THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF MATHEMATICS

A B-spline Hierarchical Model on HIV Prevalence Data

Yuan Tang
Spring 2015

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degree
in Mathematics
with honors in Mathematics

Reviewed and approved* by the following:

Le Bao
Assistant Professor of Statistics
Thesis Supervisor

Luen-Chau Li
Professor of Mathematics
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

Due to the paucity of reliable information on the incidence of HIV in most countries, sentinel surveillance systems for HIV were designed to provide information on prevalence trends, e.g. the percentage of HIV positive cases is often estimated among antenatal clinic (ANC) patients. In recent years, surveillance sites have been established which allow countries to look into the epidemics at a finer scale, e.g. at sub-national level or sub-population level. An important technical barrier is that the availability and quality of the data vary widely from area to area, and many areas lack data for deriving stable and reliable results. In countries with low-level and concentrated epidemics, HIV has spread rapidly in several high risk groups, but is not well established in the general population. Fewer data are available for those sub-populations due to the stigmatized nature of these sub-populations in many countries. To improve the accuracy of the results in areas or high risk groups with little data, we propose a Bayesian hierarchical model that utilizes information more efficiently by assuming similarity of HIV prevalence across areas and sub-populations.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The following acknowledgements are not stated in any particular order:

I would like to thank Professor Le Bao, my thesis adviser, for his insightful guidance and supervision on this thesis research project. His meticulous attitude towards conducting research truly has affected me a lot.

Thanks to Dr. Keith Sabin for his advise on the data, and to Professor Xiaoyue Niu and Ben Sheng for sharing their R code.

Also, thanks to my undergraduate advisor Professor Luen-Chau Li for his advice and encouragement throughout my undergraduate study. He has been very supportive, given me the freedom to pursue various projects, and encouraged me to take high-level graduate courses that give me not only solid foundation of my future studies, but greatly enlightened my career path.

Thanks to Schreyer Honors College for all the financial support on my study and research, all the valuable academic and non-academic activities provided that helped me grow throughout my study at Penn State University.

Thanks to Wenxuan Li, who has been giving tremondous support and help every step of the way since the first day we met.

Furthermore, I would like to sincerely thank my host parents, Daniel Janzen and Teresa Janzen, who introduced me to a whole different perspective of the world during my senior high at Michigan. They fostered my desire to pursue the truth.

Last but not least, I would like to thank my parents for their wholeheartedly love, support, and encouragement on pursuing everything I love. I wouldn't have achieved anything without them and I will always be grateful for all the sacrifices they have made to help me grow.

Chapter 1

Introduction and Background

1.1 Importance of estimating HIV epidemic

According to estimates by the World Health Organization (WHO) and the Joint United Nations Programme on HIV and AIDS (UNAIDS), 35 million people were living with HIV globally by the end of 2013. In 2013, about 2.1 million people were infected and 1.5 million died of AIDS-related causes. Tuberculosis (TB), the most common life-threatening opportunistic infection, kills nearly 36000 people living with HIV each year. It is therefore the most deadly cause among HIV-infected people worldwide, especially in Africa. Estimating HIV epidemic becomes essential in order to prevent people from being infected rapidly and get people away from high-risk groups. Estimates of HIV/AIDS prevalence are the primary measure of the state of the overall epidemic in a country. Those estimates can help both global and national program planning and decision making related to allocations of healthcare resources. Estimates of HIV prevalence are also important since they are often used to create models of the epidemic, and thus estimates of incidence and mortality can then be derived.

1.2 Challenge, sparse, and imbalanced data

Due to paucity of reliable information on the incidence of HIV in most countries, sentinel surveillance systems for HIV were designed to provide information on prevalence trends. For instance, the percentage of HIV positive cases is often estimated among antenatal clinic (ANC) patients. In recent years, increasing numbers of surveillance sites have been established, allowing countries to look into the epidemics at a finer scale, e.g. at sub-national level and sub-population level. An important technical barrier is that the availability and quality of the date vary widely from area to area, and many areas are lack of data for deriving stable and reliable results. In countries with low-level and concentrated epidemics, HIV has spread rapidly in several high risk groups, but the surveillance system is not well-established in general population in order to analyze such high HIV risk groups. Fewer data are available for those sub-populations due to the stigmatized nature of these sub-populations in many countries.

