Quantile Regression for Area Disease Counts: Bayesian Estimation using Generalized Poisson Regression

Peter Congdon*

School of Geography, Queen Mary University of London, UK

Abstract: Generalized linear models based on Poisson regression are commonly applied to count data for area morbidity outcomes, focused on modelling the conditional mean of the response as a function of a set of risk factors. Mean regression models may be sensitive to outliers and provide no information on other distributional features of the response. We consider instead a Poisson lognormal hierarchical approach to quantile regression of spatially configured count data, allowing for observed risk factors and spatially correlated unobserved risk factors. This technique has the advantage that a profile of the relative outcome risk across quantiles can be obtained, including estimates of uncertainty (e.g. the uncertainty attaching to 2.5% or 5% relative risk quantiles). An application involves counts of emergency hospitalisations for self-harm for 6791 small areas in England. Known risk factors are area deprivation, a measure of social fragmentation and a measure of rural status. It is shown that impact of these predictors varies between quantiles, and that hierarchical quantile regression generally produces narrower risk intervals, except for outlier areas, and leads to a higher number of areas being classed as high risk.

Keywords: Hierarchical quantile regression. Relative risk. Risk intervals. Elevated risk. Self-harm.

1. BACKGROUND

Community interventions to tackle health inequalities are often based on area variations in disease relative risk, or in relative risks of morbidity related outcomes (e.g. hospitalisations). For example, area-based interventions have been applied to reduce suicide and non-fatal self-harm [1]. Establishing the epidemiology of such variations is also important as a basis for targeted intervention. To this end regression methods may be applied to disease counts for administrative areas [2] to assess significant ecological risk factors.

In particular, generalized linear models (GLMs) based on Poisson regression [3] are commonly applied to data on disease counts, with the wider analytical framework based on GLMs denoted as Bayesian disease mapping [4]. In a generalized linear model, the focus is on modelling the conditional mean of the response as a function of a set of risk factors. Inferences about relative disease risks, and about areas with elevated risk, are then based on the conditional mean regression model. In particular, one may consider credible intervals for relative risks, or exceedance probabilities [5].

Mean regression models may be sensitive to outliers and provide no information on other distributional features of the response. By contrast, quantile regression considers either the conditional median or other conditional quantiles of the response variable [6]. There is little research comparing how inferences about modelled spatial patterns of disease risk are affected by using a quantile regression approach rather than conditional mean generalized linear models.

In this paper, we provide an assessment of how spatial patterns of extreme relative disease risk may be affected by using quantile regression applied to area counts of disease, mortality, or hospitalisations. A Bayesian inference and estimation strategy is used. We consider a hierarchical approach to quantile regression of spatially configured count data, allowing for observed risk factors and spatially correlated unobserved risk factors. An application involves counts of emergency hospitalisations for self-harm for 6791 small areas in England over a five year period. Known risk factors are area deprivation, a measure of social fragmentation and a measure of rural status.

2. METHODS OVERVIEW AND LITERATURE

Bayesian disease mapping is based around generalized linear models for typically discrete responses (e.g. counts of deaths or hospitalisations for administrative areas). Such models typically involve conditional mean estimation using both known risk factors and random effects to represent unknown risks or overdispersion. However, mean regression models may be sensitive to outliers and provide no information on factors affecting other distributional features of the response.

By contrast, quantile regression estimates the relationship between the α^{th} quantile $Q_Y(\alpha \mid X)$ of the

*Address correspondence to this author at the School of Geography, Queen Mary University of London, UK; Tel: +44 (0)20 7882 8200;
E-mail: p.congdon@qmul.ac.uk

response Y and covariates X [6] Quantile regression was originally developed for continuous responses as count responses do not have continuous quantiles. For $\alpha \in (0,1)$ and continuous Y, frequentist quantile regression involves minimizing $\sum_{i=1}^{n} \rho_\alpha (Y_i - X_i^T \beta_\alpha)$, where $\rho_\alpha(u_i) = u_i(\alpha - I(u_i < 0))$, and $u_i = Y_i - X_i^T \beta_\alpha$.

Following [7], a Bayesian approach to quantile regression is obtained using an Asymmetric Laplace distribution (ALD), with density function

$$\mathrm{ALD}(y \mid \eta_\alpha, \delta_\alpha, \alpha) = \frac{\alpha(1-\alpha)}{\delta_\alpha} \exp\left[-\frac{\rho_\alpha(y-\eta_\alpha)}{\delta_\alpha} \right],$$

where η_α is a regression term, α is the quantile, and δ_α is a scale parameter. This distribution can in turn be represented as a scale mixture of normals [8]. Thus for observations $i = 1,..,n,$ and assuming $Y_i \sim \mathrm{ALD}(\eta_{\alpha i}, \delta_\alpha, \alpha),$ one has

$$Y_i = \eta_{\alpha i} + \xi_\alpha W_{\alpha i} + \left[\frac{2W_{\alpha i}\delta_\alpha}{\alpha(1-\alpha)} \right]^{0.5} V_{\alpha i},$$

where $\eta_{\alpha i}$ is the regression term, $\xi_\alpha = \frac{(1-2\alpha)}{\alpha(1-\alpha)}$, $\delta_\alpha > 0, W_{\alpha i} \sim \mathrm{Exp}(\delta_\alpha)$, and $V_{\alpha i} \sim N(0,1)$.

To extend quantile regression to count data, [9] propose adding uniform noise U to count responses, giving $Z = Y + U$. With offsets ej they apply quantile regression of the form

$$Q_{Z_i}(\alpha \mid X_i) = \eta_{\alpha i} = \alpha + \exp(X_i^T \beta_\alpha) + \ln(e_i).$$

A Bayesian quantile regression adaptation to spatial count data is set out by [10], based on the procedure of [9]. Spatial quantile regression has been studied by other papers, both from Bayesian and frequentist perspectives, generally with continuous responses such as house prices, house rentals, or medical expenditures (e.g. [11], [12]). Reich *et al.* [13] adopt a spatial Bayesian approach to a continuous (positive) response, namely maximum ozone readings.

