



# Disease mapping of *tuberculosis* prevalence in Eastern Cape Province, South Africa

Davies Obaromi<sup>1</sup> · James Ndege<sup>1</sup> · Qin Yongsong<sup>1</sup>

Received: 12 February 2018 / Accepted: 18 May 2018 / Published online: 12 June 2018  
© The Author(s) 2018

## Abstract

**Aim** This is an applied study to investigate the association of selected socio-economic and demographic factors with the relative risk of tuberculosis (TB) prevalence in the Eastern Cape Province of South Africa and to produce disease maps for the spatial outlines of the disease in the province.

**Subjects and methods** This is an ecological spatial study of TB prevalence in the Eastern Cape, a province in South Africa, during the year 2014. Three socio-economic indicators and three demographic factors, all calculated per sub-district, were used to assess their relationship with tuberculosis prevalence, using a Poisson regression model.

**Results** From the analysis, the best model included all the selected covariates of the proximal model with the spatial random effects. The improvement in the goodness-of-fit statistic when the spatial structure was included confirms the spatial pattern of population density and average household size.

**Conclusion** The idea of assessing both the impact of covariates at the ecological level and spatial outlines in the same context should be encouraged in epidemiology to help with creating epidemiological surveillance systems (ESS) on a provincial basis for planning interventions and improvement of control programme efficiency.

**Keywords** Bayesian · Tuberculosis · Spatial · Multilevel · Epidemiology

## Introduction

Tuberculosis (TB), one of the first known human diseases, is still one of the major causes of death worldwide; about 2 million people die each year from the disease. TB has many indicators, affecting the bones, central nervous system, and many other organ systems, but it is primarily a pulmonary disease that is introduced by the deposition of *Mycobacterium tuberculosis*, contained in aerosol droplets, onto lung alveolar surfaces (Smith 2013). Tuberculosis is closely linked to both overcrowding and malnutrition, making it one of the principal diseases of poverty (Lawn and Zumla 2011). Many people in the developing world contract tuberculosis because of a poor immune system, largely due to high rates of HIV infection and the corresponding development of AIDS (Lawn and Zumla 2011).

Household crowding, deficient health services, and poor access to health services are well-known factors linked with TB prevalence and should be integrated in health policy planning. An effective programme for TB control should include these features not only as individual risk factors but as population determinants, comprising the basis for a territorial approach to a surveillance system (Barr et al. 2001). Strong evidence for an association between TB and poverty is already available, expressed by higher TB incidence rates in crowded urban areas and amongst low-income and illiterate populations (Waalder 2002). Health services access, in addition, is limited in the same populations, both as a result of lack of sufficient health facilities in poor areas and because poorer and illiterate people are less aware of their own health status. In a typical clinical practice, and most epidemiological studies, socio-economic determinants of disease are included in terms of an individual person's risk factors (Kawachi and Berkman 2003). In ecological studies, on the other hand, the focus is on the community as an entity in itself, an entity more complex than the sum of the individual persons who make it up (Berkman and Kawachi 2000). However, individual-level factors, such as behaviours and lifestyles, common to persons

---

✉ Davies Obaromi  
daviesobaroni@gmail.com

<sup>1</sup> Department of Statistics, University of Fort Hare, PMB X1314, Alice 5700, South Africa

across areas and subject to public health interventions, may also play a role in TB incidence. The question in ecological analysis is not about the causes of disease cases, but the causes of disease incidence (Rose 2001).

Previous disease mapping work has been based on collecting, mapping and analysing prevalence or incidence data with conventional statistical techniques, which are affected by random variation due to population variability and a loss of statistical power when cases are assigned to subgroups. The differences in geographical distribution due to random errors may be wrongly interpreted as a true variation of epidemiological interest. A Bayesian method, on the other hand, is used to model the random and true variation separately (Bergamaschi et al. 2006) and is a substitute for the frequentist methods.

