas where we do not have as much information. The adaptive KDE-based sRRF was first investigated by Davies and Hazelton (2010), who found both theoretical and practical advantages over the fixed KDE-based sRRF, and has found a successful application in a study about subclinical *Salmonella* infection in finisher pig herds (Benschop *et al*. 2008).

The present study aims to apply this novel approach of adaptive KDE-based sRRF to detect high-risk regions of schistosomiasis in the Guichi region of China, and compare it with the method of fixed KDE-based sRRF. Generally, the obtained results can shed lighton rational applications of adaptive KDE-based sRRF to identify disease-risk regions. Specifically, the results can aid in the design of more efficient schistosomiasis control strategies in the local region.

MATERIALS AND METHODS

Data sources

The study site was the Guichi region of Anhui province in eastern China and the study design was spatial case-control. Acute schistosomiasis cases from permanent residents between 1 January 2001 and 31 December 2006 were retrospectively collected from local schistosomiasis-specific hospitals and village-level clinics. The same number of controls was randomly chosen to represent the background at-risk population pattern using the sampling approach of the probability proportion to size. All the spatial coordinates of cases and controls were first obtained in the field using the hand-held global positioning system (GPS) (MobileMapper, Thales Navigation, Inc., USA) and then the spatial analysis database for schistosomiasis was created using the ArcGIS9.2 software (Environmental Systems Research Institute, Redlands, CA, USA). See our previous reports for detailed descriptions of the study area and the database (Zhang *et al*. 2008, 2009*a*).

Statistical analysis

Let x_1, x_2,…, x_{n1} denote the coordinates of n_1 schistosomiasis cases in Guichi region. The (bivariate) kernel density estimate thereof is written as (Davies and Hazelton, 2010),

$$\hat{f}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{h(x_i)^2}K\left(\frac{x-x_i}{x_i}\right) \tag{1}$$

where K is a radially symmetrical probability density function (the *kernel*) and $h(\cdot)$ is the bandwidth or smoothing parameter controlling the smoothness of the density estimator.

The bivariate Gaussian kernel is implemented for the current analysis, as the infinite tails of this function are useful in areas with sparse data. For the fixed-bandwidth approach, $h(\cdot) = h_{\text{fix}}$ (i.e. simply a scalar constant). We defined the bandwidth function of the adaptive kernel estimator as (Terrell, 1990; Davies and Hazelton, 2010),

$$h(u) = h_0\left\{ f(u)^{1/2}\left(\prod_{i=1}^{n_1} f(x_i)^{-1/2}\right)^{1/n_1}\right\}^{-1} \tag{2}$$

where h_0 is the *global bandwidth*. In practice we must replace the unknown density function f in equation (2) with a pilot estimate $\bar{f}$, which is itself a fixed-bandwidth kernel estimate of the observed data with smoothing parameter $h_{\text{fix}} = \bar{h}$; referred to as the *pilot bandwidth*.

Now let y_1, y_2,…, y_{n2} denote the coordinates of n_2 sampled controls. Conditional on the sample sizes of n_1 cases and n_2 controls, the sRRF is defined as the density ratio of cases and controls (Bithell, 1990),

$$\hat{r}(x) = \log\frac{\hat{f}(x)}{\hat{g}(x)} \tag{3}$$

where, $\hat{f}$ and $\hat{g}$ are the kernel estimates from equation (1) of the case and control data, respectively (either fixed or adaptive). The density ratio is transformed to the log scale in order to make symmetrical the treatment of the two estimates and stabilize numerical results (Kelsall and Diggle, 1995*a*, *b*). Owing to the fact that the data have been collected with respect to a finite geographical region, we also employ edge-correction techniques described by Diggle (1985) (fixed) and Marshall and Hazelton (2010) (adaptive) for $\hat{f}$ and $\hat{g}$ to reduce the boundary biases.

In terms of assigning the bandwidths, we use a common fixed bandwidth (i.e. $h_{\text{fix},(\hat{f})} = h_{\text{fix},(\hat{g})} = h_{\text{fix}}$) for the fixed KDE-based sRRF, and a common

https://doi.org/10.1017/S0031182013000048 Published online by Cambridge University Press

global bandwidth $h_{0,(\hat{f})} = h_{0,(\hat{g})} = h_0$ for the adaptive KDE-based sRRF. This is due to a resulting first-order bias cancellation in areas where $f \cong g$ (Kelsall and Diggle, 1995*a*). The pilot bandwidths in the adaptive KDE-based sRRF, however, are computed separately for the case and control data in order to assist in preserving any specific detail of use for the pilot densities and variable bandwidth calculations.