Implications of a quantile regression approach to inferences from disease mapping for count lattice data using generalized Poisson regression are relatively unexplored. Dreassi *et al.* [14] provide a semiparametric Negative Binomial M-quantile regression method applied to count lattice data, which involves frequentist estimation. Requia *et al.* [15] use lattice disease data but with a continous response (albeit based on originally count data, cardiorespiratory disease hospitalisations) and frequentist quantile regression, enabling use of the R package quantreg. Chiu *et al.* [16] also use a continous response. Transformations of count outcomes are sometimes used, but this is subject to pitfalls [17]. So a likelihood which reflects the count nature of the response is preferred.

3. QUANTILE REGRESSION FOR SPATIAL DISEASE COUNTS

In this paper, we consider quantile regression for area disease counts, overdispersed in relation to the Poisson assumption, and adopt a scale mixture version of the ALD within a hierarchical Poisson lognormal representation to account for overdispersion. Specifically, let observed and expected disease counts for areas i be denoted $\{Y_i, E_i\}$, where E_i are offsets based on demographic standardisation, subject to $\sum_i Y_i = \sum_i E_i$. The E_i are obtained either by multiplying area populations by the region-wide outcome rate, or multiplying age-specific populations by region-wide age specific outcome rates. Then subject to the necessity to take account of overdispersion, the Y_i may be taken as Poisson,

$$Y_i \sim \mathrm{Poi}(\mu_i), \qquad (1)$$

$$\mu_i = E_i \exp(\rho_i).$$

The quantities $R_i = \exp(\rho_i)$ represent relative disease risks, which have value 1 for the entire region when $\sum_i Y_i = \sum_i E_i$.

Consider a conventional Poisson-lognormal distribution (PLN) to represent overdispersion [18]. Mahaki *et al.* [19] compare the PLN and negative-binomial representations to modelling spatial relative risks for overdispersed disease counts, and mention that the Poisson-lognormal representation is advantageous in terms of readily being able to include spatially configured residual effects. Moreover the tails of the log-normal are heavier than for the gamma distribution, and for data with outliers, the PLN model may give a better fit than the negative-binomial model when the counts are overdispersed [20].

Hence the PLN is advantageous as a basis for introducing quantile regression at the second stage, as in the current paper. Here we modify the PLN representation so that a scale mixture ALD is applied at the second stage specific to quantiles α, and Poisson sampling applies at the first stage. The second stage regression is focussed on estimating conditional quantiles in relative risks, which are latent quantities of primary epidemiological interest. Thus for observed and expected disease counts $\{Y_i, E_i\}$, we specify for quantiles $\alpha = 1,..,A$,

$$Y_i \sim \mathrm{Poi}(\mu_{\alpha i}), \tag{2}$$

$$\mu_{\alpha i} = E_i \exp(\nu_{\alpha i}),$$

$$\nu_{\alpha i} \sim N(X_i\beta_\alpha + s_{\alpha i} + \xi_\alpha W_{\alpha i}, \tfrac{2W_{\alpha i}\delta_\alpha}{\alpha(1-\alpha)}),$$

$$W_{\alpha i} \sim \mathrm{Exp}(\delta_\alpha),$$

$$\rho_{\alpha i} = X_i\beta_\alpha + s_{\alpha i},$$

$$R_{\alpha i} = \exp(\rho_{\alpha i}).$$

Spatially correlated effects $s_{\alpha i}$ are included in the regression term for quantile α to represent spatially clustered, but unobserved, risk factors for areas i, these being quantile specific (cf. [11, 21]). The prior density for the spatial effects $s_{\alpha i}$ is discussed in the next section. Of substantive importance are conditional quantile regression coefficient estimates, and relative risk inferences from quantile regression, as against conditional mean regression, though quantile and conditional mean regression are obtained by mathematically distinct approaches.

The $W_{\alpha i}$ in equation (2) are measures of outlier status. Areas with higher $W_{\alpha i}$ have higher variances (lower precisions) and hence diminished influence on the likelihood. Relative disease risks for areas with high $W_{\alpha i}$ are likely to have a wide uncertainty interval.

4. DETECTING ELEVATED RELATIVE RISK

In disease mapping applications the canonical model involves observed and expected disease counts $\{Y_i, E_i\}, i = 1,..,n$, typically overdispersed. The convolution or BYM prior of [3] is often used in analyzing area disease variations and to identify areas with elevated relative risk [5]. Thus as per equation (1), one has a conditional mean regression

$$\rho_{i,BYM} = X_i\beta + u_i + s_i,$$

where u_i is an independent and identically distributed (iid) normal random effect $u_i \sim N(0, \sigma_u^2)$, and the s_i are spatially dependent, or equivalently

$$\rho_{i,BYM} \sim N(X_i\beta + s_i, \sigma_u^2). \tag{3}$$

There are two sets of random effects, with s_i representing a relatively smooth underlying spatially dependent effect, and u_i representing remaining overdispersion or idiosyncractic area effects. This model implies shrinkage and spatial smoothing to a pattern of underlying risks by borrowing strength across neighbouring locations, so producing increased precision in risk estimates [22]. With spatial interaction represented by binary adjacency, the s_i follow a conditional autoregressive prior,

$$s_i \mid s_{[i]} \sim N\left(\sum_{j \in N_i} s_j / L_i, \sigma_s^2 / L_i\right),$$

with $s_{[i]}$ denoting s effects excluding s_i, and L_i the number of areas adjacent to area i in its locality N_i .

Bayesian disease mapping models often focus on assessing elevated relative risk in different areas. For Markov Chain Monte Carlo (MCMC) iterations $t = 1,..,T$, let $R_{i,BYM}^{(t)} = \exp(\rho_{i,BYM}^{(t)})$ be sampled relative risks in area i at iteration t based on the conditional mean regression. Extreme quantiles of relative risk, such as the 2.5% and 97.5% quantiles, are estimated from the sampled $R_{i,BYM}^{(t)}$. If there are T samples of the conditional mean relative risk from the BYM model, the 0.025 percentile of these T samples is the estimate of the 2.5% risk quantile. The uncertainty associated with this estimate is not available from a single MCMC run. Elevated risk may be indicated by areas with 95% credible intervals for $R_{i,BYM}$ entirely above 1, that is with both 2.5% and 97.5% quantiles exceeding 1. Less stringent risk classification might be based on 90% or 80% credible intervals being entirely above 1.