Bayesian techniques can provide some shrinkage and spatial smoothing of raw standardised incidence ratio estimates, which are intensely influenced by the size of the population at risk, resulting in a noisy and unclear picture of the true unobserved risks (Richardson et al. 2004).

South Africa ranks as the third highest in the world burden of TB and for 3 consecutive years (2007, 2008 and 2009) the disease ranked as the number one among the ten leading underlying natural causes of death in SA (WHO 2010). Tuberculosis in the Eastern Cape Province mainly affects the economically active age group. Within the age group of 25–34 years the percentage distribution of reported TB cases was 15.9, 0.7 and 23.1% for the years 2003, 2004 and 2005 respectively. A report on the South African National Burden of Disease study 2000 Eastern Cape Province by the medical research council (MRC) showed that tuberculosis was the second leading cause of death among women and the third leading cause of death among men aged 15–44 years. Eastern Cape Province ranks as having the second highest burden of TB by province after KwaZulu Natal [National Department of Health, (NDoH) 2010 data]. In people with normal immune systems, the lifetime risk of progressing from latent TB infection to active TB disease is 10%. HIV, by weakening the immune system, increases a person's risk of progressing from latent TB infection to active TB disease by 10% per year. The province has an extremely high burden of TB, co-infection with TB and HIV (TB/HIV), and multi-drug-resistant tuberculosis (MDR-TB). In 2008, there were more than 60,000 new TB cases in the province. Of these, there were 1251 confirmed cases of MDR-TB and 385 confirmed cases of extensively drug-resistant tuberculosis (XDR-TB). In 2010, the total of new TB and re-treatment cases identified in the province stood at 62,226 (ECAC 2012).

The connection between tuberculosis (TB) and socio-economic status is well acknowledged (Souza et al. 2000;

Waalder 2002) and studies have succeeded in demonstrating an association between poverty and TB incidence rates (Krieger et al. 2003). The objective of this article was therefore to model the relationship between TB prevalence at the provincial level and socio-economic and demographic measures using Poisson regression and to discuss the capability of various indicators in guiding preventive actions and interventions.

## Methods

### Epidemiological data sources

This is a retrospective secondary data study using Eastern Cape Province TB notification and survey data. All data used were extracted from the electronic tuberculosis register (ETR) of the 24 health sub-districts of the province including the two metropolitan municipalities, namely, Nelson Mandela and Buffalo. The data obtained were for the period of 2012–2015.

The socio-economic and socio-demographic indicators and variables were obtained from publications of the Eastern Cape Socio-Economic Consultative Council (ECSECC 2014) containing reports from all local municipalities (Figs. 1 and 2).

