Investigating the UK measles data set between 1944 to 1962 using the hhh4 model

Kajal Dodhia
Supervisor: Jordan J Hood

STOR-i , Lancaster University

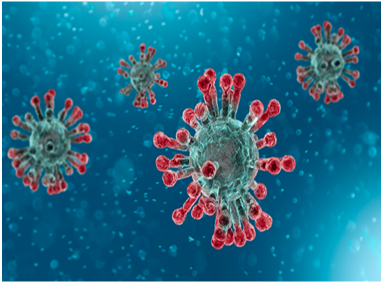
August 25, 2022

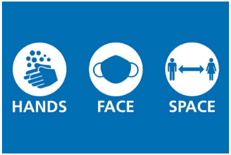
1. Background
2. Explanatory Analysis
3. Formal Analysis
4. Conclusions
5. References

1. Background
2. Explanatory Analysis
3. Formal Analysis
4. Conclusions
5. References

Introduction

SARS-CoV-2 (Covid-19) virus has had a significant impact on our lives since January 2020.





Aims

- To analyse the prevaccine UK measles data set from 1944-1962.
- To explore adding covariates and random effects to the basic model in order to find the best fitting model and look at prediction of future outbreaks.
- Analysis was carried out in R using the hhh4 package.

Basic Theory

Generalised Linear Models (GLM)

Let Y_i be independent responses from an exponential family distribution in canonical form and $\mu_i = \boldsymbol{x_i^T}\beta$ for $i, \cdots n$. A generalised linear model is a model of the form $g(\mu_i) = \boldsymbol{x_i^T}\beta$ where β is a p - dimensional parameter vector, $\boldsymbol{x_i^T}$ is the ith row of the design matrix $\boldsymbol{X}$, and $g()$ is a monotonic, differentiable function called the link function.

See [Berridge et al., 2011] for more information.

Generalised linear models extend the normal linear model by:

- allowing the response to follow distributions other than the normal distribution.
- setting a more general function g of the mean equal to the linear predictor, so that instead of $\mu = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta}$ we have $g(\mu) = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta}$.

Generalised linear mixed models extend the generalised linear model by:

- the linear predictor contains random effects in addition to the usual fixed effects.
- they inherit from GLMs the idea of extending linear mixed models to non-normal and correlated data.
- responses have equal variance conditional on the random effects and random effects are normally distributed, independent, zero mean and (not necessarily same variance).

1 Background

2 Explanatory Analysis

3 Formal Analysis

4 Conclusions

5 References

First look

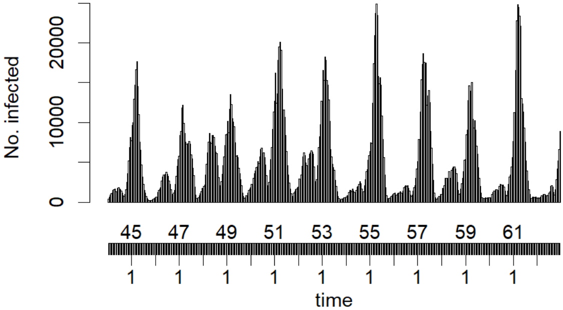


Figure 1: Total number of infections across 1944-1962, measured fortnightly

Basic Model

Endemic-epidemic multivariate time-series model

An endemic-epidemic multivariate time-series model for infectious disease counts Y_{it} from units $i = 1, \cdots, 60$ during periods $t = 1, \cdots, 493$, where i denotes city and t denotes fortnightly time. The hhh4 model assumes that $Y_{it} | \mathcal{F}_{t-1} \sim \mathcal{NB}(\mu_{it}, \psi)$, where

$$\mu_{it} = e_i \nu_t + \lambda Y_{i,t-1} + \phi \sum_{j \neq i} w_{ji} Y_{i,t-1},$$

$$\log(\nu_t) = \alpha^{(\nu)} + \beta_t t + \gamma \sin(\omega t) + \delta \cos(\omega t),$$

and overdispersion parameter $\psi_i > 0$ such that the conditional variance of Y_{it} is $\mu_{it}(1 + \psi_i \mu_{it})$. The link function is $\log(\mu_i)$.