By contrast, under the quantile regression (2), extreme conditional quantiles of relative risk may be estimated from quantile specific regression on risk factors or spatial random effects. Thus relative risks for quantile α are estimated from sampled $R_{\alpha i}^{(t)} = \exp(X_i\beta_\alpha^{(t)} + s_{\alpha i}^{(t)})$, with full posterior densities available for relative risk quantiles $R_{\alpha i}$. Elevated risk is then indicated by areas with posterior mean estimates for the 2.5% and 97.5% quantiles $\{R_{0.025,i}; R_{0.975,i}\}$ both exceeding 1. In view of the epidemiological emphasis on detecting elevated (as opposed to depressed) relative risk, particular importance attaches to impacts of predictors and spatial smoothing on lower quintiles such as $R_{0.025,i}$. If the focus is widened to detecting areas with depressed risk, then this is indicated by areas having posterior mean estimates for $\{R_{0.025,i}; R_{0.975,i}\}$ both under 1.

It should be noted that this quantile regression approach to assessing extreme relative risks offers an alternative and complementary perspective to the conventional Bayesian disease mapping (BDM) methodology based on conditional mean regression. There is no implicit assumption of mathematical equivalence between (say) the 2.5 percentile of the posterior of the relative risk estimated via BDM conditional mean regression, and the 2.5th percentile of relative risk estimated via quantile regression. Quantile regression enables one to assess influences of spatial covariates on extreme relative risks, whereas conventional BDM estimates of relative risk follow from a conditional mean regression on spatial covariates. Characterizing extreme rates or risk is important not just in health but other applications (e.g. climatic, environmental). As Zhou *et al.* [23] mention "Therefore improving the ability to characterize extreme temperature events is of critical importance. To this end, quantile regression is an important tool for characterizing the tail probabilities". Similarly Reich *et al.* [13] mention that "correctly estimating the tail probability is critically important in studying the health effects of ozone exposure, and has policy implications".

5. CASE STUDY IMPLEMENTATION

The dataset considered consists of counts Y_i of emergency hospitalisations for self-harm for $n = 6791$ small areas (middle level super output areas, MSOAs) in England over the five year period (financial years 2010/11 to 2014/15). These MSOAs are subdivisions of 326 local government units ("local authorities"). The average event count is 79, and the variance of event counts is 2815. There are four areas with unknown self-harm counts, not released as the count is 5 or under. For these areas truncated Poisson sampling is used with upper limit 5.

Known risk factors are area deprivation, a measure of social fragmentation and a measure of rural status. Deprivation (Z_1) is measured by the 2015 Index of Multiple Deprivation (IMD) [24], with most indicators being based on the period 2012/13. Social fragmentation (e.g. [25]) is defined by indicators from the 2011 UK Census, including measures of marital status, population turnover, single person households and private sector renting. Rural status is based on accessibility indicators (access to services and facilities) developed by as part of the 2015 IMD development. These are road distances to the nearest post office, nearest primary school, nearest general store/supermarket, and distance to a doctors (GP) surgery, with rural areas expected to have lower accessibility. The social fragmentation and rurality scores (Z_2, Z_3) are obtained by summing z-scores on constituent indicators. For the subsequent regression all three predictors are transformed to a common [0,1] scale, so that for areas i and index j, $X_{ij} = (Z_{ij} - Z_{j,\min}) / (Z_{j,\max} - Z_{j,\min})$.

Regression analysis of the BYM, as in (3), and hierarchical QR (HQR) model, as in (2), is carried out in WINBUGS14 [26]. Inferences are based on the second halves of 20,000 two chain runs with convergence assessed using Brooks-Gelman-Rubin diagnostics [27]. Normal N(0,100) priors are adopted on β parameters, and gamma priors with shape 1 and rate 0.001 on precision (inverse scale) parameters $1/\sigma_u^2, 1/\sigma_s^2, 1/\sigma_{\alpha s}^2$, and $1/\delta_\alpha$. Model fit is based on the Watanabe-Akaike information criterion (WAIC) [28] evaluated for each of five quantiles $\alpha = 0.025, 0.25, 0.5, 0.75$ and 0.975; a possible alternative is the Deviance Information Criterion or DIC [29]. The WAIC is based on the sum LL_{WAIC} of log posterior mean likelihoods for each observation, and a complexity measure C_{WAIC} based on summing the posterior variances of posterior mean log-likelihoods.

Model checks are provided by posterior predictive p-tests [30]. These involve sampling replicate data Y_{rep} (or $Y_{rep,\alpha}$) from the particular model being estimated and evaluating a test statistics $T(Y_{rep};\theta)$ and $T(Y;\theta)$ for replicate and actual data. The test statistics $\Pr[T(Y_{rep};\theta) > T(Y;\theta)]$ should be within the range 0.1 to 0.9 for a model that satisfactorily reproduces the actual data. Three test statistics are used: the chi square $\sum(Y_{rep,i} - \mu_i)^2 / \mu_i$ (for BYM) or $\sum(Y_{rep,\alpha,i} - \mu_{\alpha i})^2 / \mu_{\alpha i}$ (for HQR); the mean absolute deviation $\sum | Y_{rep,i} - \mu_i | / n$ (for BYM) or $\sum | Y_{rep,\alpha,i} - \mu_{\alpha i} | / n$ (for HQR); and the maximum observation, $\max(Y_{rep,i})$ (for BYM) and $\max(Y_{rep,\alpha,i})$ (for HQR). The scaled deviance is also obtained as this should be approximately n for Poisson data [31].