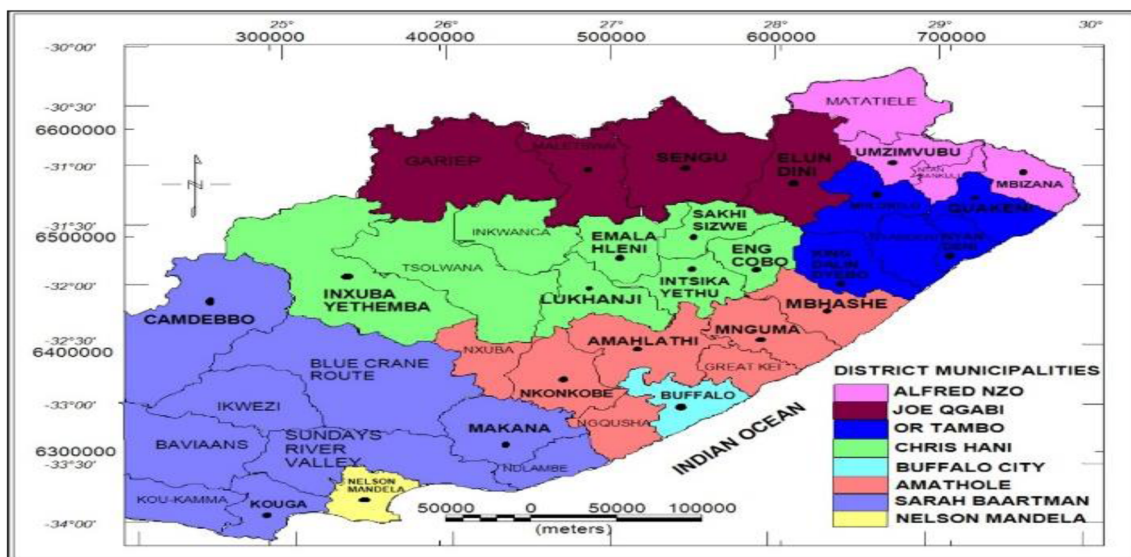
### Bayesian methods

We used a combination of a Bayesian approach and a generalised linear mixed model (GLMM) to smooth out the variability in observed disease rates and to estimate the association between TB prevalence rates, averaged over the observed period and chosen covariates of socio-economic vulnerability and demographic factors, and it therefore takes the form of:

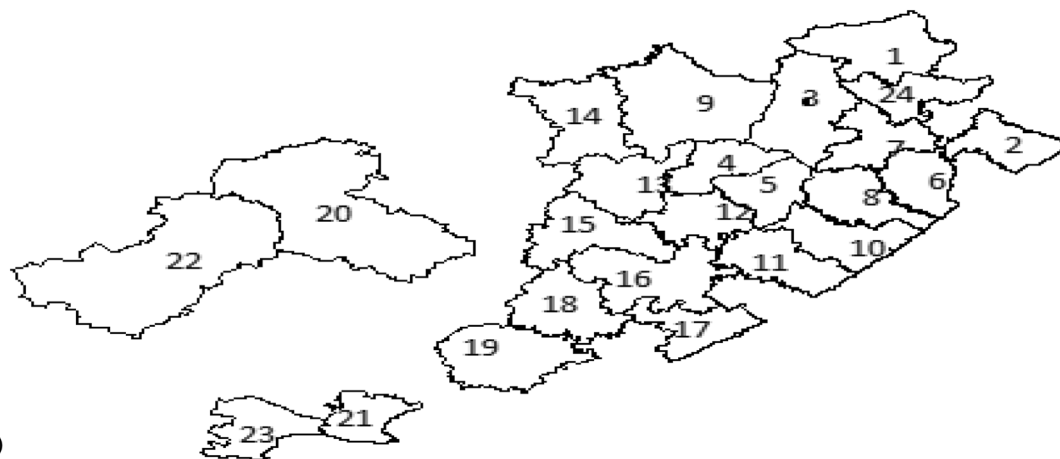
$$\log \lambda_i = \beta_0 + \sum_{i=1}^k X_i \beta_i + u_i + v_i \quad (1)$$

In the convolution regression model in (1) above,  $v_i$  is a non-spatially structured random effect, typically assumed to be independent Gaussian, with zero mean and variance  $\sigma_v^2$ , and it is usually included in models to account for extra-Poisson variation because of some non-measured important covariates. The structured spatial random effects,  $u = (u_1, \dots, u_n)$ , explains spatial dependence, with a prior distribution taken as an intrinsic conditional autoregressive model (ICAR), as seen in (2), where the mean value for  $u_i$  is the weighted average of the neighbouring random effects and the variance,  $\sigma_u^2$ , controls the strength of the local spatial dependence:

$$P(\mu_i | \mu_{j \neq i}) \sim N \left( \frac{\sum_{j \neq i} w_{ij} \mu_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma^2}{\sum_{j \neq i} w_{ij}} \right) \quad (2)$$



(a)



(b)