In the examination of the adaptive KDE-based sRRF, Davies and Hazelton (2010) made use of 2 data-driven bandwidth selection methods for estimation of the fixed, global and pilot bandwidths. The first, based on an 'oversmoothing' principle described by Terrell (1990), was elected due to its potential to control excess variability in the estimated densities. We refer to it as 'OS'. The second is the least-squares cross-validation (LSCV) approach as described by Bowman and Azzalini (1997). We repeated these calculation methods for analysis of the schistosomiasis data, with $h_{\text{fix}} \leftarrow \text{OS}$ in the fixed KDE-based sRRF (based on the pooled case-control coordinates), and $h_0 \leftarrow \text{OS}$ (again based on the pooled case-control data) and $\bar{h}_{(\hat{f})} \leftarrow \text{LSCV}$, $\bar{h}_{(\hat{g})} \leftarrow \text{LSCV}$ (based on the case and control data, separately) in the adaptive KDE-based sRRF.

The risk functions estimated by equation (3) are simply point estimates of the observed (log-) risks. Naturally, it is of interest to be able to distinguish any statistically significant risk regions from non-significant risk regions. To avoid an over-interpretation of the results, *tolerance contours* based on a pointwise p-value surface (and drawn at a significance level of $\alpha = 0{\cdot}05$) were applied from the statistical test searching for elevated risk (i.e. with respect to the hypotheses $H_0 = r(x) = 0$; $H_1 : r(x) > 0$, where r denotes the 'true' log-risk surface).

Two different approaches were used to obtain the p values in each point. The first method is the Monte-Carlo (MC) randomization test based on permutations of the case-control labels in each point (Kelsall and Diggle, 1995*a*, *b*). First, the case and control location data are pooled; then n_1 points were re-sampled without replacement to represent the simulated cases and the remaining n_2 points were used as the simulated controls. The fixed and adaptive KDE-based sRRF was then repeatedly applied on these simulated datasets. The whole process was replicated 999 times to obtain the simulated risk values $(r_1^*(x), r_2^*(x), \ldots, r_{999}^*(x))$. For each point, we get 1 observed $\hat{r}(x)$ and 999 simulated $r_i^*(x)$, so the p-value at each point based on MC method was obtained by the formula,

$$p = \left(1 + \sum_{i=1}^{999} I(\hat{r}(x) \geq r_i^*(x))\right) / 1000 \qquad (4)$$

where, $I(\ )$ is the indicator function. The second method is the z-test statistic-based asymptotic normality test (ASYN) introduced by Hazelton and Davies (2009) for the fixed KDE-based sRRF and Davies and Hazelton (2010) for the adaptive KDE-based sRRF. In this approach, the authors exploited approximations to the variances of the fixed and adaptive KDE-based sRRF in order to construct test statistics $z(x)$ at each location x. The p-value surface is then computed easily with respect to the aforementioned hypotheses, as under the asymptotic theory of the kernel estimator $z(x) \sim N(0,1)$, where $N(0,1)$ denotes the standard normal distribution. The authors found this technique to be significantly computationally cheaper than the MC method (especially for large datasets), and it also seemed to avoid some instability in the resulting tolerance contours.

To further investigate the variation of estimated risks, the point-wise standard error (S.E.) surfaces was generated for both MC and ASYN approaches, empirically over the iterated results for the former and via the asymptotic variance for the latter.

For the sensitivity analysis, different bandwidths (based on halving and doubling the appropriate OS or LSCV bandwidths) for adaptive and fixed KDE-based sRRF were also used and are displayed in the Supplementary appendix (online version only) as laid out in Table 1. For each of the various bandwidth combinations, MC and ASYN tests were applied to obtain the corresponding p-value and S.E. surfaces for delineating the significant schistosomiasis risk subregions and associated variation.

All computations and images were produced in the R software (Davies *et al*. 2011).

RESULTS

In total, 83 acute schistosomiasis cases were collected and 83 controls were sampled for the schistosomiasis dataset, which is described elsewhere (Zhang *et al*. 2008, 2009*a*). Here, only the results from adaptive KDE-based sRRF and its comparisons with fixed KDE-based sRRF are reported.