6. CASE STUDY: RESULTS

Table **1** presents comparative fit and model checks for the BYM model and the HQR model estimated at quantiles $\alpha = 0.025, 0.25, 0.5, 0.75$ and 0.975. In terms of models estimating central relative risk, the median quantile regression has a lower WAIC than the conditional mean BYM model. Posterior predictive checks are satisfactory for all the models, and credible intervals for the scaled deviance include n.

Table 1: Model Fit and Checks*

		(a) Model Fit		
HQR	**Quantile**	**LL_{WAIC}**	**C_{WAIC}**	**WAIC**
	0.025	-22884	3164	52097
	0.25	-22871	3217	52177
	0.50	-22823	3216	52078
	0.75	-22828	3252	52160
	0.975	-22840	3222	52125
BYM		-22837	3211	52096
		(b) Model checks		
	Posterior Predictive P-tests			
HQR	**Quantile**	**Chi-square**	**Mean absolute deviation**	**Maximum**
	0.025	0.25	0.38	0.59
	0.25	0.26	0.38	0.60
	0.50	0.46	0.53	0.60
	0.75	0.32	0.39	0.55
	0.975	0.23	0.30	0.56
BYM		0.42	0.49	0.57
Scaled Deviance				
HQR	Quantile	Mean	2.5%	97.5%
	0.025	6942	6699	7167
	0.25	6936	6713	7165
	0.50	6838	6609	7072
	0.75	6867	6637	7098
	0.975	6896	6682	7121
BYM		6869	6653	7101

*Abbreviations: HQR: Hierarchical Quantile Regression; BYM: Besag *et al.* Spatial Model; LL_{WAIC}: Total of log posterior mean likelihoods; C_{WAIC}: Complexity Term in WAIC; WAIC: Watanabe-Akaike information criterion.

Table 2: Regression Coefficients, BYM vs Hierarchical Quantile Regression

				Posterior Summary	
BYM			**Mean**	**2.5%**	**97.5%**
		Intercept	-0.59	-0.62	-0.56
		IMD (Deprivation)	2.00	1.94	2.06
		Rurality	-0.83	-0.90	-0.76
		Social Fragmentation	0.40	0.30	0.49
HQR	Quantile				
	0.025	Intercept	-0.71	-0.74	-0.69
		IMD (Deprivation)	2.01	1.96	2.07
		Rurality	-0.85	-0.91	-0.77
		Social Fragmentation	0.39	0.31	0.50

(Table 2). Continued.

0.25	Intercept	-0.67	-0.70	-0.64
	IMD (Deprivation)	2.02	1.98	2.07
	Rurality	-0.86	-0.93	-0.79
	Social Fragmentation	0.38	0.29	0.47
0.50	Intercept	-0.59	-0.62	-0.56
	IMD (Deprivation)	1.99	1.93	2.05
	Rurality	-0.84	-0.91	-0.77
	Social Fragmentation	0.41	0.33	0.50
0.75	Intercept	-0.52	-0.55	-0.49
	IMD (Deprivation)	1.98	1.91	2.04
	Rurality	-0.81	-0.88	-0.74
	Social Fragmentation	0.46	0.36	0.57
0.975	Intercept	-0.49	-0.52	-0.45
	IMD (Deprivation)	1.98	1.92	2.03
	Rurality	-0.78	-0.87	-0.71
	Social Fragmentation	0.47	0.39	0.55

Table **2** shows regression coefficients for the BYM and HQR regressions. All regressions show positive effects of deprivation and social fragmentation on self-harm, and that deprivation is the stronger influence. All regressions show self-harm declining in more rural areas. However, the quantile regression sequence shows impacts of social fragmentation to increase for higher quantiles, while those for rurality attenuate.

Table **3** shows the decile breaks in alternative estimates (posterior means) of central relative risk over the 6791 areas: the conditional mean under the BYM model and the conditional median under the hierarchical QR model. These are very similar between the two estimators. The correlation between the posterior means on the two estimators is 0.987. Histograms and density plots of the two estimates are shown in Figure **1**.

An England-wide map of the median relative risks from the HQR model is shown in Figure **2**. The highest self-harm levels tend to be in Northern England, though some large towns in Southern England (except the South East) are also areas with high self-harm risks. Figure **3** contains averages of MSOA relative risks within twenty local authorities, ranked from the highest.

Table **4** shows decile breaks in estimates of the 2.5% quantile of relative risk, important in assessing excess relative risk. Under the BYM model these are point estimates obtained from the MCMC samples of the conditional mean relative risk, without any indication of sampling variability. Under the quantile

Table 3: Decile Points, Estimates of Central Relative Risk

Decile	BYM (Mean)	HQR (Median)
0.1	0.43	0.46
0.2	0.54	0.56
0.3	0.64	0.65
0.4	0.73	0.74
0.5	0.84	0.84
0.6	0.98	0.97
0.7	1.16	1.14
0.8	1.40	1.38
0.9	1.76	1.71

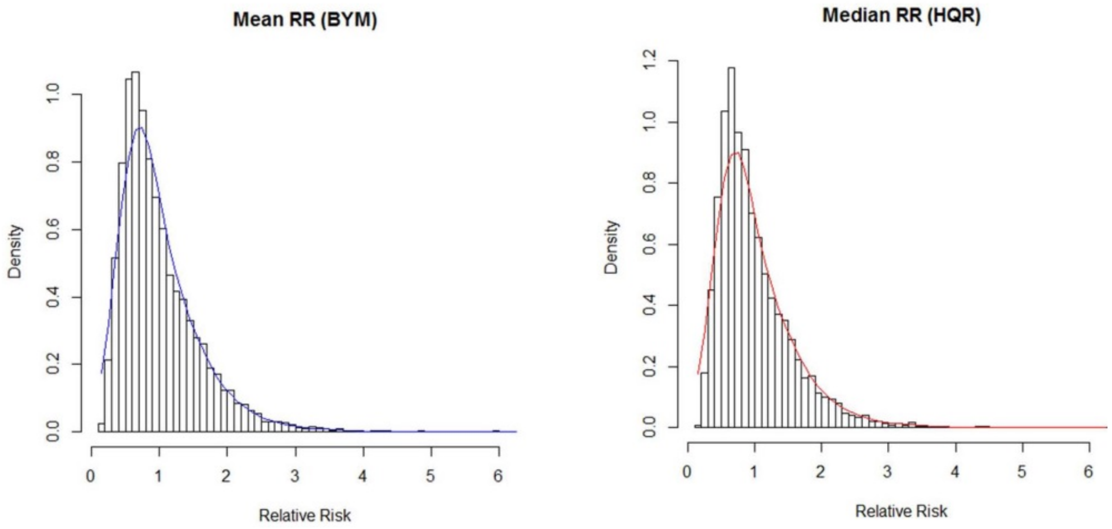

Figure 1: Estimates of central relative risk.

