Investigating Measles Dynamics in the Pre-Vaccination Era: A POMP Model Approach

Jiayi Xu

A Thesis Submitted in Partial Fulfillment
of Bachelor of Science in Honors Data Science
at the University of Michigan.

Supervised by
Professor Edward Ionides
Aaron Abkemeier

Department of Statistics
University of Michigan - Ann Arbor
United States of America
Mar 2024

Acknowledgements

I would like to thank Professor Edward Ionides of the University of Michigan's Statistics Department for his insightful advice and guidance throughout the research project. From data analysis to thesis writing, Professor Ionides' suggestions were invaluable and enabled my growth as a researcher. I am also grateful to Aaron Abkemeier, who served as my mentor during this project, for guiding and offering me detailed suggestions through the research process and introducing me to performing statistical analysis through the measlespkg. I would also like to thank our research group and our measles research subgroup for engaging discussions that truly inspired my passion for epidemiology and mechanistic models.

Contents

1 Introduction **3**

2 Data **4**

3 Method **5**
3.1 Partially Observed Markov Process 5
3.2 Panel Iterated Filtering (PIF) 6
3.3 Initial Conditions and Parameter Descriptions 7
3.4 Model 1 (Unit Specific) 9
3.5 Model 2 ($\log(\iota)$ varying linearly with log(Population Size N)) 9
3.6 Model 3 (shared parameter model with unit specific ι) 10

4 Results **10**

5 Discussion **21**

1 Introduction

Within the field of epidemiology, there has been a great interest in modeling essential mechanisms that characterize the development of an infectious disease (Keeling and Ross, 2008). The potential for devising corresponding policies that could minimize and control the impact of infectious diseases leads to the call for developing statistical methods customized for time series data. Multiple approaches have been undertaken by statisticians to model the spread of infectious diseases, such as measles. Measles disease dynamics have been an active area of research due to the high quality of readily available data and life-long immunity post-infection, which made it easier to model (Glass et al., 2003). Its unique characteristics enable researchers to test newly developed models against previous studies to gain insights into the modeling ability before applying them to model other disease outbreaks (Gibson and Renshaw, 2001). From nonlinear stochastic dynamical models to Bayesian approaches, each model offers unique insights into the disease, but with trade-offs that sacrifice modeling accuracy (Liu and West, 2001). For instance, several methods may work only under restricted assumptions and may not be applicable across various scenarios. The complex nature of disease outbreaks and the intrinsic unpredictability of infectious disease results in the challenges faced in the modeling process (Gibson and Renshaw, 2001).

Apart from the challenges mentioned above, the Partially Observed Markov Process (POMP) model aims to overcome these obstacles (Ionides et al., 2006). The unique characteristic of the algorithm for the POMP model lies in its plug-and-play property, which lifts the requirement of explicit expressions for transition probabilities and instead obtains results from dynamic model simulation (He et al., 2010). Building upon the development of POMP, the POMP package enables the implementation of such a method in a statistical analysis setting (King et al., 2015). The extension of the method to panelPOMP allows for constructing POMP models for each individual unit and incorporates shared parameters to capture disease characteristics (Breto et al., 2022).

He et al. (2010) investigated how this continuous time model, POMP, is applicable to varying population sizes while accounting for the effect of environmental and demographic stochasticity by analyzing measles outbreak data. He et al. (2010) further indicated that the data suggests little deviation from mass-action transmission and unveils the relationship between heterogeneity in transmission and population size, leading to the conclusion that extra-demographic variability played an important role in the measles outbreak. Nevertheless, the observation that γ (recovery rate for infected individuals) varies with population size, which was modeled through log relationships and the observation that Latent period and infectious period have a linear relationship with population size, both required additional investigation and further explanation to address interpretation concerns from a biological context.

Following the case study, an R package, measlespkg (available at https://github.com/AJAbkemeier/m was developed to facilitating the fitting process for PanelPOMP Models, which is used in this research paper for fitting POMP models. The Monte Carlo profile confidence intervals (MCAP) provides both theoretical and computation foundation for testing whether a parameter specified in the model proposed in the He et al.(2010) paper should be shared or unit-specific (Ionides et al., 2017). Having the parameter being shared aligns with the biological interpretation that some inherent characteristics of the

disease are common across populations with varied sizes. This study is thus prompted to have a focus on fitting shared parameter models; additionally, the MCAP extension under development could be taken to our advantage of things we learn about fitting shared parameter models. By incorporating the shared parameters, we hope to explore the possibility of a model that could better capture the underlying dynamics of disease spread. We further hypothesize that the MCAP approach could be adapted to include multiple shared parameters, which remains to be tested.

The insights and challenges from previous studies motivate this research project to provide empirical support for prior conclusions and newly proposed shared parameter modeling approaches. Thus, we apply the POMP model to the England and Wales dataset and extend the study by He et al. (2010) to include an additional 200 cities. Within the field of interest, though He et al. (2010) has conducted a case study on 20 cities, there has been no prior research done on a set of 200 cities using mechanistic model. Moreover, Becker and Grenfell (2017) suggested the challenge of using mechanistic model on large data sets by indicating that POMP may not be efficient in analyzing 100+ time series as it takes a long period of time for them to converge. Since the time needed for fitting the model scales linearly, as it took us 1.25 hours to fit one city, the model fitting process needs 5 days to complete the process of fitting 100 cities. Although the computational challenges are apparent, including 200 cities would allow us to explore measles disease dynamics across varied population sizes, investigate prior observations that required empirical evidence, and validate the application of the POMP model to large data sets. The research also attempts to investigate the relationship between different parameters and how parameter estimates vary with population size. Through this dataset analysis, we aim to demonstrate the limitations and capabilities of mechanistic models through empirical evidence.

2 Data

Since we aim to extend the scope of previous research to include more cities, we used the data from Korevaar et al. (2020) as well as the 20 UK cities data set from He et al. (2010). From a pool of 1402 cities in the England and Wales data set combining Korevaar et al. (2020) and He et al. (2010), we randomly selected 200 cities, which excluded the 20 cities used in He et al. (2010). The parameter estimates for the 20 cities presented in the He et al. (2010) paper were used to confirm the accuracy of the simulation and modeling process before proceeding to other cities. Random selection of the additional cities ensured we have good coverage of cities with varying population sizes and characteristics, where the population size range from 660 to 3,389,620. Compared to cities with a larger population size, small cities tend to have fewer observations and more fadeouts, which are more likely to result in weak identifiability for parameter estimates in the POMP model. The chosen dataset includes: weekly reported measles cases, birth rates, and yearly population data for each city covering the period from 1950 to 1964. We follow the approaches undertaken by He et al. (2010) and Fine and Clarkson (1982) to only include cases from 1950 and after to ensure the data were not disrupted by World War II and the introduction of the National Health Service. The selection of cities built a solid foundation for a comprehensive analysis of measles disease dynamics and the application of POMP models across a wide range of environments,

extending the scope of the original study. In Figure 1, we presented a visualization for the time series data for a selected 30 cities from the combined data set.