Since 1997, UNAIDS and WHO have produced country-specific estimates of HIV/AIDS biannually, which become a primary source of information about the extent and spread of the HIV epidemic. These estimates are prepared under the guidance of the UNAIDS Reference Group on Estimates, Modelling and Projection, which gives reviews on the quality and availability of epidemiological data, new diagnostic, surveillance and treatment technologies, and evolving statistical and mathematical approaches in order to improve estimates and projections over time. These estimates are based on national Spectrum files collected and maintained by experts (see more information at http://apps.unaids.org/spectrum/). UNAIDS has made these files public to make it more transparent in how those estimates are derived. We are permitted to use the data set of the three countries, Nigeria, Papua New Guinea, and Vietnam, to conduct this research project. Though real data will be used in the analysis, fake data, such as country names, area names, etc, will be shown in this paper instead. The results presented in this thesis are based on illustrative HIV prevalence data for these countries, which may not be complete. These results should therefore not be seen as replacing or competing with official estimates regularly published by countries and UNAIDS. A glance at the example data set will be presented in Data Process section in the next chapter.

1.3 Our idea

To improve the accuracy of the results in areas or high risk groups with limited data available, we propose a Bayesian hierarchical model that utilizes information more efficiently by assuming similarity of HIV prevalence across areas and sub-populations. We also apply basis splines on the time parameter to better model the prevalence trend over time. We will apply our models to three different countries, namely, Nigeria, Papua New Guinea, and Vietnam, and will present the experimental results from the models.

Chapter 2

Methods

2.1 Data Process

Below are two of the segments of the raw prevalence data for one particular country, namely Country1. Note that for confidential reasons, we are not including real names and real data in this paper. In this piece of data, we have information such as area name, risk group name (on the right of the area name, separated by a back slash), site name, active, year, sample size and prevalence that correspond to each particular site, area, and year.

Area1\IDU											
Active(Y/N)		1985	...	1996	1997	1998	1999	2000	2001	2002	...	2020
Y	Site1 all IDU (% HIV+)			2.5	1.2	3.5	10.83	12.29				
	Sample size			300	500	250	300	300				
Y	Site2 (% HIV+)				1.5	1.3	3.2	3.6				
	Sample size				315	708	1000	900				
Y	Site1 IDU in community (% HIV+)								3	4.3		
	Sample size								200	400		
Area2\MSM												
Active(Y/N)		1985	...	1996	1997	1998	1999	2000	2001	2002	...	2020
Y	Site1 old (% HIV+)				12.29	21.3	22					
	Sample size				300	300	300					
N	Site2 (% HIV+)				2.8	10	12.1					
	Sample size											
Y	Site1 new (% HIV+)								1.5	1.3		
	Sample size								315	708		

Figure 2.1: Sample Segment of the Raw Data for One Country

After an execution of a program written in R (a programming language and software environment for statistical computing and graphics), which automatically collects useful cells in the raw data, all the segments in the raw data are cleaned and formatted as a standard format, e.g. each column represents different predictors (year, site, area, prevalence, sample size, active, and risk group) while each row represents one segment in the raw data, as in the format above.

The variable prevalence is represented in percentage, which is the percentage of the people who show HIV positive in the sample population. Each risk group represents different group of individuals who have the greatest risk of becoming infected. Here we have injecting drug users (IDU), men who have sex with men (MSM), sex work clients, sex workers, male remaining population (male population that don't belong to other risk group), and female remaining population (female population that don't belong to other risk group). Note that the risk group information is only available for Vietnam, which will be used in our analysis.

2.2 Regression Splines

Regression splines introduces great flexibility with fixed degree, which greatly relieves the pressure of overfitting. To discuss *regression splines*, first let's introduce *piecewise polynomials*.

2.2.1 Piecewise Polynomials

Piecewise polynomial regression fits separate low-degree polynomials on different intervals. A piecewise cubic polynomial, for example, fits a cubic regression model in the following form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

where the coefficients of x differ on different intervals of x. *Knots* are where the coefficients change. For example, if point c is the only knot in a piecewise cubic polynomial, two distinct polynomial functions need to be fitted differently on two subsets of observations, $x_i < c$ and $x_i \geq c$, with two different sets of coefficients for x in each range, and then the two polynomial functions are fitted using least squares.