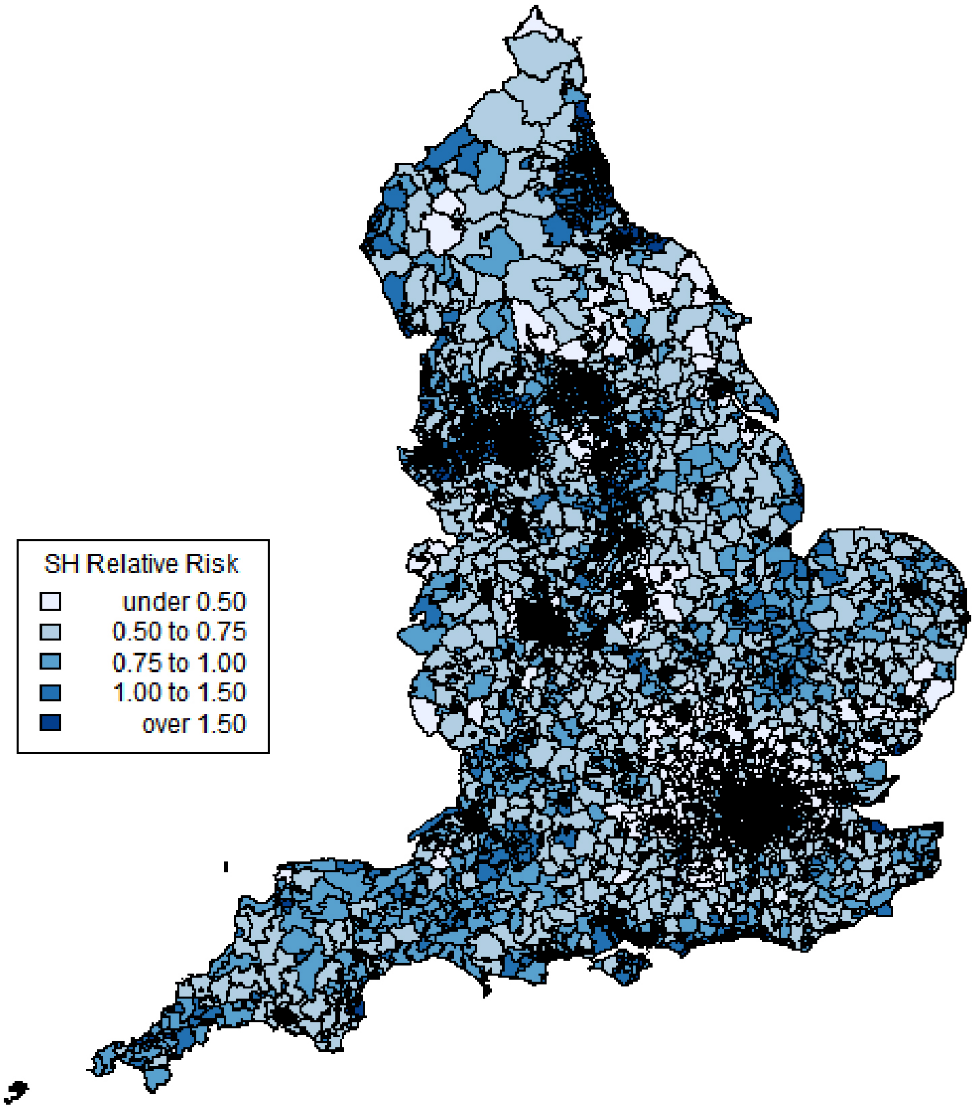

Figure 2: Median relative risks, self-harm.

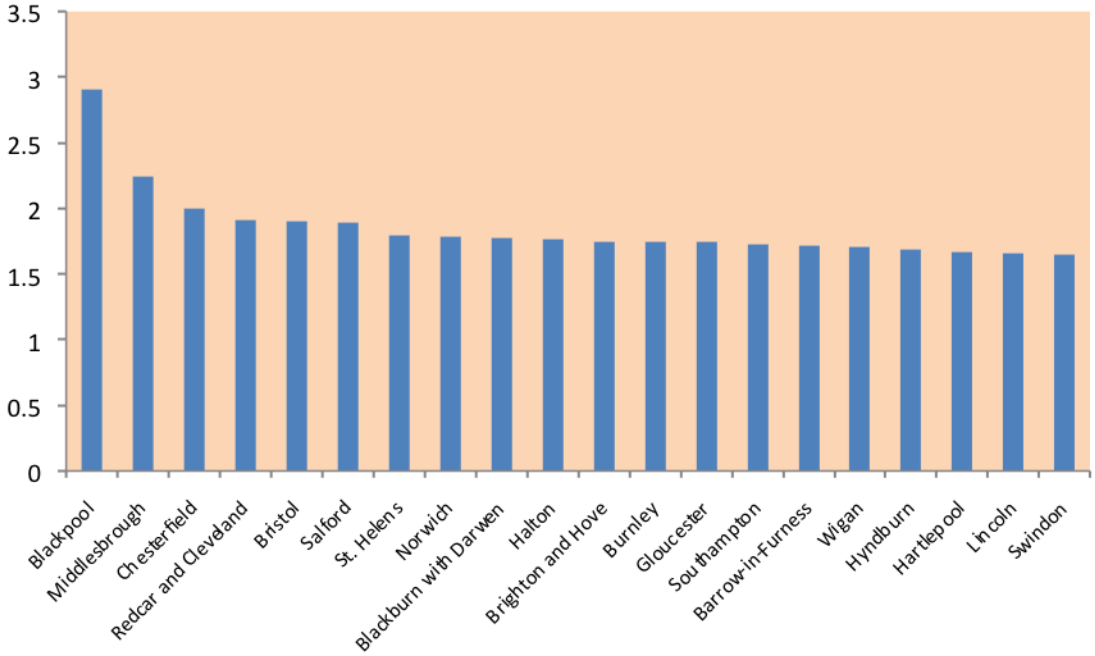

Figure 3: Average self-harm relative risk (median HQR).