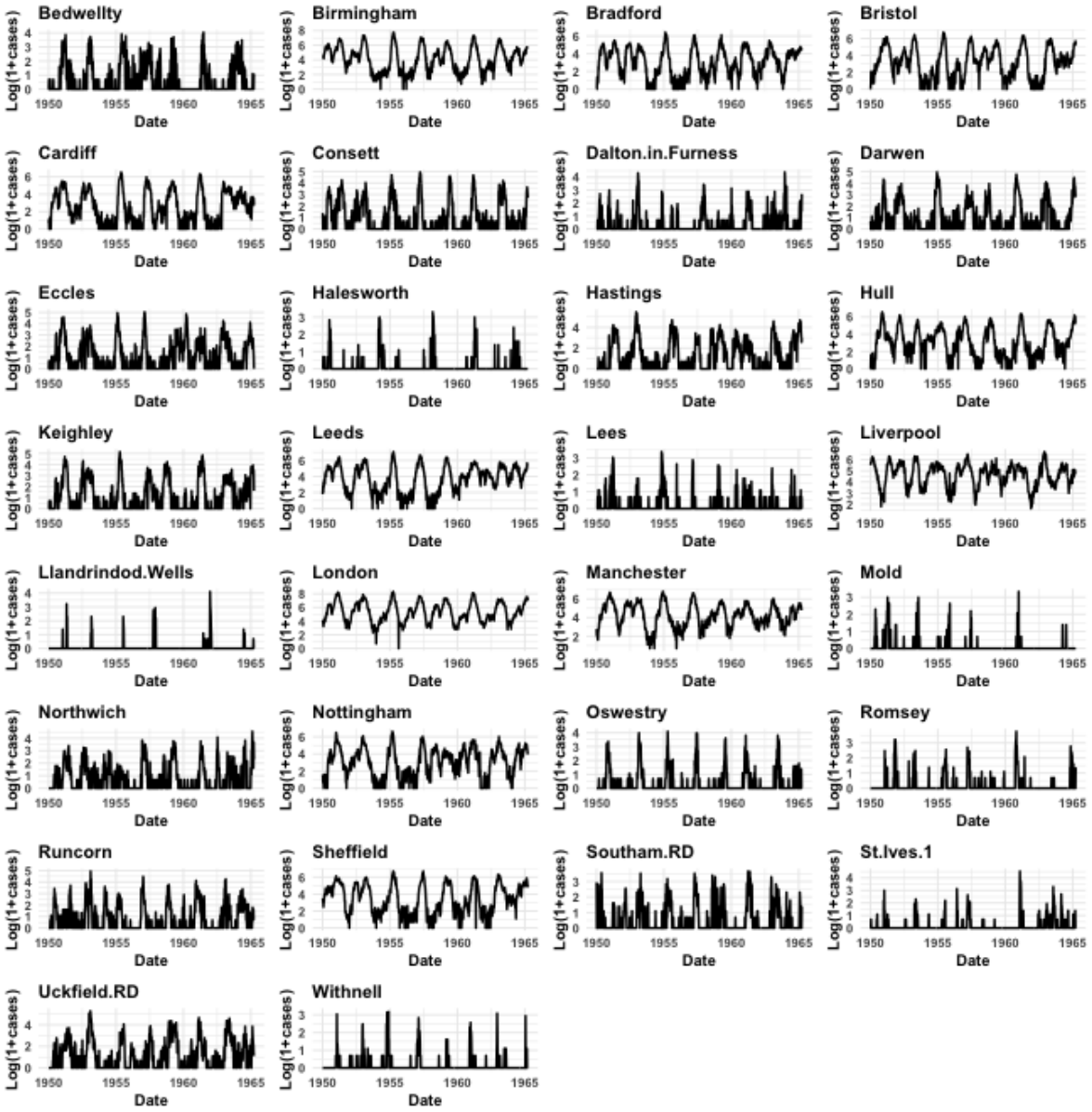

Figure 1: Time series for each selected 30 cities. (RD indicates rural areas outside of the cities)

3 Method

3.1 Partially Observed Markov Process

In epidemiological modeling, the Partially Observed Markov Process (POMP) models, also known as Hidden Markov Models or state-space models, are utilized to describe the dynamics of infectious disease transmission when the system is only partially observed.

These models are tailored for scenarios where direct observation of the system's state is obscured and must be inferred from incomplete and noisy data (Ionides et al., 2006).

Within the SEIR modeling framework for infectious disease modeling, the POMP model segments the population into compartments according to their infection status: Susceptible (S), Exposed (E), Infected (I), and Recovered (R). The count of individuals in each category at time t is denoted by $S(t)$, $E(t)$, $I(t)$, and $R(t)$, respectively. Transition between these states follows a Markov process $X_{1:N}$, reflecting the stochastic nature of disease progression.

The observable data, represented as $Y_{1:N}$ and derived from case reports, are the only source of information to infer the hidden states. Each observation Y_n is conditionally dependent only on the state X_n at that time and is conditionally independent from the other states given the current state. $S(t)$ is related to $X_{1:N}$ via $X_n = X(t_n)$ in that $S(t)$ is a part of each state $X(t_n)$ observed at specific times $t_1, t_2, \ldots, t_N$. Changes in $S(t)$ over time help inform the understanding of how the disease spreads and what interventions might be effective.

The mathematical formulation of a POMP model is defined by several densities. The transition density $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$ describes the probability of transitioning from one state to the next. The measurement density $f_{Y_n|X_n}(y_n|x_n;\theta)$ represents the probability of observing a data point given the current state. The initial density $f_{X_0}(x_0;\theta)$ specifies the starting point of the stochastic process. The likelihood function for the POMP model, which plays an important role in parameter estimation, is the product of these densities integrated over all possible state paths:

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y^*_{1:N};\theta) = \int f_{X_0}(x_0;\theta) \prod_{i=1}^{N} f_{Y_n|X_n}(y^*_n|x_n;\theta) f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta) dx_{0:N} \quad (1)$$

The PanelPOMP framework is a distinct adaptation of the POMP model, which is specialized in separating hidden and observable processes into different units. It also allows us to use shared parameters across different cities. In a dataset featuring panel structure with K discrete time series, there exist K individual units symbolized as $u_1, u_2, \ldots, u_K$. Each of these units is associated with a sequence of N_k periodic observations. These observations are encapsulated by a stochastic process $\{y_{k,1}, \ldots, y_{k,N_k}\}$ for each entity k, reflecting the systematic measurements over time. This property allows for processing each unit independently and creating individual POMP models for each unit. This framework is thus preferable for analyzing complex systems that consist of multiple independent units. In this case, PanelPomp is a crucial modeling framework for us as we have cities that are independent from one another and we model some parameters to be shared.

3.2 Panel Iterated Filtering (PIF)

The Iterated Filtering algorithm (IF2) is an iterative approach to parameter estimation for POMP models that employs a series of particle filters and a decreasing random walk variance to converge to the maximum likelihood estimates (Ionides et al., 2015). Building on IF2, the Panel Iterated Filtering (PIF) method is particularly adept at dealing with cases where the units are independent from one another and shared parameters could be incorporated into the model. The algorithm simulates the initial

state and transitions of the latent process and evaluates the likelihood of the observed data given these latent states. It iteratively refines the estimates of θ through a series of simulations and likelihood evaluations (Bretó et al., 2019). The algorithm could be summarized by the following pseudocode:

1. For each iteration m from 1 to M:
   (a) For each unit u from 1 to U:
      i. Perturb the initial parameter vector $\Theta_{u,0}^{(m)}$ using the perturbation density $h_{u,0}$.
      ii. Simulate the initial latent states $X_{u,0}^{(m)}$ from the initial state density $f_{X_{u,0}}$ with the perturbed parameters.
      iii. For each time point n from 1 to N_u:
         A. Perturb the parameters $\Theta_{u,n}^{(m)}$ for each particle using the perturbation density $h_{u,n}$ with diminishing variance.
         B. Predict the next states $X_{u,n}^{(m)}$ using the transition density $f_{X_{u,n}|X_{u,n-1}}$.
         C. Compute weights $w_{u,n}^{(m)}$ based on the measurement density $f_{Y_{u,n}|X_{u,n}}$ and the observed data.
         D. Resample the parameters and states according to the computed weights.
      iv. Update the final parameter vector for unit u in iteration m to the resampled parameters after the last observation.
   (b) Consolidate the final parameters for all units to define the parameter vector $\Theta^{(m)}$ for iteration m.
2. Yield the final parameter swarm $\{\Theta_j^{(M)}\}_{j=1}^{J}$ as the output.

The algorithm resembles the process of natural selection where parameter estimates evolve, and those fitting the data best survive and propagate. The decreasing variance of the parameter estimate perturbations automatically shrinks the scope of the search as the parameter space is explored so it can converge on the MLE. Like IF2, the algorithm benefits from the plug-and-play property, allowing for flexibility in modeling the state transitions without the need for explicit transition probabilities.

3.3 Initial Conditions and Parameter Descriptions

We adopt the standard susceptible-exposed-infectious-recovered (SEIR) framework to describe the transitions between the states of the disease, which is the same framework used in the He et al. (2010) paper. This model simulates a group of people, each of whom could be susceptible to the disease (S), exposed (E), infectious (I), and finally recovered (R), on each day (denoted as t) within the considered timeframe.

The number of individuals transitioning between states is modeled as follows, where we simplify the notation by omitting the city index and t represents time in days:

$$
\begin{aligned}
(S_t - A_{t+1}^{SE} - A_{t+1}^{SD}, A_{t+1}^{SE}, A_{t+1}^{SD}) &\sim \text{Eulermultinom}(S_t, \mu^{SE}(t), \mu^{SD}) \\
(E_t - A_{t+1}^{EI} - A_{t+1}^{ED}, A_{t+1}^{EI}, A_{t+1}^{ED}) &\sim \text{Eulermultinom}(E_t, \mu^{EI}(t), \mu^{ED}) \\
(I_t - A_{t+1}^{IR} - A_{t+1}^{ID}, A_{t+1}^{IR}, A_{t+1}^{ID}) &\sim \text{Eulermultinom}(I_t, \mu^{IR}(t), \mu^{ID}),
\end{aligned}
$$

where $A_{t+1}^{BS} \sim \text{Pois}(\mu^{BS}(t))$. A_{t+1}^{BC} is the number of transitions from state B to C that take place between times t and $t+1$.

The probability of not transitioning within the next day is given by $p_0 = \exp(-\mu_1 - \mu_2)$, and the probability of transitioning to state i given that the person is transitioning is $p_i = \frac{\mu_i}{\mu_1+\mu_2-p_0}$.

The rate $\mu_{BS}(t)$, at which individuals become susceptible, depends on the birth rate $b(t)$ and is adjusted to account for the higher likelihood of measles spread at the start of the school year. We follow the same approach taken by He et al. (2010) to include a term $\delta(t - t_0 \mod 365)$ to add a Dirac delta impulse to the recruitment rate. This occurs when t falls on the same calendar day as t_0, the designated school admission day in England would be $t_0 = 251$ (He et al., 2010). The rate $\mu_{SE}(t)$ at which individuals become exposed is determined by the force of infection $\beta(t)$, the number of infectious individuals $I(t)$, the population size $N(t)$, and a gamma white noise process $\gamma_w(t)$ with intensity σ_{SE}.

The force of infection $\beta(t)$ varies with the school term, incorporating an amplitude a and the proportion of the year p that the school term occupies: During the school term: $\beta(t) = \beta_0(1 + a(1-p)/p)$, and during vacation: $\beta(t) = \beta_0(1-a)$.