Basic Model explained

Model components

1. **Endemic log-linear predictor ν_t:**

$$\log(\nu_t) = \alpha^{(\nu)} + \beta_t t + \gamma \sin(\omega t) + \delta \cos(\omega t)$$

- Temporal variation of disease incidence incorporates an overall trend and a sinusoidal wave of frequency $\omega = \frac{2\pi}{26}$.
- Population fraction as multiplicative offset e_i.

2. **Epidemic component:**

$$\mu_{it} = e_i \nu_t + \lambda Y_{i,t-1} + \phi \sum_{j \neq i} w_{ji} Y_{i,t-1},$$

- Autoregressive: $\lambda = \exp(\alpha)^{\lambda}$. Spatio-temporal: $\phi = \exp(\alpha)^{\phi}$.
- These are assumed homogeneous across cities and constant over time and in this model the epidemic can only arrive from directly adjacent cities.

Fitting the basic model to our data

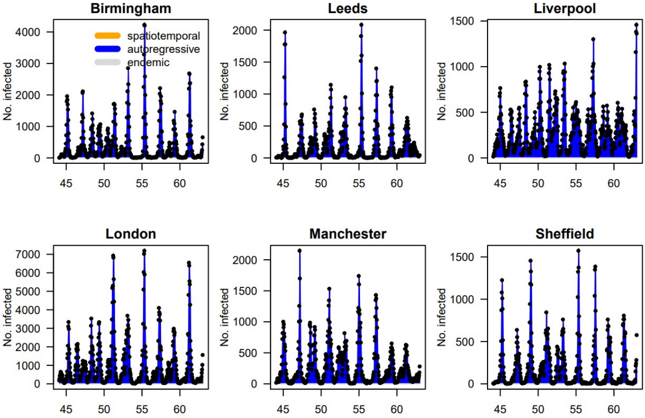


Figure 2: Fitted components in the initial model for the cities with more than 80,000 total infections. Dots are drawn for positive weekly counts.

Endemic Mean

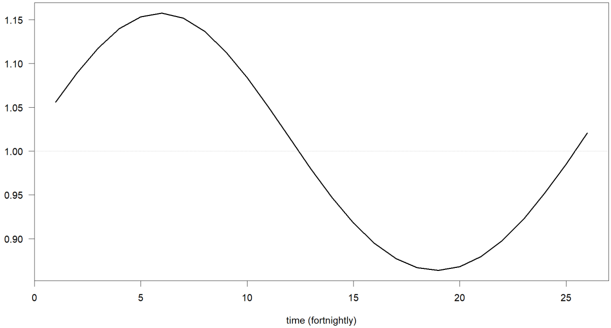


Figure 3: Estimated multiplicative effect of seasonality on the endemic mean

The multiplicative effect of seasonality increases as winter approaches and starts to decrease towards the end of winter in February.

Adding covariates

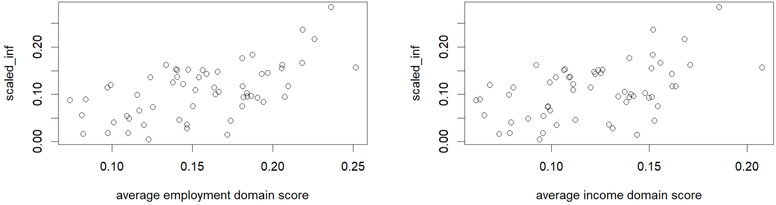


Figure 4: Scatter plots showing the relationship between scaled infections average employment domain score and average income domain score respectively. Correlation coefficients: 0.5854129 and 0.4786521

1. Background
2. Explanatory Analysis
3. Formal Analysis
4. Conclusions
5. References

Average of Employment Domain Score

To choose between endemic and/or autoregressive effects, and multiplicative offset vs. covariate modeling, by performing AIC-based model selection.

AIC Model Selection for employment

	df	AIC
Scovar|unchanged	8	218296.5
Scovar|Scovar	9	218298.2
Soffset|Scovar	8	218392.0
Soffset|unchanged	7	218395.8
unchanged|Scovar	8	218749.8
unchanged|unchanged	7	218770.4
Soffset|Soffset	7	221802.0
Scovar|Soffset	8	221803.3
unchanged|Soffset	7	221899.2

Average of Income Domain Score

AIC Model Selection for income

	df	AIC
Scovar|unchanged	9	218296.5
Scovar|Scovar	10	218298.2
Soffset|unchanged	8	218392.0
Soffset|Scovar	9	218395.8
unchanged|unchanged	8	218749.8
unchanged|Scovar	9	218770.4
Soffset|Soffset	8	221802.0
Scovar|Soffset	9	221803.3
unchanged|Soffset	8	221899.2

Leave the autoregressive component unchanged and add both employment and income to the endemic predictor in model.