Table 4: Decile Points, Estimates of 2.5% Quantile of Relative Risk

Decile	BYM	HQR
0.1	0.32	0.40
0.2	0.41	0.49
0.3	0.49	0.57
0.4	0.57	0.65
0.5	0.67	0.75
0.6	0.79	0.87
0.7	0.95	1.01
0.8	1.16	1.23
0.9	1.50	1.54

regression model, estimates of the 2.5% quantile of relative risk are based on the coefficients for $\alpha = 0.025$ shown in Table **2**, and subject to shrinkage and spatial smoothing due to random effects borrowing strength across neighboring locations. The quantile regression estimates (posterior means) are shifted slightly upwards compared to the BYM point estimates. However, for 6673 of the 6791 areas the 95% credible interval for $R_{0.025,i}$ obtained from quantile regression includes the BYM point estimate.

Table **5** cross-tabulates areas with elevated risk under BYM and HQR models. Elevated risk under the BYM models is based on areas with both 2.5% and 97.5% quantiles for $R_{i,BYM}$ exceeding 1. Elevated risk under the HQR model is based on posterior mean estimates for the 2.5% and 97.5% quantiles $\{R_{0.025,i}; R_{0.975,i}\}$ both exceeding 1. Of the 1857 areas with elevated risk according to the BYM model, such a classification is also obtained under quantile regression for 1825 areas. This demonstrates concordance in risk classification.

However, quantile regression detects a higher number of areas (2094 vs. 1857) with elevated risk than under the BYM model. This may be due, for example, to borrowing strength in HQR estimates of the 0.025 quantile. Borrowing strength also affects HQR estimates of the 0.975 quantile, so that 95% intervals for relative risk under the quantile approach will tend to be narrower than under the BYM model. The exception to this would be in areas with high $W_{\alpha i}$.

Table 5: Cross-Tabulation of Elevated Relative Risk

		BYM model		
		Not elevated	**Elevated**	**Total**
Quantile Regression	Not elevated	4665	32	4697
	Elevated	269	1825	2094
	Total	4934	1857	6791

Table 6: Relative Risk Intervals, Ashfield MSOAs

		Self-Harm		**Scaled Area Predictors**				**BYM Point Estimates of RR Quantiles**			**HQR RR Quantile Estimates at α=0.025 and α=0.975**		
MSOA Code	**MSOA Name**	**Y**	**E**	**IMD**	**Rurality**	**SF**	**Posterior mean RR (BYM)**	**0.025**	**0.975**	**Posterior Median RR (HQR estimate for α=0.5)**	**0.025**	**0.975**	**Normalized W**
5690	Ashfield 001	69	57.9	0.34	0.19	0.21	1.16	0.93	1.42	1.12	1.02	1.26	0.88
5691	Ashfield 002	33	53.6	0.17	0.16	0.17	0.65	0.48	0.86	0.68	0.61	0.76	0.95
5692	Ashfield 003	64	63.5	0.40	0.09	0.25	1.08	0.86	1.32	1.19	1.01	1.30	1.04
5693	Ashfield 004	137	102.8	0.36	0.14	0.30	1.31	1.11	1.53	1.26	1.15	1.42	0.90
5694	Ashfield 005	117	73.6	0.57	0.14	0.31	1.61	1.36	1.90	1.67	1.49	1.84	0.87
5695	Ashfield 006	80	96.4	0.21	0.15	0.17	0.82	0.66	1.00	0.82	0.74	0.91	0.92
5696	Ashfield 007	76	77.9	0.32	0.19	0.17	0.97	0.79	1.18	0.97	0.88	1.08	0.86
5697	Ashfield 008	73	59.9	0.47	0.09	0.28	1.24	0.99	1.52	1.29	1.14	1.44	0.94
5698	Ashfield 009	69	84.8	0.21	0.19	0.18	0.81	0.65	0.98	0.80	0.72	0.88	0.86
5699	Ashfield 010	74	68.2	0.20	0.12	0.18	1.02	0.82	1.25	0.94	0.85	1.08	1.03
5700	Ashfield 011	42	63.7	0.21	0.13	0.16	0.71	0.54	0.90	0.77	0.67	0.84	1.00
5701	Ashfield 012	34	57.4	0.17	0.13	0.17	0.66	0.49	0.85	0.71	0.61	0.78	0.96
5702	Ashfield 013	99	87.8	0.20	0.18	0.21	1.07	0.88	1.27	0.95	0.87	1.13	1.12
5703	Ashfield 014	106	90.8	0.32	0.09	0.34	1.17	0.97	1.39	1.18	1.05	1.33	0.88
5704	Ashfield 015	40	72.1	0.14	0.18	0.13	0.60	0.47	0.76	0.65	0.56	0.72	0.98
5705	Ashfield 016	91	70.8	0.41	0.20	0.22	1.26	1.04	1.52	1.23	1.11	1.38	0.92
MSOA Code	MSOA Name	Codes of Adjacent MSOAs											
5690	Ashfield 001	3956,3957,5691,5692,5693,5694,5751,5755											
5691	Ashfield 002	5690,5693,5694											
5692	Ashfield 003	3957,3959,5690,5694,5695,5698											

(Table 6). Continued.