In order to avoid significant distortion, the latent period (LP) and infectious period (IP) can be calculated as

$$\mu_{EI}(t) = \sigma, \quad LP = \frac{1}{1 - e^{-\frac{\mu_{EI}}{365}}} \tag{2}$$

$$\mu_{IR}(t) = \gamma, \quad IP = \frac{1}{1 - e^{-\frac{\mu_{IR}}{365}}} \tag{3}$$

The basic reproduction number R_0 is related to the transition rates $\mu_{IR}(t)$ and μ_{ID} as:

$$\beta_0 = R_0(1 - \exp(-(\mu_{IR}(t) + \mu_{ID})))$$

The remaining transition rates are assumed to be constant, with $\mu_{EI}(t)$, $\mu_{IR}(t)$, μ_{SD}, μ_{ED}, and μ_{ID} set to 0.02, which indicates a constant per-capita death rate (He et al., 2010). Table 1 provides the meaning for each notation used during this research. By investigating the spatial and temporal patterns of measles spread in the 200 cities, we employed the POMP model for a nuanced understanding of disease dynamics in varying population sizes. We wanted to look for any patterns in the estimated parameter that were taken into account previously but lack empirical evidence from a large data set. For example, one of the relationships we want to investigate is whether there exists a relationship between γ and population size. Another goal we have is to show that POMP models are useful for fitting large amounts of time series data in a reasonable time frame.

We then conducted the analysis in two main phases, where in the first stage, we mainly focused on fitting the unit-specific model (the model that was used in the He et al. (2010) research) to the 200 cities extending the scope of the original research, thus testing the validity of the POMP model developed in the He et al. (2010) paper. Following the analysis of parameter estimates from the 200 cities and insights we learned from the results, we then proposed alternative models in the second phase that modified the original model to test whether it improved the modeling of the measles outbreak and increased the identifiability of parameter estimates.

Notation	Meaning
N	Population Size
S_0, E_0, I_0, R_0	At time 0 , the proportion of individuals who are in states S, E, I, R.
μ_{ij}	Rate for an individual who transition from state i to state j.
γ	Per-capita rate for an individual who in state I to state R.
a	Amplitude of seasonality due to school term.
α	Mixing exponent.
R_0	Basic reproduction number.
c	The cohort entry fraction.
ρ	Reporting probability.
ι	Mean number of imported infectious individuals.
ψ	Measurement overdispersion parameter.
σ	Per-capita rate for an individual who in state E to state I.
σ_{SE}	Standard error of gamma white noise during transmission.

Table 1: Description of the notation used in the model.

For each model, we ran 36 replications of IF2 using 5,000 particles. We then evaluated the unit log likelihood using the particle filter 36 times, which then led us to obtain an estimate for each unit log likelihood by averaging those 36 estimates.

3.4 Model 1 (Unit Specific)

Model 1 uses the He et al. (2010) model to fit the 200 cities with each parameter being unit specific. Before fitting the 200 cities using the model with its implementation code specified in measlespkg, we first tested the implementation using the twenty cities in the He et al. (2010) data set and compared the results with the previous parameter estimates presented in the paper. In particular, we included a careful review of log likelihood estimates and model parameter estimates to ensure there was no implementation error on the coding side. The 200 additional cities were divided into 10 sets, each consisting of 20 cities. The first stage of data analysis consisted of two rounds, with round 2 built upon round 1 results. Along with reduced random walk standard deviation, we further used previous round 1 results for parameter estimates to increase the accuracy of the new rounds of fitting using restricted conditions. For each fit, we recorded and organized the parameter estimates for each city and drew insights from the diagnostic plots. In particular, we explored trace plots, simulation plots, and pair plots, which were graphed for each of the three models.

3.5 Model 2 ($\log(\iota)$ varying linearly with $\log$(Population Size N))

Motivated by the observation made in He et al. (2010) indicating that gamma has a power-law relationship with population size when tested on a 20-city data set, we decided to move forward by investigating the relationship between other parameter estimates and population size. We also explored whether there appeared to be relationships between pairs of parameter estimates. Aside from pair plots that provided a comprehensive overview of the relationships, the relationships between each individual

parameter and population size were plotted using scatterplots and modeled with linear models, using R^2 being an additional metric evaluating relationship strength. Among those relationship plots, we found ι to have a relatively strong linear relationship with population size as modeled by their log relationship, which is supported by the observation made by the He et al. (2010) paper that suggested ι varies linearly with population size based on a set of 20 cities. The scatterplot in Figure 2, the R^2 value of 0.1888, and the idea that the mean number of imported infectious individuals would be greater for larger cities lead us to conclude that we should take this log linear relationship into account when modeling. Thus, we made $\log(\iota)$ to vary linearly with $log(N)$, with the formula

$$\log(\iota_n) = \iota_0 + \iota_1 \log(N) \tag{4}$$

The slope and the intercept are made to be shared parameters. Since the cities are independent of each other and we use shared parameters, this model is considered a PanelPOMP model.

3.6 Model 3 (shared parameter model with unit specific ι)

As measles was more likely to be extinct in smaller cities than in larger ones, it would be reintroduced through the people who traveled from other cities, which is captured by the parameter ι. Since this parameter depends on other cities, its estimate is not as fine-tuned as many of the other parameters especially for smaller cities. Since, our data set contains cities with varying sizes, we have many smaller cities that are more sensitive to ι. These reasons lead us to model ι as unit-specific in this model rather than being shared in the previous model, so we could compare the efficacy of the different approaches. Additionally, we added more shared parameters as we were motivated by the model employed in the Ionides et al. (2022) paper, where all the other parameters are made to be shared, except for parameters representing initial values and reporting rates. It is worth noting that, unlike in the Ionides et al. (2022) paper, ψ is not made to be a shared parameter because it varies across populations and is well-identified, the estimated values vary across cities, and the log likelihoods are sensitive to ψ. Similar to model 2, this is also considered a PanelPOMP model as cities are independent of each other and have shared parameters.

4 Results

In table 2, the Model 1 parameter estimates for the 30 selected cities are presented. The use of Model 1 further demonstrated the validity of the unit-specific model and its capacity to be applicable across large samples. Following parameter estimation, we further explored how parameters (including γ, ι, σ, and others) varied with population size. We also investigated relationships between parameters to explore the possibility of future model reparameterization. The general results are shown in the pair plots (figure 4). Furthermore, we fit linear models to test the relationship strength and reported R^2 values as a metric for evaluation.

City	N	α	a	cohort	γ	ι	ψ	R_0	ρ	σ	σ_{SE}	IP	LP
London	3389620	0.96	0.55	0.66	27.22	4.16	0.11	62.86	0.49	32.90	0.09	13.91	11.60
Birmingham	1117900	0.99	0.33	0.81	31.67	0.01	0.18	37.37	0.56	52.43	0.07	12.03	7.47
Liverpool	802300	0.98	0.23	0.32	45.59	0.18	0.13	31.98	0.49	76.51	0.04	8.52	5.29
Manchester	704500	0.91	0.51	0.48	52.46	2.34	0.14	38.35	0.57	19.44	0.12	7.47	19.28
Sheffield	515000	1.01	0.32	0.17	57.13	0.75	0.16	31.96	0.66	66.87	0.05	6.90	5.97
Leeds	509700	1.01	0.36	0.66	24.33	1.33	0.17	56.84	0.66	39.85	0.11	15.51	9.67
Bristol	442600	1.01	0.19	0.15	54.90	0.66	0.17	27.56	0.63	185.47	0.03	7.16	2.51
Nottingham	307000	0.99	0.19	0.11	75.08	0.53	0.23	25.26	0.59	84.11	0.04	5.38	4.86
Hull	302100	0.98	0.68	0.05	23.64	0.32	0.23	84.52	0.54	22.65	0.18	15.95	16.62
Bradford	294300	1.01	0.66	0.34	24.75	1.39	0.21	84.54	0.57	21.02	0.18	15.25	17.87
Cardiff	244600	0.99	0.21	0.35	40.08	0.59	0.28	27.16	0.58	1795.13	0.03	9.62	1.01
Hastings	65690	0.97	0.50	0.04	65.38	0.13	0.42	45.26	0.69	46.91	0.11	6.10	8.29
Keighley	56980	1.03	0.36	0.01	125.03	0.11	0.31	38.22	0.59	39.33	0.08	3.44	9.78
Eccles	44370	0.97	0.39	0.66	52.99	0.31	0.33	37.86	0.64	81.66	0.09	7.39	4.98
Uckfield.RD	42230	0.94	0.17	0.13	119.02	0.08	0.45	19.26	0.94	64.76	0.09	3.59	6.15
Consett	39130	0.95	0.54	0.24	1370.03	0.01	0.35	68.81	0.66	24.34	0.12	1.02	15.50
Darwen	31030	1.02	0.05	0.49	41.91	0.29	0.42	19.78	0.87	639.49	0.04	9.21	1.21
Bedwellty	28930	0.99	0.04	0.14	51.89	0.15	0.94	19.40	0.32	1640.12	0.04	7.55	1.01
Runcorn	24000	0.97	0.02	0.24	28.97	0.18	0.57	228.36	0.57	29.03	0.38	12.11	13.07
Northwich	18330	0.97	0.68	0.26	28.18	0.34	0.34	22.02	0.89	69.52	0.24	13.46	5.77
Southam.RD	12870	0.91	0.42	0.04	88.96	0.03	0.74	68.37	0.53	41.53	0.13	4.62	9.29
Oswestry	10970	1.03	0.34	0.05	117.06	0.04	0.47	52.70	0.61	42.16	0.08	3.64	9.17
Dalton.in.Furness	10560	0.99	0.02	0.07	504.43	0.02	0.66	31.38	0.46	65.42	0.11	1.34	6.09
Romsey	6580	1.03	0.62	133.31	0.04	0.91	12.46	0.40	0.07	486.41	0.02	3.26	1.35
Mold	6409	1.07	0.20	0.43	79.33	0.05	2.43	13.68	0.15	332.26	0.01	5.12	1.67
Lees	4247	0.97	0.01	0.49	23.66	0.28	0.63	104.50	0.64	76.58	0.48	15.93	5.28
Llandrindod.Wells	3586	0.96	0.01	0.88	51.52	0.00059	2.0031	2015.29	0.31	925.80	0.44	7.59	1.08
St.Ives.1	3310	1.04	0.83	0.12	34.17	0.61	0.67	73.51	0.64	37.48	1.32	11.18	10.24
Withnell	2870	1.01	0.47	0.47	21.07	0.11	0.69	139.89	0.55	6386.61	0.53	17.82	1
Halesworth	2171	0.99	0.47	0.01	170.09	0.0071	0.60	92.90	0.78	29.60	0.24	2.68	12.84