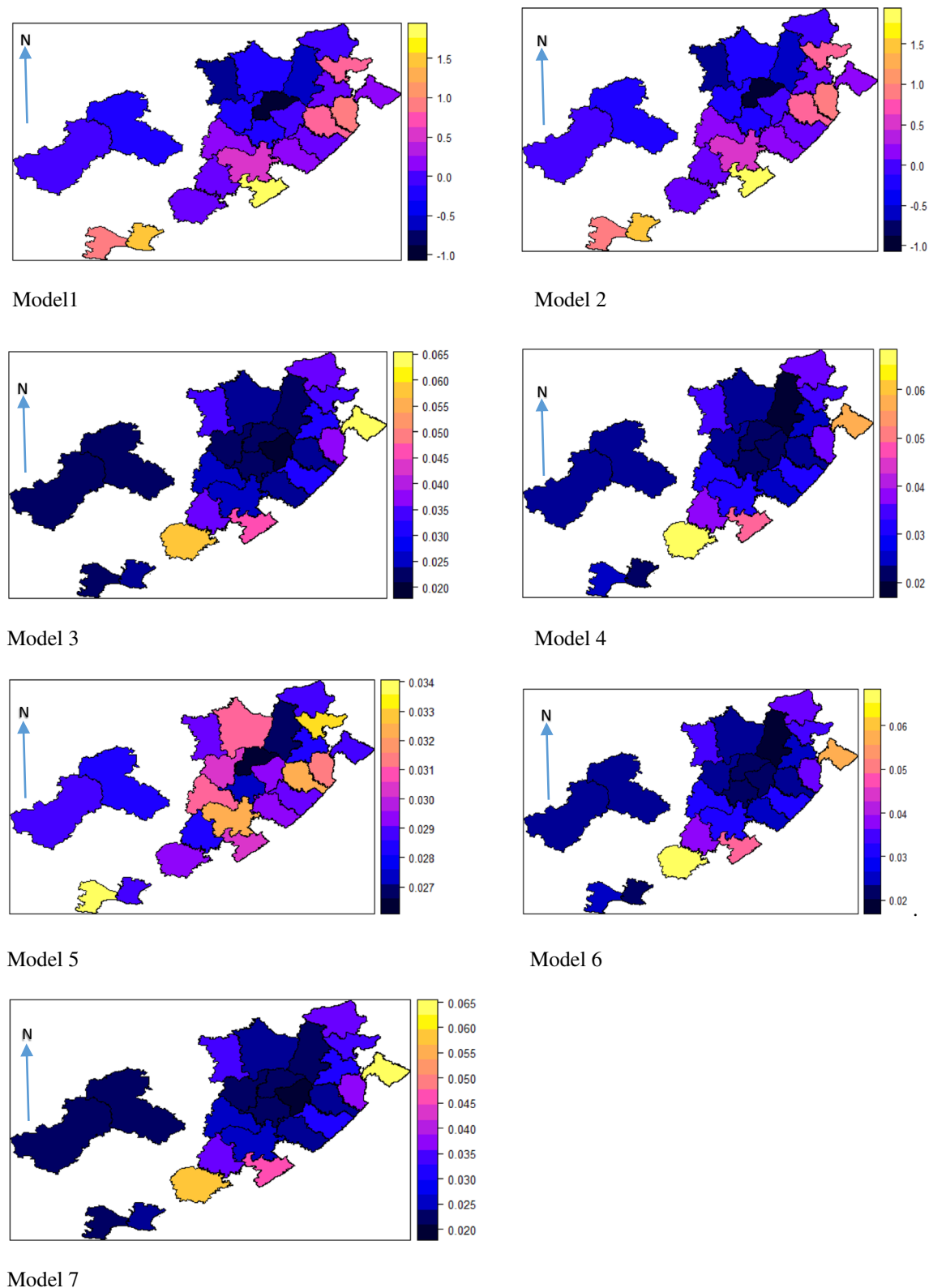
**Fig. 1** Map a: Eastern Cape Province showing 37 district municipalities and 2 metropolitans. Map b: 24 health sub-districts for the TB data set

Bayesian modelling depends on the ability to compute posterior distributions to provide estimates for all the corresponding model parameters. The majority of these posterior distributions are straightforward to calculate. Distributions with a conjugate prior typically have a posterior distribution that follows a standard distributional form. In many cases, however, the computation required is more complex and a more advanced method is essential to calculate the posterior distribution. These advanced approaches usually make use of some form of numerical simulation, generally by drawing a sample of parameter values from an approximation of the posterior distribution  $f(\theta|Y)$  to allow estimation of model parameter distribution.