Figure 1 shows the estimated values of adaptive and fixed KDE-based sRRF using the automatically determined optimum bandwidths, and the significant tolerance contours from asymptotic normality and MC tests were superimposed upon the density maps. The adaptive KDE-based sRRF estimated higher risks on the high-risk regions (e.g. significant risk regions) and lower risks on the low-risk regions (e.g. nonsignificant risk regions) than the fixed KDE-based sRRF.

For the fixed bandwidth contours, we observed little difference for the contours corresponding to the ASYN and MC methods of computation. Two clear subregions of interest were identified, one on the northern border where the Qiupu River feeds into the Yangtze River, and the other in the southeastern corner. The ASYN and MC contour methods for the adaptive KDE-based sRRF, however, are more distinct. The adaptive-surface ASYN contours

https://doi.org/10.1017/S0031182013000048 Published online by Cambridge University Press

Table 1. Different bandwidths used for the sensitivity analysis

No	Fixed KDE	Adaptive KDE		
		Pilot (case)	Pilot (control)	Global
S1	0·5 × OS	0·5 × LSCV	0·5 × LSCV	0·5 × OS
S2	0·5 × OS	0·5 × LSCV	0·5 × LSCV	2 × OS
S3	0·5 × OS	2 × LSCV	2 × LSCV	0·5 × OS
S4	0·5 × OS	2 × LSCV	2 × LSCV	2 × OS
S5	2 × OS	0·5 × LSCV	0·5 × LSCV	0·5 × OS
S6	2 × OS	0·5 × LSCV	0·5 × LSCV	2 × OS
S7	2 × OS	2 × LSCV	2 × LSCV	0·5 × OS
S8	2 × OS	2 × LSCV	2 × LSCV	2 × OS