Table 2: Parameter Estimates for Selected 30 Cities using model 1

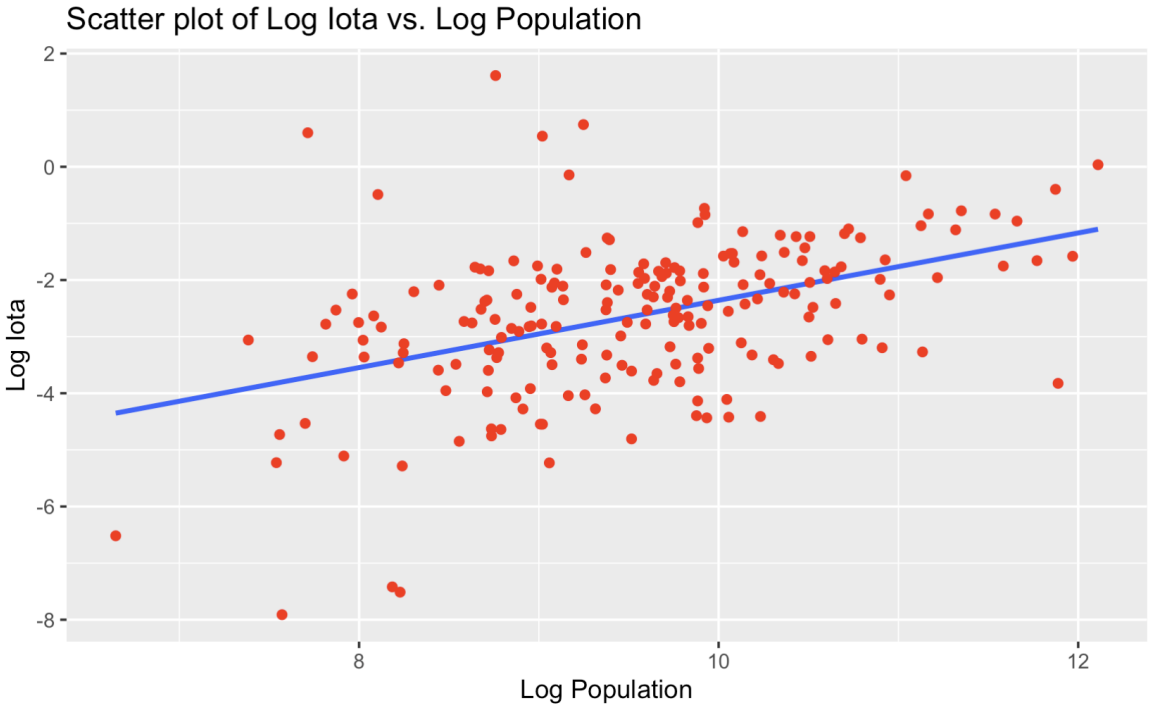

Figure 2: Scatterplot showing log-linear relationship between ι and population

Among those parameter estimates, we also paid close attention to the relationship

between γ, σ, and R_0 due to their collective influence on the disease dynamics. Specifically, R_0 and γ together determine the infection rate, while σ and γ describe the disease progression cycle. We also found α to be consistently estimated to be around 1, which is consistent with the estimates from He et al. (2010). Additionally, we wanted to highlight two key relationships that are of interest. In He et al. (2010), a power-law relationship between γ and population size was observed but was questioned due to a lack of biological interpretation. Since this observation was made on a relatively small sample of 20 cities and is closely related to how we should improve our previous model, we wanted to test this hypothesis on a larger sample of 200 cities. The larger sample showed that the observed linear relationship between γ and population size in He et al. (2010) is due to random chance. Specifically, the scatterplot in Figure 3 did not show a strong relationship between the two variables, and the corresponding low R^2 value of 0.00443 suggested weak relationship strength.

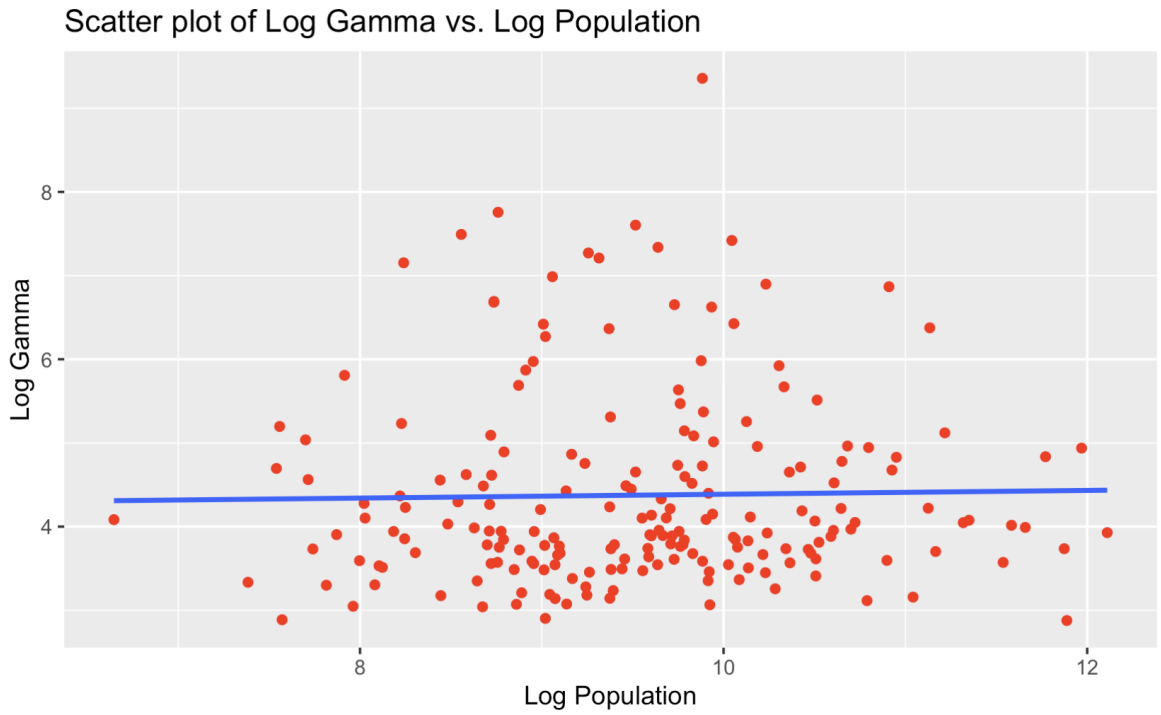

Figure 3: Scatterplot showing relationship between γ and population

The parameter estimates for Model 2 are shown in table 3. In order to determine whether shared parameters could better model the measles and improve the accuracy of parameter estimates, we then developed Model 3, with the parameter estimates presented in Table 4.