For our model, six covariates (socio-economic and demographic factors) were taken as explanatory variables for the relative risk of the disease. Let  $\lambda_i$  be the number of new TB

cases in area  $i$ ,  $x_1$  = Gini coefficient (a measurement of how income or poverty is equally distributed),  $x_2$  = poverty rate (% number of individuals living below the poverty line, although there is no official poverty line defined for South Africa),  $x_3$  = unemployment rate (%),  $x_4$  = no schooling (%; for persons aged 20+ years),  $x_5$  = average household size and  $x_6$  = population density of the regions/municipalities. The Gini coefficient and poverty and unemployment rates are considered in this study as distal factors, while no schooling, average household size, and population density are taken as proximal factors.

Two spatial random effects of  $u_i$  and  $v_i$  were used as an unstructured random effect to measure for spatial heterogeneity and as a structured random effect to measure for spatial dependency among the regions respectively. In this study, seven separate multilevel models including/excluding the covariates/spatial random effects were developed and treated



**Fig. 2** Posterior estimated relative risks maps of the seven models

as non-independent Poisson random variables with means  $\lambda = (\lambda_1, \dots, \lambda)$  to investigate whether the covariates influenced part or all of the spatial correlations to TB risks.

Model comparisons were carried out using the deviance information criterion (DIC), which combines a measure of fit and a measure of model complexity based on the effective

number of parameters. Smaller DIC values show a more fitting model (Spiegelhalter et al. 2002). The R codes used are presented in the Appendix Table 2.

### Results

Table 1 shows the estimates from the combination of the Bayesian approach and the GLMM to assess spatial heterogeneity in the TB relative risk for the 2014 data set by investigating its relationship with some socio-economic and demographic variables in the Eastern Cape Province of South Africa. In 2014, Eastern Cape had 37,365 notified TB cases from 24 health sub-districts, with about 91.1% bacteriological coverage [ratio of the number of pulmonary tuberculosis (PTB) patients diagnosed by bacteriological tests to the total number of PTB patients reported, excluding children 0–4 years].

In this study, seven separate multilevel models including/excluding the covariates/spatial random effects were developed and treated as non-independent Poisson random variables with means  $\lambda_i = (\lambda_1, \dots, \lambda_n)$  to investigate whether the covariates influenced part or all of the spatial correlation in the TB relativity. In Table 1, the measures of association and their respective means and standard errors are presented for all covariates in all models. In the distal models, the most significant explanatory variable was poverty. The Gini coefficient, unemployment, and no schooling, which are three of the most widely used socio-economic indicators in South Africa, were not significant in any of the models and had very low standard errors. This was possibly due to one of the indirect effects of the trio in TB occurrence through their

influence on income (Singh-Manoux et al. 2002). For the proximal models in model 2 and the rest of the models, average household size was significant, indicating that each additional person in the house increased the risk of TB. Population density also played a significant role in the disease incidence, although with a low degree of association (models 2, 3, 5, and 6) and indirectly influenced the average household size. In the models without the spatial random effects (models 1, 2, and 3), the effects of poverty and average household size were vital.

The largest significant factor, however, was average household size.