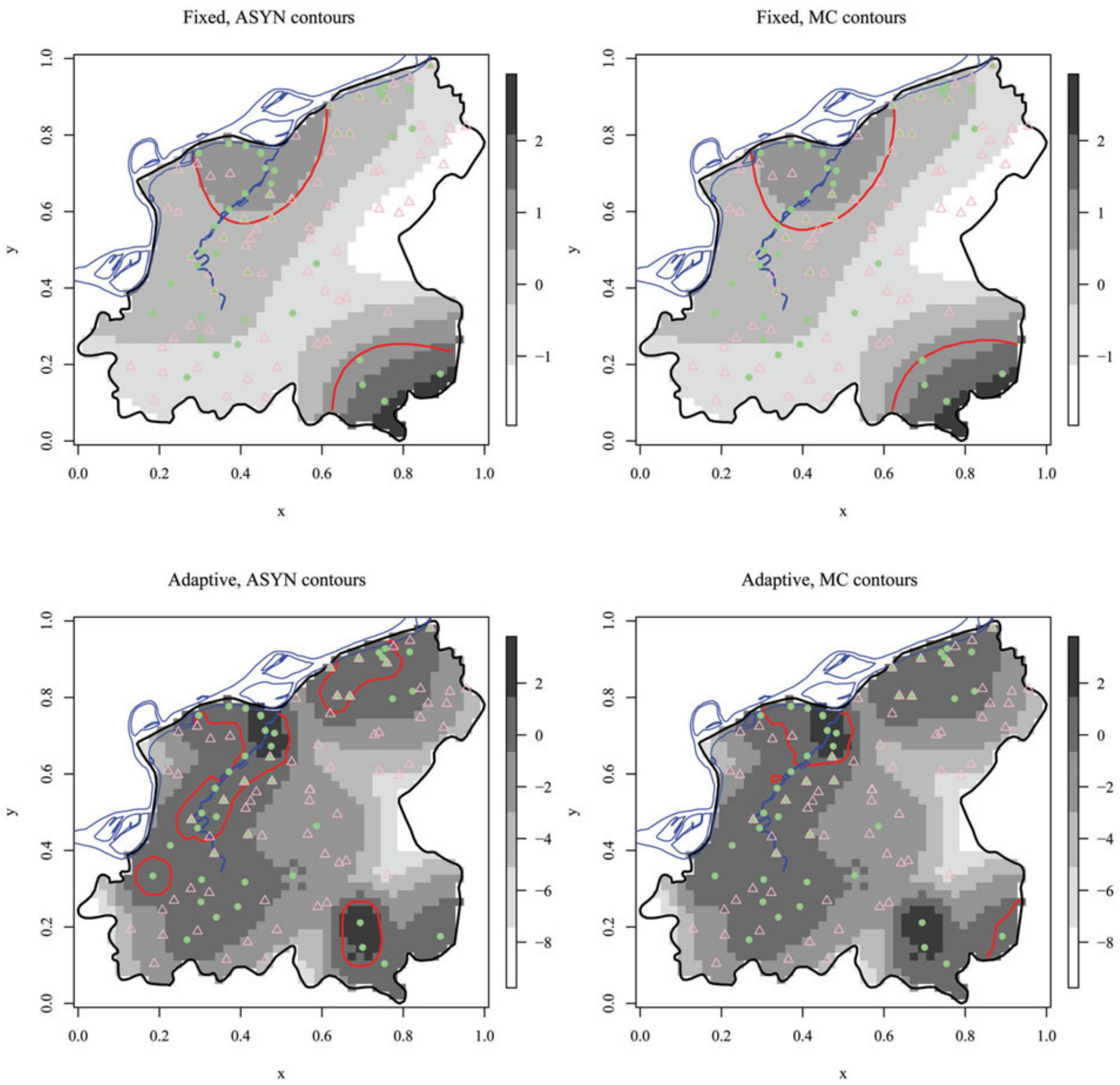

Fig. 1. Spatial relative risk functions of schistosomiasis using fixed and adaptive KDE, and with 5% significant tolerance contours computed using the MC and ASYN methods. The cases and controls are displayed as dots and triangles, respectively.

track the cases falling along the Qiupu River for a considerable distance. In addition to this, the contours seem to highlight 3 other areas. One, in the south-eastern region, approximately matches the location of the southeastern significant region on the fixed KDE-based sRRF plots. The other two, initiatedin one instance by a single case of schistosomiasis, should be interpreted with caution. It must be kept in mind that

https://doi.org/10.1017/S0031182013000048 Published online by Cambridge University Press