Random effects? - GLMM

- Cities exhibit heterogeneous incidence levels not explained by observed covariates, and especially if the number of cities is large (60).
- An example of unobserved heterogeneity in the measles data set is under-reporting.
- Allowing for city-specific intercepts in the endemic or epidemic components is expected to improve the model fit.
- Disadvantages: runtime increases considerably and random effects invalidate simple AIC based model comparisons. See [Czado et al., 2009].

Updated model with covariates and random effects

The Final Model

The final model incorporates the covariates for average of employment and income domain score and independent random effects in all three components.

$$\alpha_i^{(v)} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha^{(v)}, \sigma_v^2),\ \ \alpha_i^{(\lambda)} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha^{(\lambda)}, \sigma_\lambda^2) \text{ and } \alpha_i^{(\phi)} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha^{(\phi)}, \sigma_\phi^2),$$

$$\mu_{it} = e_i \nu_t + \lambda Y_{i,t-1} + \phi \sum_{j \neq i} w_{ji} Y_{i,t-1},$$

$$\log(\nu_t) = \alpha_i^{(\nu)} + \beta_t t + \gamma \sin(\omega t) + \delta \cos(\omega t) + \beta_E \log(E_i) + \beta_i \log(I_i).$$

Final Model Updated graphs

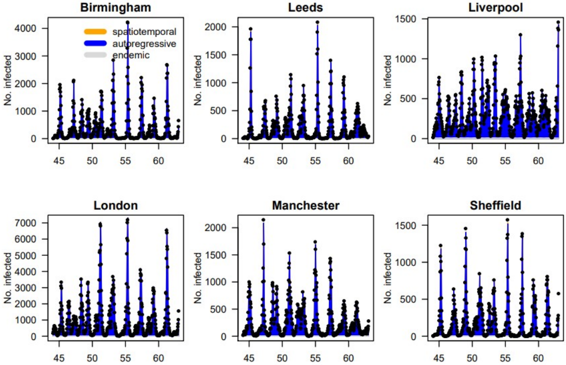


Figure 5: Fitted components in the random effects model for the cities with more than 80,000 total infections. Dots are drawn for positive weekly counts.

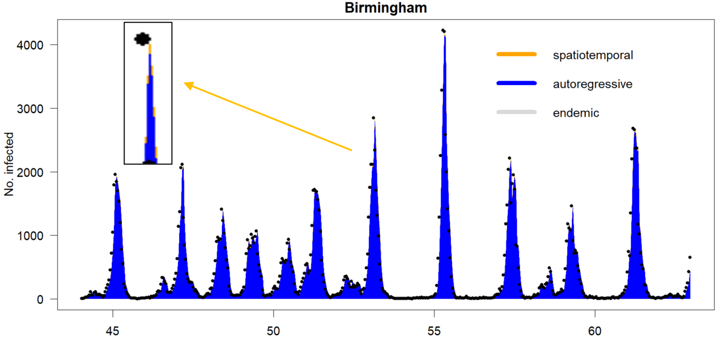


Figure 6: Fitted components in the random effects model for Birmingham

There is a slight increase in the proportion of fitted mean captured by the spatio-temporal component for Birmingham.

Predicting Model Assessment - Test Period: 1962

Scoring Methods

- Squared error score (ses)
- Logarithmic score (logs)
- Ranked probability score (rps)
- Dawid-Sebastiani score (dss)

➜ Lower scores correspond to better predictions.
See review [Gneiting and Katzfuss, 2014].

Goodness of fit test and true two week ahead prediction

Goodness of fit assessment

	logs	rps	dds	ses
measlesFit_basic	3.067134	8.858451	6.281844	1065.896
measlesFit_emp	3.076321	8.818000	7.708370	1085.562
measlesFit_emp+inc	3.070956	8.822474	7.335209	1089.699
measlesFit_final	3.005305	8.702937	5.696284	1095.460

The final model gave the smallest mean score for most of the scoring methods, hence it is the best fitting model.