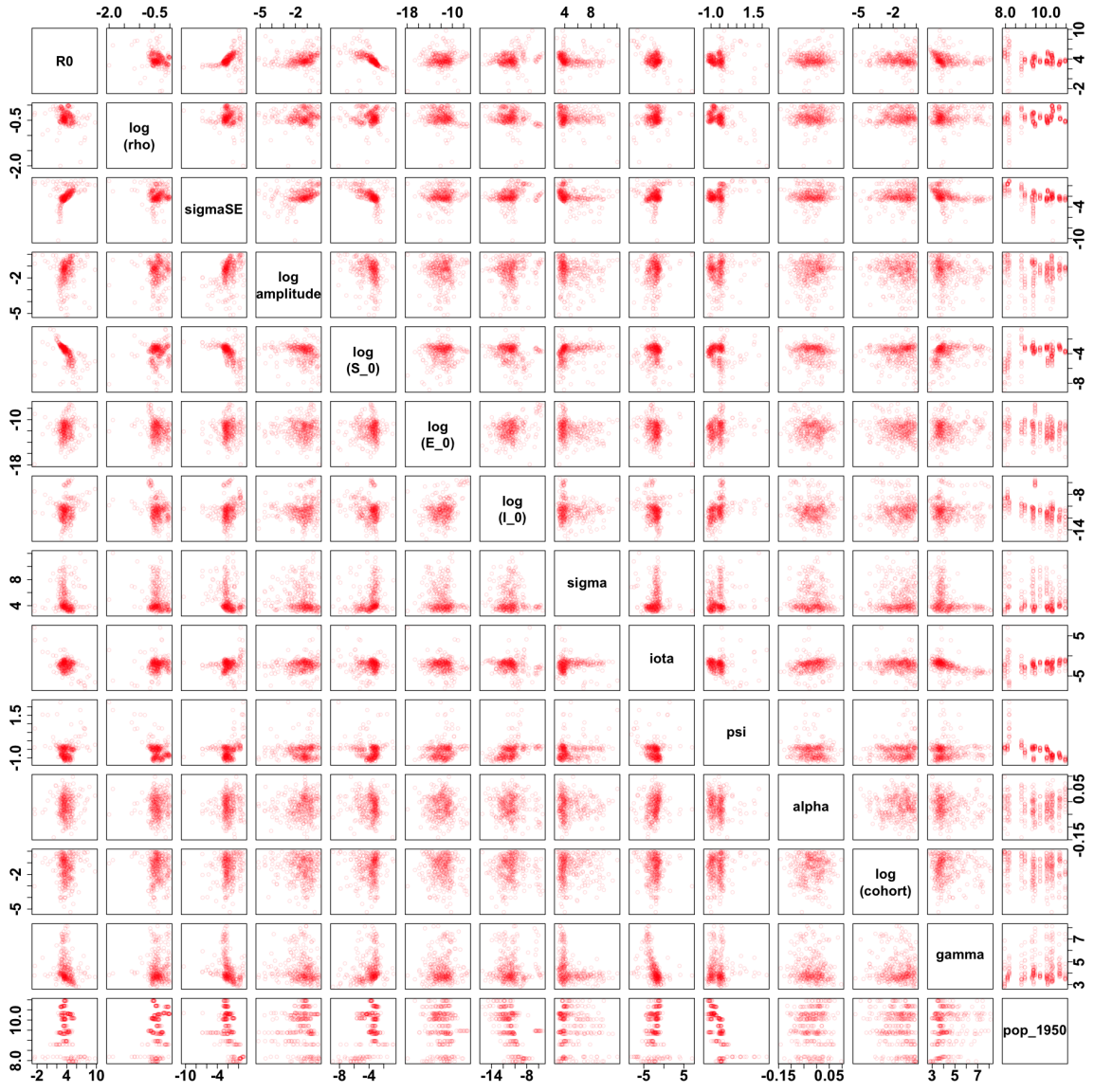

Figure 4: Pair Plot showing relationships between parameters and population size using the 200 cities

Parameter	Value
Amplitude	0.20
Cohort	0.19
γ	34.81
ι_1	-3.91
ι_2	0.86
R_0	27.68
σ	43.90
σ_{SE}	0.06
IP	10.05
LP	8.82

(a) Shared Parameter Values across All Cities

City	ψ	ρ
London	0.10	0.49
Birmingham	0.12	0.55
Liverpool	0.13	0.50
Manchester	0.13	0.63
Sheffield	0.12	0.66
Leeds	0.12	0.73
Bristol	0.19	0.60
Nottingham	0.27	0.63
Hull	0.24	0.56
Bradford	0.21	0.63
Cardiff	0.24	0.60
Hastings	0.37	0.73
Keighley	0.34	0.57
Eccles	0.42	0.63
Uckfield.RD	0.45	0.82
Consett	0.39	0.67
Darwen	0.39	0.84
Bedwellty	0.91	0.33
Runcorn	0.60	0.53
Northwich	0.48	0.76
Southam.RD	0.61	0.54
Oswestry	0.49	0.60
Dalton.in.Furness	0.94	0.46
Romsey	0.77	0.43
Mold	8.49	0.05
Lees	0.56	0.67
Llandrindod.Wells	11.70	0.10
St.Ives.1	0.65	0.57
Withnell	0.65	0.58
Halesworth	0.69	0.84

(b) Unit-Specific Parameters

Table 3: Parameter Estimates and Values using Model 2: (a) Parameter values across all 30 units. (b) Parameter estimates for selected 30 cities.

Parameter	Value
Amplitude	0.28
Cohort	1.25e-11
γ_1	38.25
R_0	28.27
σ	242766264
σ_{SE}	0.07
IP	10.95
LP	8.82

(a) Shared Parameter Values across All Cities

City	ι	ψ	ρ
London	1.60	0.09	0.48
Birmingham	2.51	0.14	0.60
Liverpool	3.06	0.14	0.49
Manchester	0.76	0.13	0.57
Sheffield	1.31	0.11	0.64
Leeds	1.61	0.14	0.67
Bristol	0.58	0.17	0.62
Nottingham	0.62	0.26	0.59
Hull	0.42	0.21	0.57
Bradford	1.22	0.18	0.59
Cardiff	0.39	0.26	0.61
Hastings	0.24	0.35	0.72
Keighley	0.33	0.31	0.59
Eccles	0.46	0.37	0.63
Uckfield.RD	0.26	0.41	0.97
Consett	0.33	0.38	0.64
Darwen	0.32	0.36	0.81
Bedwellty	0.15	0.81	0.32
Runcorn	0.27	0.55	0.58
Northwich	0.31	0.36	0.77
Southam.RD	0.15	0.71	0.51
Oswestry	0.11	0.49	0.57
Dalton.in.Furness	0.15	0.96	0.41
Romsey	0.05	1.01	0.36
Mold	0.05	6.03	0.07
Lees	0.17	0.55	0.66
Llandrindod.Wells	0.01	26.16	0.06
St.Ives.1	0.04	0.65	0.62
Withnell	0.09	0.71	0.54
Halesworth	0.05	0.55	0.81

(b) Unit-Specific Parameters

Table 4: Parameter Estimates and Values using Model 3: (a) Parameter values across all 30 units. (b) Parameter estimates for selected 30 cities.

Along with parameter estimation, we plotted trace plots and simulation plots to facilitate data analysis. An example simulation plot is shown in figure 5, in which we plotted simulated observations and compared how those resemble the actual observed data. The simulation plot contains the results of two simulations. The y-axis is the number of cases simulated from the model, and the x-axis represents the time in which those cases took place.

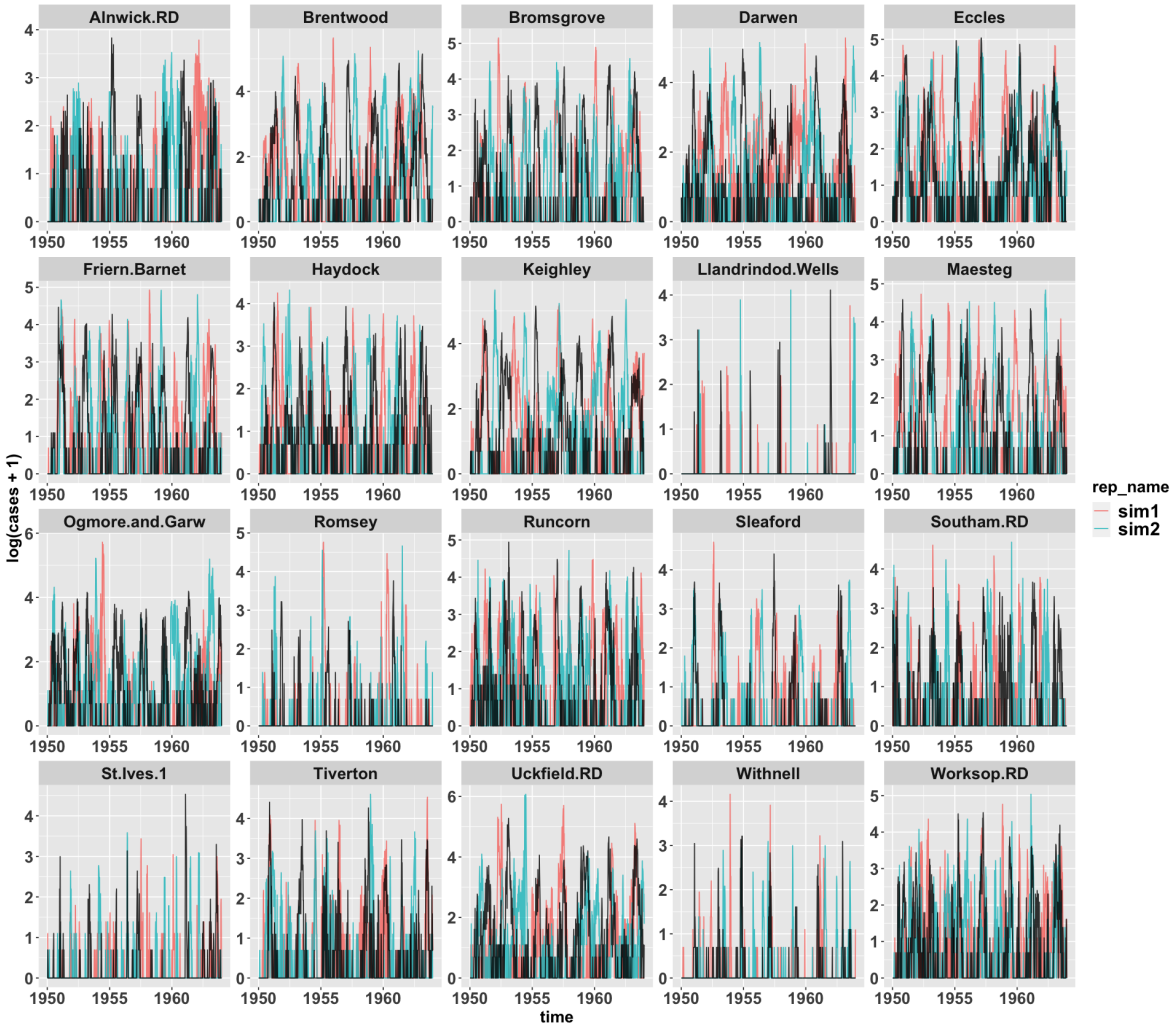

Figure 5: Simulation Plot showing the data points simulated from the model, used to compare with the actual data to determine whether the two align

While trace plots were plotted for each parameter in all three models, figure 6, 7, and 8 show examples of the traces of plots for R_0 across the three models. All 36 replications were plotted against each other and showed trends that indicated how the parameter space is explored. The trends of these traces thus suggested that our estimates for R_0 converge from the 36 replications, rather than diverging to different values. For instance, in figure 7, we can see that the 36 iterations eventually converge to the range between 30-40. The estimation for R_0 from the three models is consistently higher than commonly found R_0 values for measles, which ranges from 12 to 18 (Keeling and Ross, 2008). Nevertheless, this pattern for higher estimation is consistent with the R_0 estimation made in the He et al. (2010) paper. This could be explained by the high values of R_0 required by homogeneous population models to be consistent with the data (He et al., 2010).