5693	Ashfield 004	5690,5691,5694,5696,5755,5758
5694	Ashfield 005	5690,5691,5692,5693,5695,5696,5698
5695	Ashfield 006	5692,5694,5698
5696	Ashfield 007	5693,5694,5697,5698,5732,5758,5764
5697	Ashfield 008	5696,5698
5698	Ashfield 009	3959,5692,5694,5695,5696,5697,5699,5700,5732
5699	Ashfield 010	5698,5700,5701,5704,5719,5732
5700	Ashfield 011	3937,3959,5698,5699,5701
5701	Ashfield 012	3937,3939,5699,5700,5719
5702	Ashfield 013	2791,5703,5704,5705,5732,6693
5703	Ashfield 014	5702,5704,5705
5704	Ashfield 015	5699,5702,5703,5705,5719,5721,5722,5732
5705	Ashfield 016	2791,2793,5702,5703,5704,5722,6764

In sub-regions where many areas have high central risk, borrowing strength at the lower quantiles under the HQR model may result in more areas classified as high risk. Additionally borrowing strength at both low and high quantiles will produce narrower 95% risk intervals. This pattern (more high risk areas and narrower risk intervals under quantile regression) occurs especially within cities such as Birmingham, Bristol, Nottingham, Sheffield and Leeds.

Another illustration is provided by the MSOAs in the town of Ashfield (see Table **6**), intermediate between Sheffield and Nottingham. Table **6** shows predictors and risk intervals under BYM and HQR models, and outlier indicators expressed in the normalized form $W_{\alpha i} / \overline{W_{\alpha}}$. Apparent are narrower risk intervals under quantile regression. Regarding classification as high risk, an example is Ashfield003 with a relative 95% risk interval (0.86, 1.32) under the BYM model, but (1.01, 1.30) from quantile regressions at $\alpha = 0.025$ and $\alpha = 0.975$. It can be seen that this area has neighbours within the city with estimated $R_{0.025,i}$ above 1 (especially Ashfield005), and that this area has low rurality, a factor that also affects quantile regression estimates of $R_{0.025,i}$.

Narrower risk intervals do not, however, apply to areas with high outlier indicators (see Table **7**, where NA for the first area indicates that Y_i is unavailable, and outlier indicators at $\alpha = 0.50$ are normalized). High outliers are typically associated with discrepancies between the area's observed MLE relative risk Y_i / E_i, and the MLE relative risk in the area's locality N_i. They may also be associated with discrepancies between observed risk and the area's predictor profile (e.g. Newcastle upon Tyne 013). Median quantile regression estimators shrink the modelled relative risk closer to the neighbourhood risk level than the BYM regression. However, risk intervals are wider under median quantile regression.

7. CONCLUSIONS

In this paper, a model for quantile regression within a hierarchical framework is proposed for overdispersed area counts of disease, mortality, or hospitalisations. This technique has the advantage that a profile of the relative outcome risk across quantiles can be obtained, including estimates of uncertainty (e.g. the uncertainty attaching to 2.5% or 5% relative risk quantiles).

An assessment of how spatial patterns of relative morbidity risk (especially elevated levels) may be affected by using quantile regression involves self-harm data for English MSOAs. It was shown that impact of area predictors (deprivation, rurality,

Table 7: Outlier Areas

	Self-Harm									
	Count	Expected	Area Predictors			MLE		BYM, Estimated RR		
Area Name	Y	E	IMD	Rurality	SF	RR	Spatial Lag RR	Mean	2.5%	97.5%
Chelmsford 001	NA	66.2	0.16	0.34	0.14	NA	0.60	0.30	0.21	0.40
Leeds 006	91	66.6	0.06	0.45	0.13	1.37	0.59	1.10	0.89	1.33
Gloucester 002	431	94.5	0.31	0.17	0.44	4.56	1.80	4.37	3.97	4.79
Newcastle upon Tyne 013	47	165.5	0.04	0.09	0.74	0.28	1.13	0.37	0.30	0.46
Wokingham 009	89	96.8	0.07	0.30	0.16	0.92	0.30	0.78	0.63	0.95
South Staffordshire 006	117	87.5	0.11	0.23	0.19	1.34	0.63	1.17	0.97	1.39
Crawley 004	276	107.5	0.19	0.12	0.34	2.57	1.02	2.44	2.16	2.73
Chorley 007	130	79.9	0.12	0.29	0.24	1.63	0.73	1.44	1.21	1.69
Birmingham 096	89	246.8	0.25	0.09	0.84	0.36	0.88	0.41	0.34	0.49
Cambridge 007	133	235.4	0.07	0.14	0.56	0.56	1.45	0.63	0.53	0.73

	HQR (Median) Estimated RR					HQR Spatial Effects at α		
	Mean	2.5%	97.5%	W (Median QR)	Spatial lag in RR (Median QR)	0.025	0.5	0.975
Chelmsford 001	0.50	0.35	0.71	4.70	0.61	-0.54	-0.20	-0.18
Leeds 006	0.61	0.45	0.81	3.78	0.64	0.29	0.29	0.69
Gloucester 002	2.27	1.58	3.20	3.70	1.59	0.73	0.74	1.29
Newcastle upon Tyne 013	0.62	0.44	0.83	3.25	1.18	-0.51	-0.21	-0.20
Wokingham 009	0.48	0.32	0.68	3.08	0.35	-0.15	-0.10	0.29
South Staffordshire 006	0.71	0.49	0.99	3.07	0.67	0.10	0.13	0.48
Crawley 004	1.47	1.00	2.09	3.00	1.02	0.55	0.54	0.95
Chorley 007	0.89	0.65	1.19	2.99	0.82	0.32	0.37	0.71
Birmingham 096	0.70	0.44	1.08	2.99	0.93	-0.92	-0.57	-0.54
Cambridge 007	1.04	0.74	1.43	2.97	1.45	0.02	0.35	0.40

fragmentation) varies between quantiles. Hierarchical quantile regression generally produced narrower risk intervals, except for outlier areas, and led to a higher number of areas being classed as high risk, although there was concordance in risk classification with the results of the BYM model [3].

REFERENCES

[1] McDaid D, Bonin E, Park A, Hegerl U, Arensman E, Kopp M, Gusmao R. Making the case for investing in suicide prevention interventions: estimating the economic impact of suicide and non-fatal self harm events. Injury Prevention 2010; 16(Suppl 1): A257-8.

