Appendix: an example of a surveillance system combining designed and routinely collected data to reconstruct the underlying space-time incidence of a disease

Peter J Diggle and Sylvia Richardson

May 27, 2020

Here, we describe a possible joint modelling and analysis framework for synthesizing information from designed and routinely collected records in order to reconstruct the underlying space-time incidence of a disease. For ease of exposition we ignore various potential complications including imperfect sensitivity and specificity of the diagnostic test used for the designed data-source, covariate information on place or person, local or national policy interventions. We assume that the routine records have lower sensitivity and specificity than the diagnostic test used in the designed study, and that additional data has been collected enabling calibration of the relative sensitivity and specificity of the routine data with respect to the diagnostic test.

1. Data

We suppose that our region of interest is partitioned into N small-areas, and that in each week t we have access to two kinds of incidence data as follows.

1) Designed data $(x_{it}, m_{it}, Y_{it}) : i = 1, ..., N$, where x denotes the geographical location, m the number of individuals tested by the diagnostic test, and Y the number of positive test reports. Each x_i corresponds to one of the N small-areas, for example the small-areas might be LSOA and the locations their population-weighted centroids.

2) Routinely collected data $Z_{jt} : j = 1, ..., N$, where Z denotes the number of routinely reported new "cases", for example calls to NHS111 classified as "possible Covid-19".

2. Models

2.1 Designed data

For data of the kind described, the natural sampling distribution is binomial, with probabilities P_{it} and denominators m_{it}. We assume that, conditional on a zero-mean, latent spatio-temporal Gaussian process $S(x, t)$, the counts Y_{it} are independent, binomially distributed with

$$\log\{P_{it}/(1 - P_{it}\} = \alpha + S(x_i, t). \tag{1}$$

One candidate model for $S(x, t)$ is a spatially continuous stationary process with variance σ^2 and separable spatio-temporal correlation $\mathrm{Corr}\{S(x, t), S(x', t')\} = \rho(||x - x'||)\phi(|t - t'|)$, where each of $\rho(\cdot)$ and $\phi(\cdot)$ is a member of the Matérn family (Matérn, 1960). A second possibility is a Gaussian Markov Random Field (Rue and Held, 2005).

2.2 Routinely collected data

Here, the natural sampling distribution is again binomial, but now with denominators so large that a Poisson approximation would give an excellent approximation. Writing M_j for the size of the population at risk in the jth small-area we assume that, conditional on a second zero-mean, latent spatio-temporal Gaussian process $V(x, t)$, the counts Z_{it} are independent, Poisson-distributed with means μ_{jt}, where

$$\log(\mu_{jt}) = \beta + V(x_i, t) \tag{2}$$

2.3 Linking the two data-sources

Linkage requires a calibration relationship between $S(x,t)$ and $W(x,t)$. Acknowledging the precedence of the designed data, we assume that

$$V(x,t) = \gamma S(x,t) + Z(x,t), \tag{3}$$

where $Z(x,t)$ is white noise, variance τ^2. Information on γ will be derived by modelling the additional data relating the diagnostic test results to the routine records. This data will be modelled jointly with (2) and (3).

3. Analysis

Monte Carlo maximum likelihood or Bayesian inference for the model defined by equations (1), (2) and (3) would require some modification of existing `R` packages, for example the `PrevMap` package (Giorgi and Diggle, 2017). The inferential target is the joint predictive distribution of $S(x_j,t) : j = 1, ..., N$, given data from weeks up to and including week t. Depending on the strength and scale of the estimated temporal correlation structure of $S(x,t)$, reliable predictions in week t could be obtained using data from a rolling time-window, to circumvent restricted storage and/or computing time; for an example, see Diggle, Rowingson and Su (2005).

References

Diggle, P.J., Rowingson, B. and Su, T-L. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–34.

Giorgi, E. and Diggle, P.J. (2017). PrevMap: an R package for prevalence mapping. *Journal of Statistical Software*, **78**, 1-29, doi:10.18637/jss.v078.i08

Matérn, B. (1960). *Spatial variation.* Meddelanden fran Statens Skogsforsknings institut, Stockholm. Band 49, number 5.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* London: CRC Press.