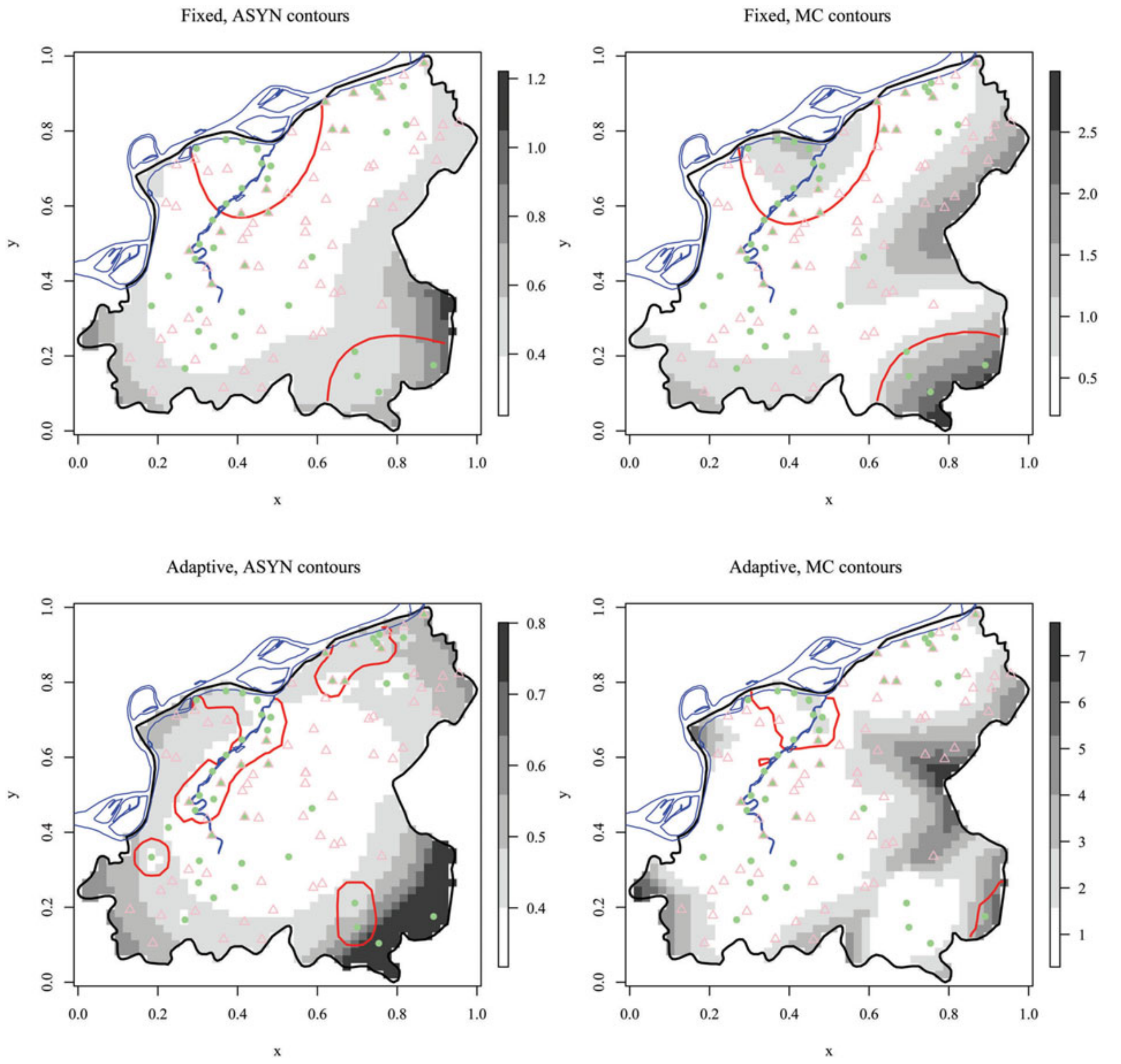

Fig. 2. Point-wise S.E. surfaces of the spatial relative risk functions in Fig. 1 (data and contours contained). MC S.E.S are empirical, ASYN S.E.S are theoretical.

the ASYN method relies on asymptotic properties, and the sample size in this application may need to be increased in order to confirm the existence of the additional 'hotspots'. This view is supported by MC contours for the adaptive KDE-based sRRF – they are less sensitive to isolated case observations for the relatively small sample size.

These interpretations warrant the inspection of the variability observed in the risk surfaces with respect to the tests of significance. Figure 2 depicts the pointwise S.E. surfaces of the sRRF obtained during the tests for significance of risk for the adaptive and fixed KDE-based smoothing approaches. Comparing the fixed and adaptive variation surfaces for the ASYN test, we note a similar pattern in that the lowest variation occurs in the interior of the region, with variation increasing near the border. This can be expected, as edge-correction can affect the performance of the risk estimator in these areas. We also note that the fixed ASYN variation reaches larger values than the adaptive ASYN variation. This can, at least in part, be attributable to the asymptotic variance stabilization for the adaptive KDE-based sRRF; this does not occur for the fixed KDE-based sRRF. This is encouraging, as it indicates that the asymptotic properties seem to be realized to at least a minimal degree for even the relatively small sample size in this application. The fixed and adaptive surface variations for the MC tests are a different story. While in both cases the MC variation is larger than the variation in their ASYN counterparts, here the adaptive MC S.E.S are much larger than the fixed MC values. A possible explanation is that the adaptive KDE-based sRRF generates higher peaks and lower troughs in the case/control density estimates than the fixed KDE-based sRRF, and when the random case/control permutations are performed for the MC tests, it is these higher peaks and lower troughs that drive the variation up in the adaptive KDE-based sRRF.

The results of sensitivity analysis using different bandwidths (as given in Table 1) (images given in the Supplementary appendix, Online version only) show that the fixed and adaptive KDE-based sRRF significance tests were both affected by the bandwidths. Examination of the differences in contour appearance between the bandwidths for the fixed KDE-based sRRF (Supplementary appendix Figs S1–S8 and Figs S9–S16, Online version only) shows that even halving and doubling the OS bandwidth had only a minimal impact on the identified significant regions. The impact of varying the pilot and global bandwidths for the adaptive KDE-based sRRF is more pronounced, although this is to be expected given the small sample size. Overall, the adaptive MC contours were more resistant to bandwidth change than the ASYN contours. Again, the small sample sizes could well be affecting the validity of the asymptotics here. For both adaptive ASYN and MC contours, altering the global bandwidth was more important than changing the pilots (e.g. Supplementary appendix Figs S1–S2 and Figs S3–S4, Online version only). However, in the event of a small global bandwidth for example, large pilots can help stabilize the surface and contours (e.g. comparing the adaptive surfaces in Supplementary appendix Figs S1–S2 and Figs S5–S6, Online version only). The variation of the tests was consistently higher for the MC test compared with the ASYN test for both smoothing regimens. For the ASYN tests only, the S.E.S reached markedly higher values for the fixed surfaces in Supplementary appendix Figs S1–S10 (Online version only); these maximums were only slightly lower for the fixed surfaces in Supplementary appendix Figs S11–S16 (Online version only). This means that even with excessive over-smoothing in the fixed surfaces to minimize variability, the adaptive surfaces (with less smoothing and hence less bias) still appear to provide competitive values of variability.