(True two week ahead) prediction

	logs	rps	dds	ses
measlesFit_basic	3.070532	8.861904	6.350042	1073.087
measlesFit_emp	3.081296	8.822825	7.985044	1093.475
measlesFit_emp+inc	3.076591	8.826998	7.659704	1097.729
measlesFit_final	3.029018	8.727300	6.409551	1106.274

The most parsimonious model is the final model which gives the best two-week-ahead predictions in terms of overall mean scores.

Predictive Model Assessment

Paired t-test - for predictive performance

$\mathbf{H_0}$: The difference between the mean scores of the basic model and final model are zero,

$\mathbf{H_1}$: The difference between the mean scores of the basic model and final model are not equal

P value: 0.00052.

Calibration Test:

$\mathbf{H_0}$: The model is well calibrated,

$\mathbf{H_1}$: The model is miscalibrated

P value: $2.2e^{-16}$.

See more: [Wei and Held, 2014]

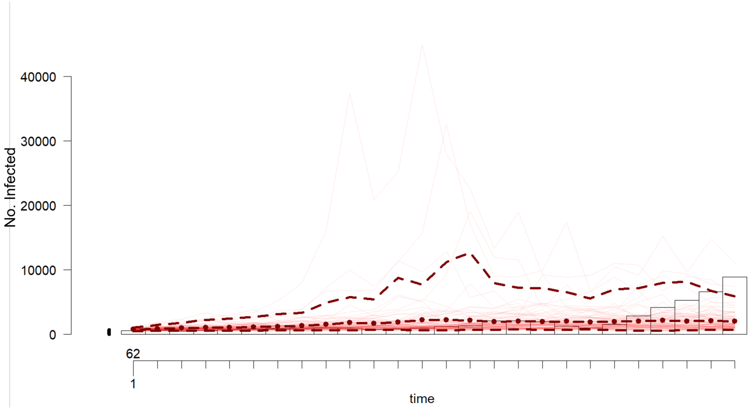


Figure 7: Simulation-based forecast of 1962 starting from the last second last week in 1961 (vertical bar on the left), showing the counts aggregated over all cities. The fortnightly mean of the simulations is represented by dots, and the dashed lines correspond to the pointwise 2.5% and 97.5% quantiles. The actually observed counts are shown in the background.

1. Background
2. Explanatory Analysis
3. Formal Analysis
4. **Conclusions**
5. References

Conclusions

- The final model which incorporates the covariates for average of employment and income domain score and random effects is the best fit for the measles data set.
- However, in predicting, it does not capture the large number of cases for the year 1962.
- The largest portion of the fitted mean results from the within-city autoregressive component, a very small spatio-temporal and almost negligible endemic component to the data.

Further research

- **Look at specific cities more closely to see if there is a seasonal component to the data** - find out why there is negligible endemic component
- **Look at other data sets with more variables that are relevant to the recent COVID-19 pandemic such as:**
  - effect of policy containment measures that limit social mobility
  - number of people vaccinated
  - long-range transmission of cases
- **Check if there is any evidence for residual spatial or temporal dependence.** Further model generalizations may be useful, but may require a Bayesian approach and more advanced MCMC techniques for statistical inference.
  Example: Allowing λ to change over time using Bayesian change-point model for time changing situations.

Thank you for listening!

Are there any questions?

1. Background
2. Explanatory Analysis
3. Formal Analysis
4. Conclusions
5. References

[Berridge et al., 2011] Berridge, D., Crouchley, R., and Grose, D. (2011).
Multivariate generalized linear mixed models using R.
CRC Press Boca Raton, FL.

[Czado et al., 2009] Czado, C., Gneiting, T., and Held, L. (2009).
Predictive model assessment for count data.
Biometrics, 65(4):1254–1261.

[Gneiting and Katzfuss, 2014] Gneiting, T. and Katzfuss, M. (2014).
Probabilistic forecasting.
Annual Review of Statistics and Its Application, 1(1):125–151.

[Held et al., 2005] Held, L., Höhle, M., and Hofmann, M. (2005).
A statistical framework for the analysis of multivariate infectious disease surveillance counts.
Statistical Modelling, 5(3):187–199.

[Meyer et al., 2017] Meyer, S., Held, L., and Höhle, M. (2017).
Spatio-temporal analysis of epidemic phenomena using the r package surveillance.
Journal of Statistical Software, 77(11):155.

[Wei and Held, 2014] Wei, W. and Held, L. (2014).
Calibration tests for count data.
TEST, 23(4):787–805.