In addition to being an indicator of poverty and population density, it was also an important covariate in the household transmission of TB. The addition of spatial random effects to the full model (comprising the distal and proximal models) decreased the significance of the effects of the previous covariates on TB incidence. By adding the conditional autoregressive random (CAR) effects, we also found that they can bring about significant changes in the posterior mean and variance of fixed effects compared with the non-spatial regression model. On the basis of the DIC, model 5 performed better than all the other models in which only the proximal factors and random effects were considered. In this best fit model, average household size and population density had a positive association with the relative risk of TB prevalence in the province. The impression is that models with a smaller DIC should be chosen in preference to models with a larger DIC. Both models are penalised by the value of  $\bar{D}$ , which favours a good fit, but also [in common with Akaike information criterion (AIC) and Bayesian information criterion (BIC)] by the effective number of

**Table 1** Posterior estimated means and SD of risk factors associated with TB by a GLMM model for Eastern Cape Province

Covariates	Model 1 Distal model	Model 2 Proximal model	Model 3 Proximal and distal	Model 4 Distal with random effects	Model 5 Proximal and random effects	Model 6 Proximal, distal and random effects	Model 7 Spatial random effects
Intercept	8.31 (0.035)	5.99 (0.062)	7.19 (0.072)	8.08 (0.760)	6.68 (0.422)	7.38 (1.071)	7.13 (0.132)
Gini coeff.	-0.33 (0.042)		-0.81 (0.045)	-0.60 (0.852)	-	-0.65 (0.647)	-
Poverty	0.10 (0.032)		0.08 (0.038)	0.33 (0.840)	-	0.06 (0.674)	-
Unemploym.	-0.03 (0.001)		-0.02 (0.001)	-0.02 (0.012)	-	-0.02 (0.009)	-
Pop. dens.	-	0.002 (0.000)	0.002 (0.000)		0.005 (0.000)	0.003 (0.001)	-
No school		-496.34 (16.00)	-388.89 (17.128)		-5.17 (31.302)	-5.40 (31.512)	-
Ave hhold size	-	0.44 (0.020)	0.35 (0.021)		0.02 (0.116)	0.10 (0.266)	-
ui-struc std.	-	-	-	18,369.33	1.21	1.55	1.841e+04
vi-unstruc	-	-	-	0.84	18,373.35	18,342.02	7.319e-01
pD	4.51	4.20	7.10	23.95	23.80	23.92	23.95
DIC	13,853.45	7123.56	6206.43	263.17	262.99	263.14	263.16
Log L	-7186.85	-3848.68	-3336.43	-219.18	-211.30	-227.39	-206.25

parameters,  $pD$ . Since  $\bar{D}$  will decrease as the number of parameters in a model increases, the  $pD$  term compensates for this effect by favouring models with a smaller number of parameters.

When the GLMM was adjusted for the proximal and distal spatial random effects models, it was seen to affect largely only the influence of poverty, which resulted in the decrease of the effect of almost all the other covariates. This outcome can be understood not only as a result of the collinearity between average household size and poverty, but also because of the fact that poverty is spatially clustered.

## Discussion

The model-by-model approach used to include variables in connected blocks enabled us to identify the significance of each covariate associated with the spread of the disease and the effects of spatial random terms. First, their inclusion in the models also significantly improved the fit of the models by the values from the DIC (Souza et al. 2007). Further, we established that socio-economic variables like the Gini coefficient and unemployment, which are distal factors, cannot on their own explain the incidence of TB. Proximal variables classified as demographic factors, which are related to population density and average household size during the period under study, were found to be significant in this study. The results showed the multilevel effects of socio-economic and demographic factors on TB, and it was generally observed that the relative risks imposed by poverty, population density, and average household size were strong and positively related to the disease outcome. We also found that the Gini coefficient, unemployment, and no schooling were not positively associated with TB prevalence in any of the models.

The combination of some socio-economic and demographic factors as distal and proximal factors, respectively, showed that both risk factors have interplaying effects on TB prevalence, as individuals tend to converge with relation to their similar economic and demographic resemblance. The Eastern Cape Province has a predominantly Black population; hence, there is a stronger tendency to converge as a community, which makes healthy individuals susceptible to the disease. Together, these variables show the role of different social and demographic practices respectively.