DISCUSSION

Kernel density estimation is a nonparametric and popular approach to identify high-risk regions of disease (Zhang *et al*. 2009*a*; Davies and Hazelton, 2010). This technique has been explored in the field of schistosomiasis (Zhang *et al*. 2009*a*; Galvao *et al*. 2010). This study was the first to apply the newly developed adaptive KDE-based sRRF to identify and highlight areas of elevated schistosomiasis risk, which is helpful in guiding rational disease control strategies and in allocating resources effectively (Brooker *et al*. 2006; Zhang *et al*. 2009*a*). The results also provide some useful guidance on using this approach effectively in a wider sense.

It is generally accepted that the amount of smoothing (for either fixed or adaptive KDE) is of paramount importance in the 'quality' of our estimate in terms of its proximity to true (unknown) density. While it has been observed that benefits such as reduced bias and variance stabilization for the adaptive version of the relative risk estimator can be realized in practice (Davies and Hazelton, 2010), there are difficulties in terms of now needing to choose multiple 'initial' values in the form of pilot and global bandwidths. Also, as many of the theoretical properties have been examined by large-sample approximations, small sample sizes can have the potential for an obviously detrimental effect on estimate and testing quality. While we note that the adaptive KDE-based sRRF has a better ability than the fixed KDE-based sRRF to depict the heterogeneity of risk regions, due to the apparent instability of the adaptive risk function for small 'case' sample sizes observed in the sensitivity imagery, we do not recommend its use for numerator sample sizes of less than, say, 100. It would be interesting to investigate further, with the help of pre-defined problem scenarios, how varying the pilot and global bandwidth impacts 'performance' statistics such as integrated square errors with respect to spatial relative risk surfaces.

Two methods of testing were used to identify the significant high-risk regions of schistosomiasis. The ASYN test was better than the MC test in terms of generally lower variation for both adaptive and fixed KDE-based sRRF. The result of the ASYN test of adaptive KDE-based sRRF seemed to be more rational considering the detected high-risk regions of schistosomiasis. For example, schistosomiasis is a waterborne disease, so the shape for schistosomiasis risk regions close to the Qiupu River should be along the direction of water flows, which is clearer for the detected Northern risk region of schistosomiasis (see Fig. 1). However, the adaptive surfaces and corresponding tolerance contours appeared more sensitive to varying the originally defined OS and LSCV bandwidths due to the relatively small samples (refer to the imagery in the Supplementary appendix, Online version only). So, we suggested combining the results of the ASYN test from fixed and adaptive KDE-based sRRF to draw a conclusion. The idea that different spatial statistical methods should be used to identify disease-risk regions was previously suggested by Ward and Carpenter (2000) and Zhang *et al*. (2008) for the reason that the spatial pattern of disease risk may be very complicated in reality. We may conclude that 2 common risk regions in the Northern and Southeast parts were 'true' risk regions of Guichi region and effective measures should be taken immediately to control schistosomiasis there; while 2 different risk regions in the Northeast and West could be regarded as 'potential' risk regions of schistosomiasis that should be monitored closely, which is consistent with previous reports (Zhang *et al*. 2008, 2009*a*, *b*).

https://doi.org/10.1017/S0031182013000048 Published online by Cambridge University Press

Adaptive KDE-based sRRF is, in essence, an approach of KDE, so kernel function is not so important based on previous studies (Kelsall and Diggle, 1998; Davies *et al*. 2010) and the conventional bivariate Gaussian kernel was used in this study. But, bandwidth selection and edge effects are two critical issues that were always discussed. A larger bandwidth tends to result in a smoother surface, and thus certain features in the data may not be captured; while a smaller bandwidth will get more local peaks and troughs, which is also not helpful for detecting the spatial variation in risk as a whole (Vieira *et al*. 2002; Zhang *et al*. 2009*a*). Edge effects are caused by the unrecorded cases outside the studied region, so bias is possible and sometimes serious for the places close to the boundaries. Interested readers are encouraged to see the discussions in our previous reports (Zhang *et al*. 2009*a*). Here, we just point out two future research questions related to the adaptive KDE-based sRRF. One question concerns the impact of bandwidths on the adaptive KDE-based sRRF. It has three different bandwidths and is more complicated than the fixed KDE-based sRRF (1 bandwidth), so determining their relative impact on the adaptive KDE-based sRRF is useful for better understanding its performance caused by bandwidths and applying it more effectively. A series of well-designed simulations is needed to illuminate this issue. Another question is about the method used to correct the edge effects of adaptive KDE-based sRRF. This study used the newly developed edge correction method, which is the only available approach that can be used for the adaptive KDE-based sRRF. Some new methods to correct the edge effects for the adaptive KDE-based sRRF need to be developed and an evaluation of the effectiveness of these edge correction methods should be conducted.