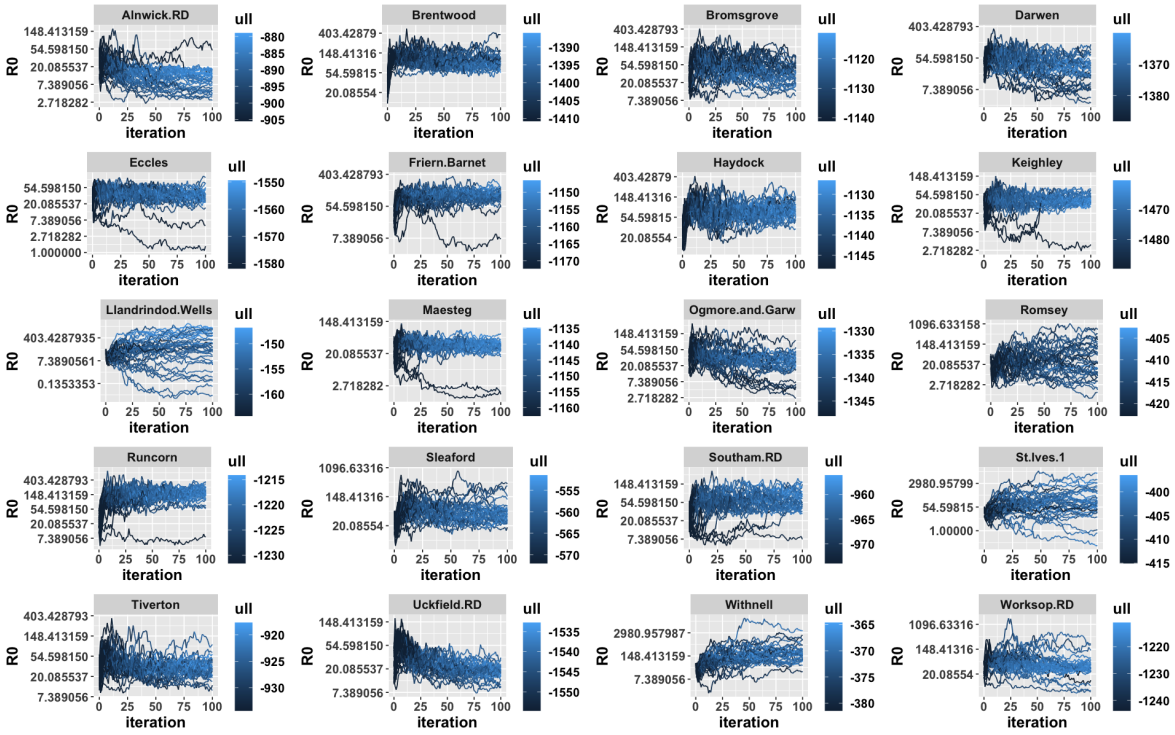

Figure 6: Trace Plot for R0 from model 1

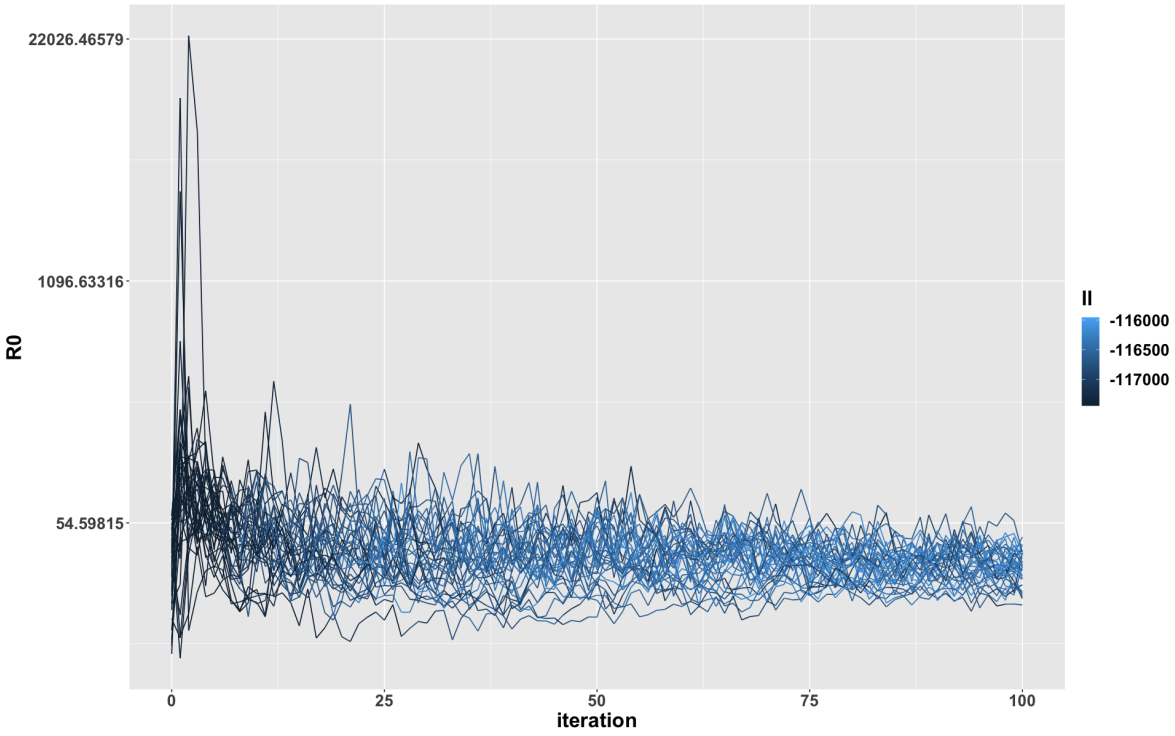

Figure 7: Trace Plot for R0 from model 2

Similar to the findings in He et al. (2010), we found discrepancies between the estimation for infectious periods and latent periods in our three models and those in

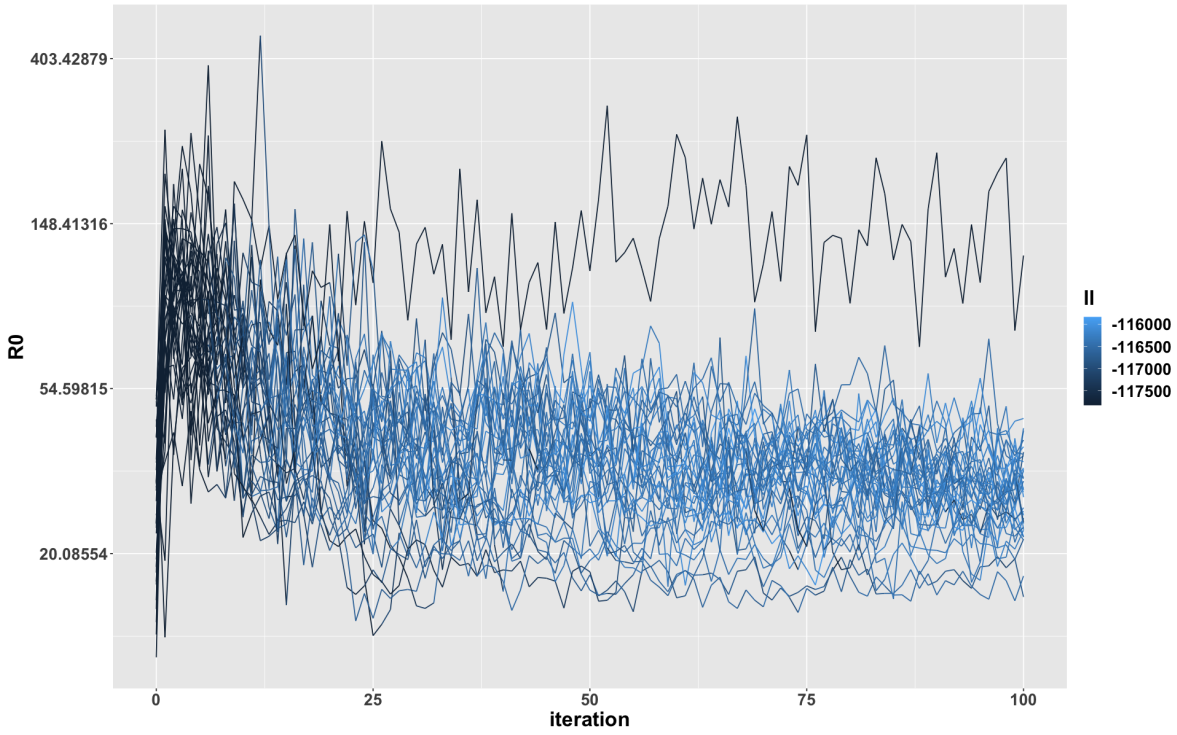

Figure 8: Trace Plot for R0 from model 3

the existing literature. While clinical observations indicate that LP should range from 8-10 days and IP should range from 4-6 days Keeling and Ross (2008), our estimates for LP and IP across all three models are not consistent, which are shown in table 2, 3, and 4. This observation suggests weak identifiability of the two estimates. Though He et al. (2010) concludes that LP and IP share relationships with population size where small populations tend to have estimates that are close to what was clinically found, our estimates obtained using a larger data sets suggest that this is more likely due to random chance.