[2] Wakefield J. Disease mapping and spatial regression with count data. Biostatistics 2006; 8(2): 158-83. https://doi.org/10.1093/biostatistics/kxl008

[3] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 1991; 43(1): 1-20. https://doi.org/10.1007/BF00116466

[4] Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Statistical Methods in Medical Research 2005; 14(1): 35-59. https://doi.org/10.1191/0962280205sm388oa

[5] Richardson S, Thomson A, Best N, Elliott P. Interpreting posterior relative risk estimates in disease-mapping studies. Environmental Health Perspectives 2004; 112(9): 1016. https://doi.org/10.1289/ehp.6740

[6] Koenker R, Hallock K. Quantile regression: An introduction. Journal of Economic Perspectives 2001; 15(4): 43-56. https://doi.org/10.1257/jep.15.4.143

[7] Yu K, Moyeed RA. Bayesian quantile regression. Statistics & Probability Letters 2001; 54(4): 437-47. https://doi.org/10.1016/S0167-7152(01)00124-9

[8] Tsionas EG. Bayesian quantile inference. Journal of Statistical Computation and Simulation 2003; 73(9): 659-74. https://doi.org/10.1080/0094965031000064463

[9] Machado JA, Silva JS. Quantiles for counts. Journal of the American Statistical Association 2005; 100(472): 1226-37. https://doi.org/10.1198/016214505000000330

[10] Lee D, Neocleous T. Bayesian quantile regression for count data with application to environmental epidemiology. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2010; 59(5): 905-20. https://doi.org/10.1111/j.1467-9876.2010.00725.x

[11] Yue Y, Rue H. Bayesian inference for additive mixed quantile regression models. Computational Statistics & Data Analysis 2011; 55(1): 84-96. https://doi.org/10.1016/j.csda.2010.05.006

[12] Neelon B, Li F, Burgette LF, Benjamin Neelon SE. A spatiotemporal quantile regression model for emergency department expenditures. Statistics in Medicine 2015; 34(17): 2559-75. https://doi.org/10.1002/sim.6480

[13] Reich B, Fuentes M, Dunson DB. Bayesian spatial quantile regression. Journal of the American Statistical Association. 2011; 106(493): 6-20. https://doi.org/10.1198/jasa.2010.ap09237

[14] Dreassi E, Ranalli MG, Salvati N. Semiparametric M-quantile regression for count data. Statistical Methods in Medical Research 2014; 23(6): 591-610. https://doi.org/10.1177/0962280214536636

[15] Requia WJ, Koutrakis P, Roig HL, Adams MD, Santos CM. Association between vehicular emissions and cardiorespiratory disease risk in Brazil and its variation by spatial clustering of socio-economic factors. Environmental Research 2016; 150: 452-60. https://doi.org/10.1016/j.envres.2016.06.027

[16] Chiu C, Wen TH, Chien LC, Yu HL. A probabilistic spatial dengue fever risk assessment by a threshold-based-quantile regression method. PLOS One 2014; 9(10): e106334.

[17] O'Hara RB, Kotze DJ. Do not log-transform count data. Methods in Ecology and Evolution 2010; 1(2): 118-22. https://doi.org/10.1111/j.2041-210X.2010.00021.x

[18] Trinh G, Rungie C, Wright M, Driesener C, Dawes J. Predicting future purchases with the Poisson log-normal model. Marketing Letters 2014; 25(2): 219-34. https://doi.org/10.1007/s11002-013-9254-1

[19] Mahaki B, Mehrabi Y, Kavousi A, Mohammadian Y, Kargar M. Applying and comparing empirical and full Bayesian models in study of evaluating relative risk of suicide among counties of Ilam province. Journal of Education and Health Promotion 2015; 4: 50. https://doi.org/10.4103/2277-9531.162331

[20] Connolly SR, Dornelas M, Bellwood DR, Hughes TP. Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. Ecology 2009; 90(11): 3138-49. https://doi.org/10.1890/08-1832.1

[21] Al-Hamzawi R, Yu K, Pan J. Prior elicitation in Bayesian quantile regression for longitudinal data. J Biomet Biostat 2011; 2: 115.

[22] Ancelet S, Abellan JJ, Del Rio Vilas VJ, Birch C, Richardson S. Bayesian shared spatial-component models to combine and borrow strength across sparse disease surveillance sources. Biometrical Journal 2012; 54(3): 385-404. https://doi.org/10.1002/bimj.201000106

[23] Zhou J, Chang HH, Fuentes M. Estimating the health impact of climate change with calibrated climate model output. Journal of Agricultural, Biological, and Environmental Statistics 2012; 17(3): 377-394. https://doi.org/10.1007/s13253-012-0105-y

[24] Department of Communities and Local Government (DCLG) The English Indices of Deprivation. Office of National Statistics and DCLG, London 2015.

[25] O'Farrell IB, Corcoran P, Perry IJ. The area level association between suicide, deprivation, social fragmentation and population density in the Republic of Ireland: a national study. Social Psychiatry and Psychiatric Epidemiology 2016; 51(6): 839-47. https://doi.org/10.1007/s00127-016-1205-8

[26] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 2000; 10(4): 325-37. https://doi.org/10.1023/A:1008929526011

[27] Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 1998; 7(4): 434-55. https://doi.org/10.1080/10618600.1998.10474787

[28] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 2010; 11: 3571-94.

[29] Reed, C., Yu, K. A partially collapsed Gibbs sampler for Bayesian quantile regression. Brunel University, Dept of Mathematics Research Papers 2009.

[30] Berkhof J, Van Mechelen I, Hoijtink H. Posterior predictive checks: Principles and discussion. Computational Statistics 2000; 15(3): 337-54. https://doi.org/10.1007/s001800000038

[31] Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS Book: A Practical Introduction to Bayesian Analysis. CRC Press 2012.

Received on 17-04-2017 Accepted on 27-07-2017 Published on 04-08-2017

https://doi.org/10.6000/1929-6029.2017.06.03.1