In summary, we have conducted the first study applying the adaptive KDE-based sRRF to identify high-risk regions of schistosomiasis, and comparisons of the results between it and the fixed KDE-based sRRF were performed. Our application of adaptive kernel estimation of relative risk and associated significance tests has shed new light not only on the ‘hotspots’ for schistosomiasis in Guichi region, but also where we expect the novel statistical methodology to perform well and where it may struggle. While both fixed and adaptive KDE-based sRRF were shown in this example to possess advantages and disadvantages, we conclude that simultaneous application of the two approaches could warrant the best possible identification of high-risk regions of disease.

ACKNOWLEDGEMENTS

We sincerely thank the staff from Guichi Station of Schistosomiasis Prevention and Control for their help on data collection and field management, without whose cooperation this project would not have been possible.

FINANCIAL SUPPORT

Z.J.Z and Q.W.J. were supported by the Important National Science and Technology Major Programme (grant numbers: 2012ZX10004-220 and 2008ZX10004-011), the National Natural Science Foundation of China (grant numbers: 81102167 and 81172609), Specialized Research Fund for the Doctoral Programme of Higher Education, SRFDP (grant number: 20110071120040) and A Foundation for the Author of National Excellent Doctoral Dissertation of PR China (FANEDD) (201186). T. M. D. was funded by a Top Achiever's Doctoral Research Scholarship (New Zealand Tertiary Education Commission).

REFERENCES

Abramson, I. S. (1982). On bandwidth estimation in kernel estimates – a square root law. *Annals of Statistics* **10**, 1217–1223.

Benschop, J., Hazelton, M. L., Stevenson, M. A., Dahl, J., Morris, R. S. and French, N. P. (2008). Descriptive spatial epidemiology of subclinical Salmonella infection in finisher pig herds: application of a novel method of spatially adaptive smoothing. *Veterinary Research* **39**, 2.

Berke, O. and Grosse Beilage, E. (2003). Spatial relative risk mapping of pseudorabies-seropositive pig herds in an animal-dense region. *Journal of Veterinary Medicine Series B – Infectious Diseases and Veterinary Public Health* **50**, 322–325.

Bithell, J. F. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine* **9**, 691–701.

Bithell, J. F. (1991). Estimation of relative risk functions. *Statistics in Medicine* **10**, 1745–1751.

Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis-The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Brooker, S., Leslie, T., Kolaczinski, K., Mohsen, E., Mehboob, N., Saleheen, S., Khudonazarov, J., Freeman, T., Clements, A., Rowland, M. and Kolaczinski, J. (2006). Spatial epidemiology of *Plasmodium vivax*, Afghanistan. *Emerging Infectious Diseases* **12**, 1600–1602.

Davies, T. M. and Hazelton, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine* **29**, 2423–2437.

Davies, T. M., Hazelton, M. L. and Marshall, J. C. (2011). Sparr: analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. *Journal of Statistical Software* **39**, 1–14.

Diggle, P. (1985). A kernel-method for smoothing point process data. *Applied Statistics – Journal of the Royal Statistical Society Series C* **34**, 138–147.

Galvao, A. F., Favre, T. C., Guimaraes, R. J., Pereira, A. P., Zani, L. C., Felipe, K. T., Domingues, A. L., Carvalho, O. S., Barbosa, C. S. and Pieri, O. S. (2010). Spatial distribution of *Schistosoma mansoni* infection before and after chemotherapy with two praziquantel doses in a community of Pernambuco, Brazil. *Memórias do Instituto Oswaldo Cruz* **105**, 555–562.

Hazelton, M. L. and Davies, T. M. (2009). Inference based on kernel estimates of the relative risk function in geographical epidemiology. *Biometrical Journal* **51**, 98–109.