Piecewise polynomial is more flexible when using more knots in the model. However, some knots may lead to discontinuity at those corresponding data points when the fitted curve is too flexible. To deal with this, we restrict both the first and second derivatives of the piecewise polynomials to be continuous, in other words, to be very *smooth*.

2.2.2 Basis of Regression Splines

A *cubic spline* with K knots can be described as follows:

$$y_i = \beta_0 + \beta_1 f_1(x_i) + \beta_2 f_2(x_i) + \cdots + +\beta_{k+3} f_{K+3}(x_i) + \epsilon_i,$$

where $f_1, f_2, \cdots, f_{K+3}$ are the choice of basis functions. Similar to piecewise polynomials, this cubic spline can then be fitted using least squares.

Alternatively, cubic splines can be represented more directly by starting with a basis for a cubic polynomial, such as x, x^2, x^3, and then a *truncated power basis* function is added per knot, which is defined as:

$$f(x, \xi) = \begin{cases} (x-\xi)^3 & x > \xi \\ 0 & otherwise, \end{cases}$$

where ξ represents the knot. By adding a term in the form $\beta_4 f(x, \xi)$ to the model for a cubic polynomial we described earlier, a discontinuity will occur only at ξ for the third derivative of the function and the function itself will remain to be continuous, in this case, with continuous first and second derivatives at each knot.

Furthermore, extremely high variance can occur at the boundary region of the predictors in splines model. *Natural spline* is a regression spline that has additional boundary constraints so that the function is linear at the boundary, which produces a more stable results at the boundaries.

Regression splines usually give better results than polynomial regression does. One reason is that polynomial regression usually needs to have a high degree to generate relatively flexible fitted curve, which is very possible to overfit the data points. Regression splines, on the other hand, introduce great flexibility with fixed degree by adding the number of knots. Often we place more knots in regions where the function changes very rapidly and place few knots in regions where the function is mostly stable. Regression splines, especially natural cubic splines for example, provide fairly plausible fit to the data points while avoiding or reducing the possibility of overfitting.

We now give an example program in R to illustrate how to fit regression splines. In order to fit the regression splines by an appropriate matrix of basis functions. An R library called *splines* offers a *bs()* function which can generate a whole matrix of basis functions that represent the family of piecewise polynomials for splines with pre-defined set of knots. Fitting *prevalence* and *year* in our data set using a regression splines is simple:

```
> require(splines)
> fit <- lm(prevalence ~ bs(year, knots=c(1988, 2001, 2003)),
       data=prevalenceData)
> pred <- predict(fit, newdata=list(year=year.grid),se=T)
> plot(year, prevalence, col="gray")
> lines(year.grid, pred$fit, lwd=2)
```

In the above code, we have pre-defined the set of knots at year 1988, 2001, and 2003. Note that *bs()* function produces cubic splines in default. One can also specify *df*, degree of freedom, as an option in *bs()* rather than *knots* to produce a spline with knots at uniform quantiles of the data. For instance, in one of the following sections – hierarchical spline model setup – we constructed three bases of regression splines of time t with 3 degrees of freedom, denoted as $f_1(t), f_2(t), f_3(t)$. They can be programmed in R as simply *bs(t, df=3)*.

Additionally, if we want to fit a natural spline, we can use the *ns()* function, instead of *bs()*, using the same syntax as illustrated above. Recall that boundary constraints are added on the basis of regression splines to let the function be linear at the boundary, in order to produce a more stable results.

2.3 Hierachical Model

Multiple levels exist in data in many fields, such as in education sector. For instance, student, classroom, and school are often the possible levels in data in such field. Experimental results may differ a lot if conducted in different levels. *Hierarchical Linear Modeling (HLM)* now come in handy when conducting this kind of statistical experiments, taking the hierarchy or level into account. HLM is a more complicated form of ordinary least squares regression, which can be used for analyzing variance in the dependent variables when independent variables are at different hierarchical levels. In case of the data in education sector, for example, students in a classroom share variance from the same teacher and the same classroom while students in a school share variance from attributes that belong to the entire school, such as the location of the school, population of

the school, etc.

