Discussion: 16 June 2022 Peter J Diggle (CHICAS, Lancaster University)

Both of tonight's papers describe important work that has contributed and, with sufficient political will, can ans should continue to contribute to public health policy-setting and decision-making. The term "semi-mechanistic" could be applied to the modelling approaches used in both, the models being informed by a combination of data and context-specific scientific knowledge. I find it hard to disagree with this as a general philosophy but striking the right balance is a challenge in any specific application. At root, an epidemic is a realisation of a spatio-temporal point process, but is rarely modelled as such; for an exception see Neal and Roberts (2004). More often, context-specific knowledge feeds the temporal formulaton of the model as a set of discrete-time stochastic processees, with the role of space reduced to the indexing of a small number of discrete units (Norwegian counties in Storvik et al, unspecified in Bhatt et al). Discretising time is usually innocuous, both because of its simple geometry and the availabilty of data at a time-resolution fine enough to capture the dynamics of the process. Discretising space seems to me more questionable, certainly if the inferential goals include estimation of parmeters that relate to the effects of spatially varying risk-factors.

Taking the purely spatial case for illustration, let $Y_i : i = 1, ..., n$ be case-counts in a set of n areal units A_i that partition the study region, let $S(x)$ be the (unobserved) disease risk at location x, and write $S_i = \int_{A_i} S(x)dx$. A very widely used class of models for mapping the spatial variation in risk (Besag, York and Mollié, 1991) specifies the Y_i as conditionally independent Poisson variates with means

$$M_i = O_i S_i = O_i \exp\{d_i'\beta + W_i\}, \tag{1}$$

where the W_i form a Gaussian Markov random field, d_i is a vector of covariates and the offset, O_i, is the number of people at risk in A_i. Arguably a more natural model is a log-Gaussian Cox process, in which case the Y_i are again conditionally independent Poisson variates, but now with means

$$M_i = \int_{A_i} o(x)S(x)dx\} = \int_{A_i} o(x) \exp\{d(x)'\beta + W(x)\}, \tag{2}$$

where $o(x)$ is population density and $W(x)$ is a spatially continuous Gaussian process. Model (2) delivers the joint predictive distribution of the complete surface $S(x)$ (in practice, represented by a suitably fine pixel grid), from which the joint predictive distribution of the S_i follows immediately if that is what is required (Diggle, Moraga, Rowlingson and Taylor, 2013). Inferences drawn from (1) and (2) can differ materially if important covariates are available at finer spatial scales than the A_i as a consequence of the long-studied effects of ecological bias (Greenland and Morgenstern, 1990). For example, the current Covid-19 epidemic in England has typically been mapped at the level of England's 331 Lower Tier local Authorities (e.g. Riley et al, 2021) but potentially important covariates, including measures of deprivation and ethnicity, are available for each of the country's 34,753 Lower Super Output Areas (LSOAs); and they can vary substantially between LSOAs within the same LTLA.

Reproduction numbers, however defined, might similarly vary over much smaller spatial scales than the ones to which epidemic models are typically fitted, which leads me to wonder exactly what property of an epidemic a spatially agnostic R-number is estimating. Contrarily, I am surprised that the time-trajectories of estimated R-numbers in Figure 2 of Storvik et al show so much high-frequency variation, even allowing for the widths of their credible intervals. Do

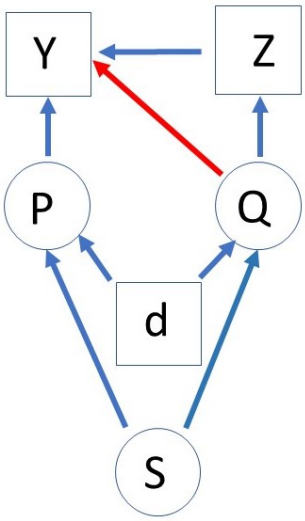

Figure 1: Causal diagram for informative non-response. Squares and circles represent observed and unobserved quantities, respectively. Observed quantities are covariates, d, number of responders, Z, and number of cases among responders, Y. Unobserved quantities are the response rate, Q, true prevalence, P, and a latent process, S, representing unobserved factors that potentially affect both Q and P. Non-response is uninformative if the red arrow is absent.

the dynamics of the epidemic really change so quickly? And if they do, how useful is an estimate of the current R-number as a guide to real-time public health decision-making?

Finally, I would like to comment on an issue that is perhaps beyond the scope of either paper, namely the reliability of the case-numbers that inform the models. A recent WHO report gives world-wide, country-level estimates of the very substantial under-counts in reported Covid-19 deaths (Knutson et al, 2022), and the problem can only be worse for case-counts of a disease with, thankfully, relatively low mortality. Nicholson et al (2022) show how to de-bias case-count data by jointly modelling routinely reported case-counts together with data from a randomised (hence unbiased) prevalence study, the REACT study (Riley et al, 2021). REACT was an England-wide repeated cross-sectional study of the spatio-temporal variation in Covid-19 prevalence over the course of the epidemic, based on Covid-19 positive/negative test results from samples, each of approximately 100,000 individuals drawn, on each of 19 rounds over almost two years, as geographically stratified random samples of the NHS-registered population. The response rate was of the order of 20 to 25 per cent, which the REACT analysis accommodated by adjusting for imbalance in the responders with respect to potemtial risk-factors. Nevertheless, this leaves a niggling doubt that in studies of this kind responders within a risk-group may not be representative of the whole group. In the setting of longitudinal studes, this corresponds to the problem of informative missingness, on which topic many papers and at least one book (Daniels and Hogan, 2008) have been written. The prevalence survey counterpart is harder to address because, unlike in longitudinal studies, there no independent replication. Figure 1 outlines a possible solution. The predictive target is the true prevalence, P, but in the case of informative non-response the empirical prevalence depends on both P and Q.

References

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Daniels, M.J. and Hogan, J.W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton: Chapman and Hall/CRC

Diggle, P.J., Moraga, P., Rowlingson, B. and Taylor, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, **28**, 542–563.

Greenland, S. and Morgenstern, H. (1990). Ecological bias, confounding and effect modification. *International Journal of Epidemiology*, **18**, 269–74.

Knutson, V., Aleshin-Guendel, S., Karlinsky, A., Msemburi, W. and Wakefield, J. (2022). Estimating country-specific excess mortality during the Covid-19 epidemic. *Annals of Applied Statistics* (to appear)

Neal, P.J. and Roberts, G.O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, **5**, 249–261.

Nicholson, G., Lehmann, B., Padellini, T., Pouwels, K.B., Jersakova, R., Lomax, J., King, R.E., Mallon, A-M., Diggle, P.J., Richardson, S., Blangiardo, M.and Holmes, C. (2022). Local prevalence of transmissible SARS-CoV-2 infection: an integrative causal model for debiasing fine-scale targeted testing data. *Nature Microbiology*, **7**, 97–107.