When only the goodness of fit as estimated by the DIC was considered, it was found that the models with spatial random effects were better fitted, even when the model had no independent variable. Better specific estimates of

the relative risk of the disease, as established by the inclusion and analyses of spatial random effects, and good fits were found in models 4, 5, and 6, thereby ratifying the assumption that the risk of occurrence is associated with the neighbourhood effect of each spatial unit. It is important to note that spatial random effects models on their own do not describe any disease incidence.

Implementing a fully categorised or multilevel Bayesian model to assess the relative risk of the incidence of TB was a suitable technique for this study. It is known that ‘all models are incorrect, but some are useful’; the possibility to assess the impact of explanatory variables at the ecological level as well the spatial patterns should be encouraged and supported in epidemiology in spite of the characteristic problems of this method (Hampel 1987).

Results from this study present two different alternatives for dealing with observing TB in the Eastern Cape Province. The use of socio-economic and demographic variables related to the prevalence of TB in the province points to the indicators to be examined and considered at the provincial level, while the mapping of risk areas show the need for health service interventions to target groups of the population at higher risk of contracting the disease. In addition, poverty, population density, average household size, and the spatial structure emphasise the associations between socio-economic and demographic deprivation and TB in this setting.

A restraint faced by many studies using this methodology is that most TB identification surveys (especially in heavy-burden and low-income countries like South Africa) do not have access to high-level statistical programmes, nor do they often have specialists to perform such analyses satisfactorily. Moreover, identifying the social and demographic processes that exacerbate circumstances of collective risk for the disease are essential to balanced planning of health and government mediations and may aid with creating ESSes on a territorial basis.

**Acknowledgments** We are grateful to data and research department, administrators, staff, and members of Eastern Cape Department of Health for releasing the data of the province. Our profound gratitude also goes to the Tertiary Education Trust Fund (TETFund) and the University of Fort Hare for their financial support.

**Authors' contributions** Davies Obaromi designed and formulated the study analysed and interpreted the data and also drafted the manuscript. Qin Yongsong and James Ndege supervised the entire concept of the study and revised the manuscript critically for intellectual criticisms.

**Funding** This study was funded by the University of Fort Hare, Alice, Eastern Cape Province.

## Compliance with ethical standards

**Conflicts of interest** The authors declare that they have no conflict of interest.

## Appendix

**Table 2** TB 2014 analysis data set

District	GINI	Poverty	TB	Noschool	POPDEN	AVGHouse	UNEMP	Population	Logpop
1	0.69	0.8005	843	0.00094	61.68	3.7	20.69	268,428	5.428828
2	0.65	0.636	2458	0.0008	92.29	3.8	28.53	231,401	5.364365
3	0.71	0.5205	1579	0.001	28.46	3.5	30.86	121,432	5.084333
4	0.5	0.6843	1185	0.00212	88.68	4.1	37.11	270,865	5.432753
5	0.51	0.6405	1272	0.00114	94.34	3.5	26.46	312,077	5.494262
6	0.38	0.6923	1042	0.00072	37.16	3.4	49.24	138,513	5.141491
11	0.39	0.6253	777	0.00188	36.51	3.7	63.84	129,886	5.113562
12	0.72	0.6803	926	0.00197	62.78	4	41.78	141,911	5.152016
13	0.2	0.7201	737	0.00146	62.36	3.5	49.78	189,883	5.278486
14	0.5	0.3521	818	0.00107	4.39	3.4	25.68	50,924	4.706923
15	0.44	0.6195	1418	0.00078	51.91	3.5	36.97	221,281	5.344944
16	1.01	0.615	388	0.00127	25.78	3.7	52.99	57,997	4.763406
17	0.58	0.6363	611	0.00159	25.63	3.5	21.05	130,174	5.114524
18	0.64	0.1868	533	0.0011	10.42	3.4	26.99	45,463	4.657658
19	0.74	0.3944	792	0.00145	17.23	3.5	35.63	126,195	5.101042
23	0.48	0.6087	1355	0.00208	120.59	4.7	23.55	296,808	5.472476
21	0.62	0.4849	2517	0.0014	156.27	4	21.02	456,447	5.65939
22	0.64	0.599	1164	0.00147	87.33	4.2	24.05	246,656	5.392092
24	0.51	0.5907	2838	0.00182	133.71	4.6	40.61	330,461	5.51912
8	0.54	0.0931	899	0.0009	5.98	3.8	18.91	43,292	4.636408
9	0.72	0.2649	2640	0.00049	31.47	3.2	14.62	76,091	4.881333
10	0.54	0.4738	1109	0.00063	17.08	3.4	36.51	74,793	4.873861
7	0.54	0.4653	5568	0.00049	311.26	3.2	21.14	785,330	5.895052
20	0.69	0.442	3896	0.0003	569.32	3.4	26.77	1,111,767	6.046014