In general, HLM takes the shared variance in hierarchically structured data into account and it is currently widely used in many areas such as health, social network, and business. HLM can efficiently examine relationships for both within and between hierarchical levels of grouped data and therefore account for variance among different predictors in various levels.

When conducting a HLM analysis, there are generally three things that need to be considered by researchers. First, one has to choose the predictors to be included in the model analysis, e.g. some of them are correlated and some of them do not have strong cause-and-effect relationship with the dependent variables. Second, the researcher must conclude whether the relationships among the predictors are *fixed* or *random*. Furthermore, the researcher need to choose an appropriate model fit statistics to assess diffrent models. We introduce different types of models that can be conducted in the HLM analysis and then give some example model fit statistics in the following subsections.

2.3.1 Model Types

A *random intercepts model* is a model whose intercepts can change, which leads to the change in the predicted dependent variable for each observation across different groups. One assumption of this model is that the slopes are the same in different situations. This model is usually used in the first place and is very helpful in analyzing whether a HLM is required or needed. A *random slopes model* is a model whose slopes can change over different predictors that represent groups. It assumes that the intercepts of the model are fixed.

It is usually very realistic and effective to have a model containing both random intercepts and random slopes, even though the model might seem very complicated. The intercepts and slopes of this mixed model are allowed to vary across different groups. Therefore, it is the most flexible and meaningful in various situations where the relationships among predictors and groups are more obscure and harder to model than using a single model.

When conducting a HLM analysis, one usually starts with the simplest model with fixed slopes and intercepts. Once a researcher thinks it is possible that one predictor might be correlated with another predictor, that is, the effect of one predictor is in chain of the effect of another predictor. The researcher can make the corresponding change in the model, such as making one fixed slope to vary across groups, we can see whether the model fit has improved or not by looking at the model fit statistics for those two models. Let us take a look at different model fit statistics as follows.

2.3.2 Model Fit Statistics

Likelihood-ratio test is one such statistical test used to compare the fit of two models, such as comparing the simplest model with fixed slopes and intercepts with the model with random slopes model. Let's take the first model as a null hypothesis H_0 and the second model to be alternative

hypothesis H_1. The likelihood function of the null hypothesis is defined as $f(x|H_0)$ and $f(x|H_1)$ is for the alternative hypothesis. $\Lambda(x)$ is the likelihood of the null hypothesis over the alternative hypothesis, defined as follows:

$$\Lambda(x) = \frac{f(x|H_0)}{f(x|H_1)} = \frac{L(H_0|x)}{L(H_1|x)}.$$

If $\Lambda(x) > c$, we accept the null hypothesis while rejecting it if $\Lambda(x) \leq c$, where c is the threshold.

Another important model fit statistic we want to mention here is *deviance information criterion (DIC)*. It is a hierarchical modeling generalization of AIC (Akaike information criterion) and BIC (Bayesian information criterion). We only introduce DIC here. DIC is an aysmptotic approximation when the sample size becomes very large and it is only a valid criterion if the posterior distribution is approximately multivariate normal. DIC is defined as follows:

$$D(\theta) = -2log(p(y|\theta)) + C,$$

where y is the data, θ is the unknown parameter in the model, $p(y|\theta)$ is the likelihood function as we mentioned before, and C is an unknown constant but is usually canceled out when comparing different models.

DIC is very useful when selecting Bayesian models where the *Markov chain Monte Carlo (MCMC)* method is used to obtain the posterior distributions of the models. We will introduce it in the following section. We use MCMC when building the HLM model and therefore DIC is more suitable for our HLM analysis.

2.3.3 Markov Chain Monte Carlo (MCMC)

In order to build a HLM, one needs to obtain the posterior distribution, which often requires us to integrate functions in high dimensions. This approach is very difficult and the computational cost is high. One approach to conquer or relieve this situation is through *Markov Chain Monte Carlo (MCMC)* methods. MCMC methods try to simulate sample draws directly from the distribution of interest, which is often very complicated. The name of MCMC methods comes from the fact that these methods use previous sample to randomly simulate the next sample, which is so-called a *Markov chain*. One special MCMC method, the *Gibbs sampler*, is the most widely applicable method for many Bayesian problems. Also, the Metropolis algorithm is the root of the MCMC methods, which has been used by physicists to compute complex integrals by drawing samples from the distribution. We will introduce these two methods later in this section. Let's first introduce Markov chains in order to understand those two methods.