The estimated values for σ were not consistent across the three models, with Model 1, 2, and 3 having unusually high values. Moreover, compared to the R_0 trace plots that indicate convergence across the different replications, the trace plots shown in figure 9, 10, and 11 for σ did not suggest good convergence of the 36 replications. For instance, in figure 11, the traces for sigma are diverging to different values ranging from 1×10^4 to 4.6×10^{28}. Among the three models, Model 2 with ι varies log-linearly with population size had some traces converging around 1×10^4, while others diverging to different values. Furthermore, among the three models, the σ estimate for Model 2 is the lowest, which suggests that we could explore how shared parameter models could be adjusted further to produce better estimates for σ. These evidences led us to conclude that σ is a weakly identified parameter that required further modification.

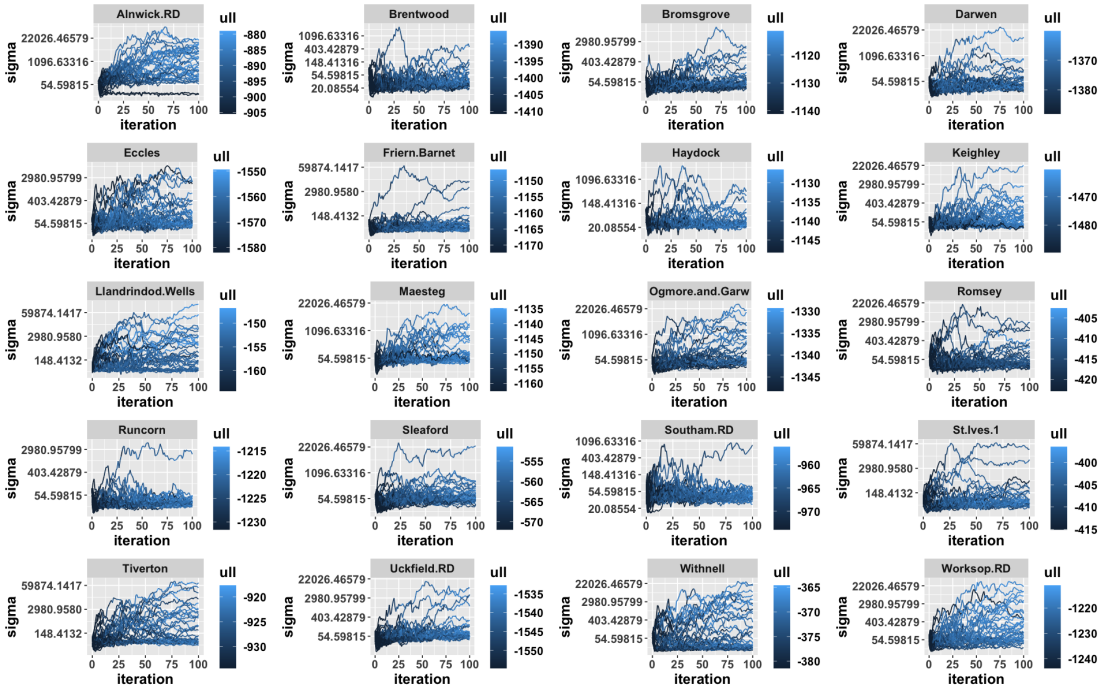

Figure 9: Trace Plot for σ from model 1

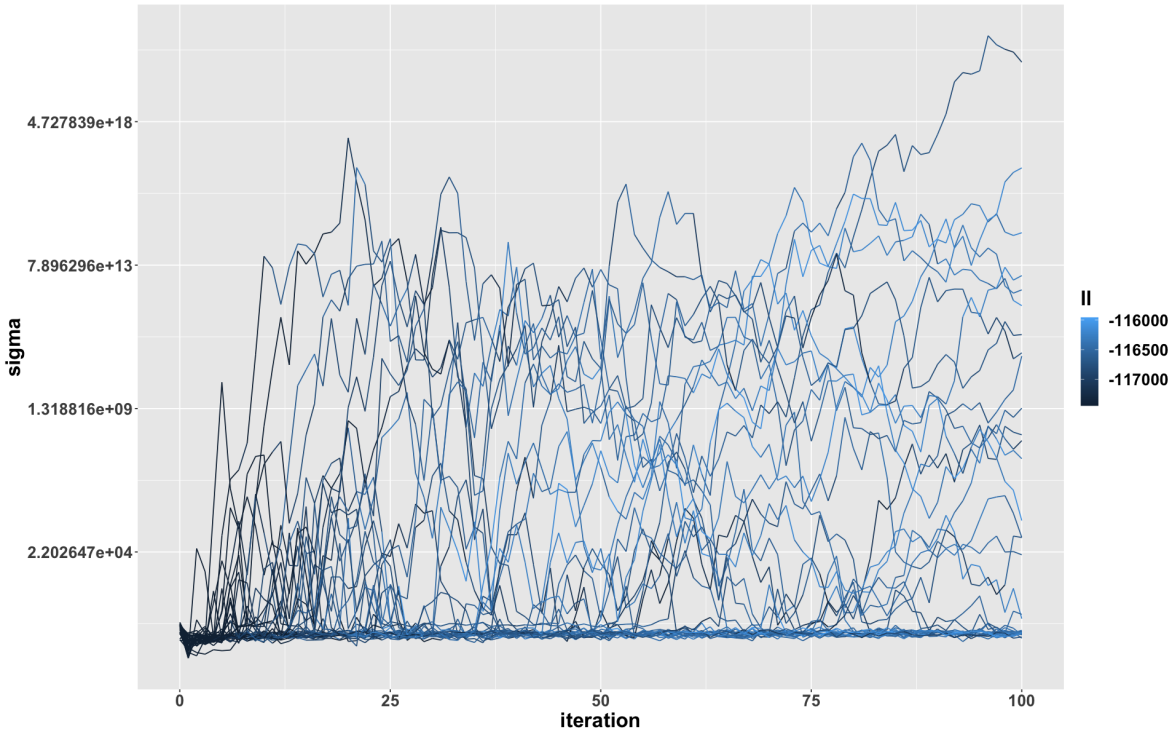

Figure 10: Trace Plot for σ from model 2

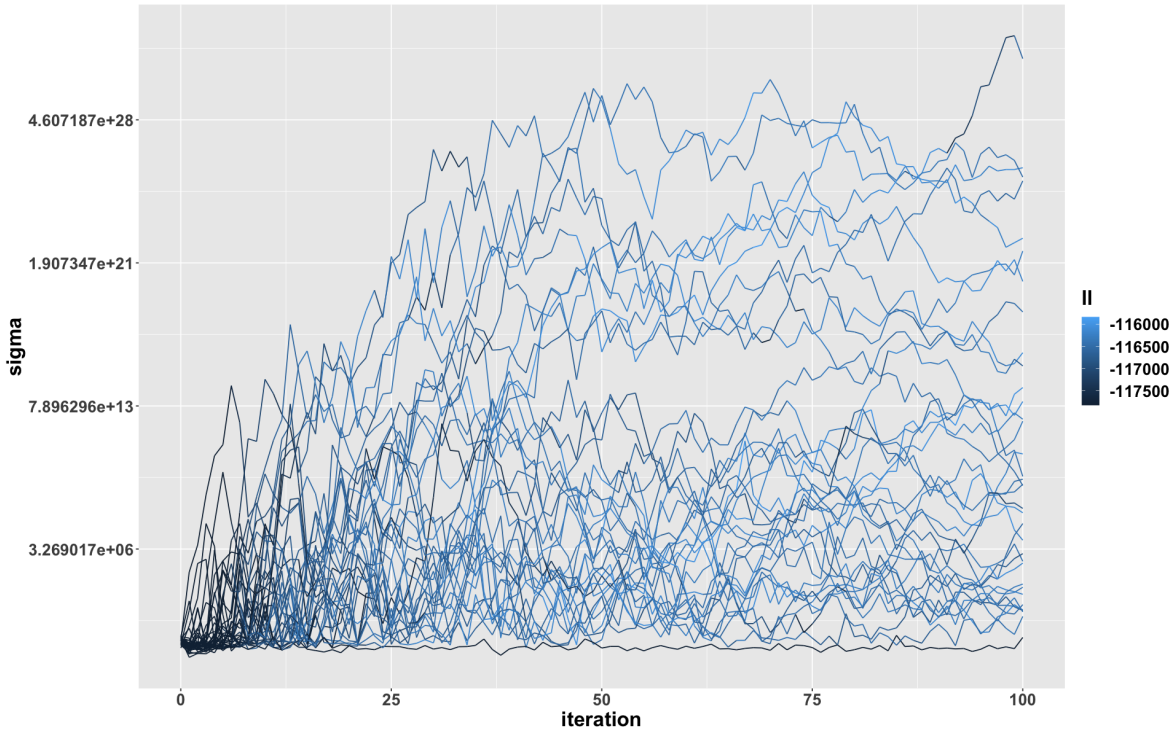

Figure 11: Trace Plot for σ from model 3

The design of the three models considers whether parameters were shared or unit-specific, which can lead to the differences in the number of parameters being included in the model and its dimensionality. If we had 20 units in the model, using a unit specific parameter would add 20 to the model dimension, while employing a shared parameter would only increase the model dimension by 1. Thus, we experienced the tradeoff between model specificity and overfitting. Due to this case, we established our goal to find a balance between the two, thus increasing predictive power and fitting accuracy. The two model evaluation metrics we used were unit log-likelihood and Akaike information criterion (AIC), to compare models' ability to quantitatively describe the observed data. Table 5 shows the comparison across unit log likelihood between the three models for a selection of 30 cities. The value for log-likelihood reflects how well the models were able to estimate the parameters and produce better fits for the data. When holding the unit fixed, Model 1 showed consistently higher unit log-likelihood than Model 2 and Model 3. Moreover, Model 1 had the lowest AIC value when compared to Model 2 and Model 3, as shown in table 6. The low AIC value indicates that Model 1 is capable of balancing the trade-off between goodness-of-fit and complexity.