Kelsall, J. E. and Diggle, P. J. (1995*a*). Kernel estimation of relative risk. *Bernoulli* **1**, 3–16.

Kelsall, J. E. and Diggle, P. J. (1995*b*). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine* **14**, 2335–2342.

Kelsall, J. E. and Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics* **47**, 559–573.

Marshall, J. C. and Hazelton, M. L. (2010). Boundary kernels for adaptive density estimators on regions with irregular boundaries. *Journal of Multivariate Analysis* **101**, 949–963.

Peng, W. X., Tao, B., Clements, A., Jiang, Q. L., Zhang, Z. J., Zhou, Y. B. and Jiang, Q. W. (2010). Identifying high-risk areas of schistosomiasis and associated risk factors in the Poyang Lake region, China. *Parasitology* **137**, 1099–1107.

Prince, M. I., Chetwynd, A., Diggle, P., Jarner, M., Metcalf, J. V. and James, O. F. (2001). The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology* **34**, 1083–1088.

Sabel, C. E., Gatrell, A. C., Loytonen, M., Maasilta, P. and Jokelainen, M. (2000). Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. *Social Science and Medicine*, **50**, 1121–1137.

https://doi.org/10.1017/S0031182013000048 Published online by Cambridge University Press

Terrell, G. R. (1990). The maximal smoothing principle in density-estimation. *Journal of the American Statistical Association* **85**, 470–477.

Utzinger, J., Bergquist, R., Shu-Hua, X., Singer, B. H. and Tanner, M. (2003). Sustainable schistosomiasis control – the way forward. *Lancet* **362**, 1932–1934.

Utzinger, J., Raso, G., Brooker, S., De Savigny, D., Tanner, M., Ornbjerg, N., Singer, B. H. and N'Goran, E. K. (2009). Schistosomiasis and neglected tropical diseases: towards integrated and sustainable control and a word of caution. *Parasitology* **136**, 1859–1874.

Vieira, V., Webster, T., Aschengrau, A. and Ozonoff, D. (2002). A method for spatial analysis of risk in a population-based case-control study. *International Journal of Hygiene and Environmental Health* **205**, 115–120.

Ward, M. P. and Carpenter, T. E. (2000). Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. *Preventive Veterinary Medicine* **45**, 257–284.

Wheeler, D. C. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996–2003. *International Journal of Health Geographics* **6**, 13.

Zhang, Z., Carpenter, T. E., Chen, Y., Clark, A. B., Lynn, H. S., Peng, W., Zhou, Y., Zhao, G. and Jiang, Q. (2008). Identifying high-risk regions for schistosomiasis in Guichi, China: a spatial analysis. *Acta Tropica* **107**, 217–223.

Zhang, Z., Clark, A. B., Bivand, R., Chen, Y., Carpenter, T. E., Peng, W., Zhou, Y., Zhao, G. and Jiang, Q. (2009*a*). Nonparametric spatial analysis to detect high-risk regions for schistosomiasis in Guichi, China. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **103**, 1045–1052.

Zhang, Z. J., Carpenter, T. E., Lynn, H. S., Chen, Y., Bivand, R., Clark, A. B., Hui, F. M., Peng, W. X., Zhou, Y. B., Zhao, G. M. and Jiang, Q. W. (2009*b*). Location of active transmission sites of *Schistosoma japonicum* in lake and marshland regions in China. *Parasitology* **136**, 737–746.

Zhou, X. N., Guo, J. G., Wu, X. H., Jiang, Q. W., Zheng, J., Dang, H., Wang, X. H., Xu, J., Zhu, H. Q., Wu, G. L., Li, Y. S., Xu, X. J., Chen, H. G., Wang, T. P., Zhu, Y. C., Qiu, D. C., Dong, X. Q., Zhao, G. M., Zhang, S. J., Zhao, N. Q., Xia, G., Wang, L. Y., Zhang, S. Q., Lin, D. D., Chen, M. G. and Hao, Y. (2007). Epidemiology of schistosomiasis in the People's Republic of China, 2004. *Emerging Infectious Diseases* **13**, 1470–1476.

Zhou, X. N., Bergquist, R., Leonardo, L., Yang, G. J., Yang, K., Sudomo, M. and Olveda, R. (2010). *Schistosomiasis japonica* control and research needs. *Advances in Parasitology* **72**, 145–178.

https://doi.org/10.1017/S0031182013000048 Published online by Cambridge University Press