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Barr RG, Diez-Roux AV, Knirsch CA, Pablos-Méndez A (2001) Neighborhood poverty and the resurgence of tuberculosis in New York City, 1984–1992. *Am J Public Health* 91:1487–1493
- Bergamaschi R, Montomoli C, Candeloro E, Monti MC, Cioccale R, Bernardinelli L, Fratino P, Cosi V, PREMS (Pavia Register of Multiple Sclerosis) Group (2006) Bayesian mapping of multiple sclerosis prevalence in the province of Pavia, northern Italy. *J Neurol Sci* 244(1–2):127–131. 17
- Berkman LF, Kawachi I (2000) *Social epidemiology*. Oxford University, New York
- Eastern Cape AIDS Council (2012) Provincial strategic plan for HIV and AIDS, STIs and TB 2012–2016
- Eastern Cape Socio-Economic Consultative Council (2014) Buffalo City Metro-Eastern Cape socioeconomic profile. ECSECC
- Hampel FR (1987) Data analysis and self-similar processes. In: *Proceedings of the 46th Session of the International Statistical Institute book 4*, Tokyo. International Statistical Institute 10 Voorburg, NL, 235–254
- Kawachi I, Berkman LF (2003) *Neighbourhoods and health*. Oxford University Press, New York
- Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian SV (2003) Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures—the public health disparities geocoding project (US). *Public Health Rep* 118:240–260
- Lawn SD, Zumla AI (2011) Tuberculosis. *Lancet* 378(9785):57–72. [https://doi.org/10.1016/S0140-6736\(10\)62173-3](https://doi.org/10.1016/S0140-6736(10)62173-3)
- National Department of Health (2010) Management of drug-resistant tuberculosis: policy guidelines [http://www.tbonline.info/media/uploads/documents/mdr-tb\\_sa\\_2010.pdf](http://www.tbonline.info/media/uploads/documents/mdr-tb_sa_2010.pdf). Accessed 4 Nov 2013

- Richardson S, Thomson A, Best N, Elliott P (2004) Interpreting posterior relative risk estimates in disease mapping studies. *Environ Health Perspect* 112:1016–1025
- Rose G (2001) Sick individuals and sick populations. *Int J Epidemiol* 30: 427–432
- Singh-Manoux A, Clarke P, Marmot M (2002) Multiple measures of socio-economic position and psychosocial health: proximal and distal measures. *Int J Epidemiol* 31:1192–1199
- Smith I (2013) *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev.* <https://doi.org/10.1128/CMR.16.3.463-496>
- Souza WV, Ximenes RAA, Albuquerque MFM et al (2000) The use of socioeconomic factors in mapping tuberculosis risk areas in a City of Northern Eastern Brazil. *Rev Panam Salud Publica* 8:403–410
- Souza WV, Carvalho MS, Albuquerque Mde F, Barcellos CC, Ximenes RA (2007) Tuberculosis in intra-urban settings: a Bayesian approach. *Trop Med Int Health* 12(3):323–330
- Spiegelhalter D, Best N, Carlin BP et al (2002) Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Methodol* 64:539–583
- Waller HT (2002) Tuberculosis and poverty. *Int J Tuberc Lung Dis* 6: 745–746
- World Health Organization (2010) Tuberculosis Fact sheet N°104"