Unit	m1_ull	m2_ull	m3_ull
London	-3803.61	-3934.44	-3971.18
Birmingham	-3238.24	-3334.11	-3302.75
Liverpool	-3399.41	-3500.80	-3521.16
Manchester	3238.61	-3395.42	-3400.28
Sheffield	-2813.63	-2852.86	-2852.97
Leeds	-2916.02	-3026.36	-2984.99
Bristol	-2680.42	-2706.91	-2738.73
Nottingham	-2705.18	-2717.37	-2724.92
Hull	-2717.54	-2741.95	-2754.81
Bradford	-2580.42	-2637.23	-2626.38
Cardiff	-2363.83	-2392.78	-2385.18
Hastings	-1583.19	-1589.62	-1587.18
Keighley	-1460.37	-1462.70	-1463.39
Eccles	-1550.48	-1560.80	-1555.03
Uckfield.RD	-1533.71	-1545.90	-1541.14
Consett	-1356.28	-1364.57	-1367.98
Darwen	-1361.99	-1359.07	-1362.15
Bedwellty	-1125.01	-1131.06	-1127.97
Runcorn	-1214.65	-1226.00	-1228.31
Northwich	-1191.51	-1201.15	-1197.14
Southam.RD	-956.54	-963.39	-961.64
Oswestry	-695.04	-703.03	-706.53
Dalton.in.Furness	-726.71	-738.37	-740.03
Romsey	-404.21	-411.58	-413.50
Mold	-296.37	-306.09	-301.42
Lees	-547.87	-554.92	-553.12
Llandrindod.Wells	-147.27	-165.11	-156.90
St.Ives.1	-398.08	-413.15	-412.83
Withnell	-366.15	-370.16	-372.85
Halesworth	-315.51	-317.78	-317.38

Table 5: Comparison for unit log likelihood across the three models

Model 1 AIC	Model 2 AIC	Model 3 AIC
231980.16	233769.81	233809

Table 6: Total AIC for the three models

5 Discussion

This research project demonstrates the ability of POMP models to be applicable across cities of varying sizes. By extending the original research scope of 20 cities to 200 cities, we showcased the reliability of POMP models in an epidemiology mechanistic modeling

context. However, the scale and depth of the project is still limited to an extent and has room for improvement. From parameter estimation and model comparison for the 200 cities, we found consistencies that aligns with prior research as well as new insight that could empower future investigations. As the observation made in He et al. (2010) suggests a power-law relationship between γ and population size, based on the data analysis conducted on the 200 cities, we provided empirical evidence to conclude that this observation is likely due to random chance.

On the other hand, the ι log-linear relationship with population size that was concluded in He et al. (2010) was supported by the evidence found by fitting 200 cities using unit-specific models, which was then taken into consideration for model fitting. Similar to He et al. (2010), we found a pattern of relatively large value of R_0 estimation across all three models. This phenomenon can be explained by several factors, including the disproportionate impact of peak incidences on R_0 and reporting rates, which results in higher R_0 values aligning with the peaks in observed epidemics. Additionally, the overestimation of R_0 could reflect the initial stages of epidemic spread, potentially capturing the higher contact rates among school-aged children rather than across the entire population. The model's tendency to extend these rates to adults, who are less represented in the data, could further inflate R_0 estimates. The assumptions related to the transmission's age structure also contribute to the variability in R_0 calculations, with the model's limitations in fully accounting for age-specific transmission dynamics. Moreover, the larger R_0 figures may be influenced by the spatial distribution of transmission, particularly in more populated urban settings, where the model's estimates diverge from expected values based on household data (He et al., 2010).

Another shared pattern across the three models is the differences in the infectious period and latent period as compared to previous literature, which was also found by He et al. (2010) when conducting analysis on the 20-city data set. However, when fitting on this large data set, we found the relationship between LP, IP, and population size that was suggested in He et al. (2010) is more likely due to random chance. Our results do show the inverse relationship between the two, as IP and LP should add up to be about 14 days (Keeling and Ross, 2008). Thus, more effort in the future could be put into improving the weak identifiability of LP and IP, potentially by adding more shared parameters and re-parameterizing the model to capture the disease characteristics more accurately.

To compare the models' ability to quantitatively describe the observed data, we used AIC and log likelihood to facilitate our evaluation. The larger log likelihood value and smaller AIC value for Model 1 when compared to the other two models suggests that Model 1 produced parameter estimates of high accuracy and better fitting. However, this result does not lead us to directly conclude that models with shared parameters should not be considered in future modeling processes. We should instead fit a second round for the shared parameter models using previously learned information from the first round of fitting. For unit-specific models, we could choose the best parameter estimates from all 36 replications, thus leading to high log likelihood. However, this optimization strategy is not applicable to shared parameter models, as using shared parameter estimates from different fits and combining different shared parameter estimates could lead to lower log likelihood, making it less ideal. Unfortunately, due to time constraints, we only had the chance to conduct 1 round of fitting for the cities. In future investigations, we would take the top fit for the shared model and then replicate

it 36 times. While fixing the shared parameter, we want to find the best parameter estimates for the unit specific parameters, thus improving the log likelihood estimate.

Since the raw data set contains 1,402 cities, it would be interesting to see how the POMP models could be adapted to fit all those cities and produce parameter estimates of high accuracy. Although the data was carefully collected to record measles cases back in the 1950s, it might still contain a few outliers that could have resulted from multiple factors, including hospital closure during holidays or misrecorded cases in small cities. Since researchers in the He et al. (2010) paper conducted detailed outlier removal for the 20 cities and concluded that several data points were considered outliers, future research studies could consider cleaning the raw data for the 1,402 cities through rigorous outlier removal procedure to further improve the accuracy of model fitting and parameter estimation. The decision to make a parameter shared or unit specific could still be an interesting topic to explore in the future with ongoing development of the MCAP method. Thus, it would be worthwhile to explore the possibility of reparameterizing other parameters in the context of measles, further improving model fitting.

References

Becker, A. D. and Grenfell, B. T. (2017). tsir: An r package for time-series susceptible-infected-recovered models of epidemics. *PLOS ONE*, 12(9):e0185528.

Bretó, C., Ionides, E. L., and King, A. A. (2019). Panel data analysis via mechanistic models. *Journal of the American Statistical Association*, 115(531):1178–1188.

Breto, C., Ionides, E. L., and King, A. A. (2022). *panelPomp: Statistical Inference for PanelPOMPs (Panel Partially Observed Markov Processes)*. R package version 0.16.0.

Fine, P. E. M. and Clarkson, J. A. (1982). Measles in england and wales—i: an analysis of factors underlying seasonal patterns. *Int. J. Epidemiol.*, 11:5–14.

Gibson, G. and Renshaw, E. (2001). Likelihood estimation for stochastic compartmental models using markov chain methods. *Statistics and Computing*, 11:347–358.

Glass, K., Xia, Y., and Grenfell, B. T. (2003). Interpreting time-series analyses for continuous-time biological models: measles as a case study. *J. Theor. Biol.*, 223:19–25.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *Journal of The Royal Society Interface*, 7(43):271–283.

Ionides, E. L., Breto, C., Park, J., Smith, R. A., and King, A. A. (2017). Monte carlo profile confidence intervals for dynamic systems. *Journal of The Royal Society Interface*, 14(132):20170126.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA*, 103:18438–18443.

Ionides, E. L., Nguyen, D., Atchadé, Y., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724.

Ionides, E. L., Ning, N., and Wheeler, J. (2022). An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters. *arXiv preprint arXiv:2206.03837*.

Keeling, M. and Ross, J. (2008). On methods for studying stochastic disease dynamics. *J. R. Soc. Interface*, 5:171–181.

King, A., Nguyen, D., and Ionides, E. (2015). Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software*, 69(12):1–43.

Korevaar, H., Metcalf, C. J., and Grenfell, B. T. (2020). Structure, space and size: competing drivers of variation in urban and rural measles transmission. *Journal of The Royal Society Interface*, 17(168):20200010.

Liu, J. and West, M. (2001). Combining parameter and state estimation in simulation-based filtering. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 197–224. Springer, New York, NY.