The Markov Chains

Let (X_t) be a discrete time stochastic process with $\mathbb{R}$ as its *state space*. If the transition probabilities between different values in state space depend solely on the current state of the random

variable, we call this stochastic process a *Markov process*, i.e,

$$Pr(X_{t+1} = s_j | X_0 = s_k, \cdots, X_t = s_i) = Pr(X_{t+1} = s_j | X_t = s_i).$$

In other words, the only information we need in the past to predict the next state is the current state of the random variable. Knowing about the information on earlier states does not affect the *transition probability*. This sequence of random variables, namely, $(X_0, \cdots, X_n)$, generated by this Markov process, is called a *Markov chain*. One particular chain of the Markov chain is said to be its *transition probabilities*, or the probability that a process at state space s_i moves to state s_j in a single step, defined as follows:

$$P(i, j) = P(i \to j) = Pr(X_{t+1} = s_j | X_t = s_i).$$

We use $\pi_j(t) = Pr(X_t = s_j)$ to denote the probability for the chain being in state j at time t. We denote the state space probability at time t as $\pi(t)$, which is a row vector. The chain starts at the row vector $\pi(0)$, with often all the elements being zero except for one of it.

The probability that a chain has value s_i at time $t + 1$ is given by the *Chapman-Kolomogrov equation*, as shown in the following:

$$\begin{aligned} \pi_i(t+1) &= Pr(X_{t+1} = s_i) \\ &= \sum_k Pr(X_{t+1} = s_i | X_t = s_k) \cdot Pr(X_t = s_k) \\ &= \sum_k P(k \to i)\pi_k(t) = \sum_k P(k, i)\pi_k(t) \end{aligned} \tag{2.1}$$

Note that the above equation sums over the probability of being in a particular state at current step and the transition probability from that state into state s_i. Let P be the *probability transition matrix* whose i, jth element is $P(i, j)$, or the transition probability from state i to state j. Then the above Chapman-Kolomogrov equation becomes

$$\pi(t + 1) = \pi(t)P.$$

Let's first iterate this equation as follows:

$$\pi(t) = \pi(t - 1)P = (\pi(t - 2)P)P = \pi(t - 2)P^2.$$

Then we can easily see that $\pi(t) = \pi(0)P^t$. The *n-step transition probability* is defined as the probability of the process being in state j, given the condition that it was in state i n steps ago:

$$p_{ij}^{(n)} = Pr(X_{t+n} = s_j | X_t = s_i).$$

It is then obvious that $p_{ij}^{(n)}$ is just the ij-th element of P^n. The Markov chain is described as *irreducible* if there exists a positive integer such that $p_{ij}^{(n)} > 0$ for all i, j. In other words, any of the state can go to any other state in a number of steps. A chain is *aperiodic* when the number of steps needed to move from one state to another state is not a multiple of an integer. Alternatively

speaking, there's no fixed length cycle between certain states in the chain.

Furthermore, the discrete-state Markov chain can be generalized into a continuous-state Markov chain by using a probability kernel $P(x, y)$ that satisfies the following:

$$\int P(x, y)dy = 1,$$

and then the Chapman-Kolomogrovo equation for the continuous-state Markov chain becomes

$$\pi_t(y) = \int \pi_{t-1}(x)P(x, y)dy.$$

When the chain is irreducible and aperiodic, the Markov chain may reach a *stationary distribution* π^*, characterized by

$$\pi^* = \pi^* P,$$

which basically means the vector of probabilities of being in any given state is independent of the initial condition of the Markov process. Moreover, at the equilibrium, the stationary distribution for a continuous-state Markov chain satisfies the following:

$$\pi^*(y) = \int \pi^*(x)P(x, y)dy.$$

The Metropolis-Hasting Algorithm

Mathematical physicists attempt to obtain random samples from some complex probability distribution $p(x)$ in order to compute complex integrals. In order to do this, the *Metropolis-Hastings algorithm* was developed to draw samples from some distribution $p(\theta)$ where $p(\theta) = f(\theta)/K$, where K is the normalizing constant, which may be unknown and hard to calculate. The *Metropolis algorithm* generates a sequence of draws from the distribution as illustrated in the following steps:

1. Initialize value θ_0 that satisfies $f(\theta_0) > 0$.

2. Compute $\alpha = \min\left(\frac{f(\theta^*)}{f(\theta_{t-1})}, 1\right)$.

3. Accept a candidate point with probability α, which is the probability of a move.

4. Generate a Markov chain $(\theta_0, \theta_1, \cdots, \theta_k, \cdots)$ as the transition probability from θ_t to θ_{t+1}, depending solely on θ_t.

5. After a sufficient number of iterations, the chain reaches its stationary distribution, and then the samples from the vector $(\theta_{k+1}, \cdots, \theta_{k+n})$ are samples from $p(x)$.

The Metropolis algorithm was later generalized by Hastings using an arbitrary transition probability function $q(\theta_1, \theta_2) = Pr(\theta_1 \to \theta_2)$. The acceptance probability for a candidate point is set to be

$$\alpha = \min\left(\frac{f(\theta^*)q(\theta^*, \theta_{t-1})}{f(\theta_{t-1})q(\theta_{t-1}, \theta^*))}, 1\right),$$

which is called the *Metropolis-Hastings algorithm*.

Usually during the process of applying Metropolis-Hastings sampler, we often throw out the first 1000 to 5000 elements, which is called *burn-in*, and then a convergence test is used to see whether the stationary distribution has been reached. The burn-in process will then make the chain independent of the starting values by removing the effects of the initial sampling values. However, if the starting values are bad, the necessary burn-in time will be increased. One way to relieve this situation is to start the chain as close to the center, e.g. mean or mode, of the distribution as possible. A *poorly mixing* chain happens when the chain stays in a small regions of the parameter space for a long time. In contrast, a *well mixing* chain explores the space more flexibly than a poorly mixing chain.

The Gibbs Sampler

The *Gibbs sampler* is a special case of Metropolis-Hastings sampling where the random value is always accepted, that is to say, $\alpha = 1$. How can we construct a Markov chain whose values converge to the target distribution?

Let's consider a bivariate random variable (X, Y) and compute the marginals $p(x)$ and $p(y)$. The idea of Gibbs sampler is that we can avoid calculating the the integral of joint density $p(x, y)$ in order to obtain the marginals $p(x) = \int p(x, y)dy$ and $p(y) = \int p(x, y)dx$, which sometimes extremely hard. In stead, we can generate a sequence of conditional distributions $p(x|y)$ and $p(y|x)$. Firstly, the sampler choose the initial value y_0 for y to obtain x_0 by generating a random variable from $p(x|Y = y_0)$. Then the sampler uses the generated x_0 to obtain y_1 by generating a new random variable from $p(y|X = x_0)$, and then use the following functions to repeatedly update x_i and y_i:

$$x_i \sim p(x|Y = y_{i-1})$$
$$y_i \sim p(y|X = x_i).$$

A *Gibbs sequence* of length k would be generated after k iterations of the above process. The points (x_j, y_j) are then the simulated draws from the joint distribution. After a sufficient burn-in of removing the effects of initial sampling values, this Gibbs sequence will converge to a stationary distribution, which is independent of the starting values, and this stationary distribution is then the target distribution we are trying to simulate.

For multivariate random variable, we have the generalized version with the value of the k-th variable being drawn from the distribution $p(\theta^{(k)}|\Theta^{(-k)})$ where $\Theta^{(-k)}$ is a vector containing all the variables but k-th one. We use the following function to update the value of $\theta_i^{(k)}$:

$$\theta_i^{(k)} \sim p(\theta^{(k)}|\theta^{(1)} = \theta_i^{(1)}, \cdots, \theta^{(k-1)} = \theta_i^{(k-1)}, \theta^{(k+1)} = \theta_{i-1}^{(k+1)}, \cdots, \theta^{(n)} = \theta_{i-1}^{(n)}).$$

2.4 Hierarchical Spline Model Setup

We propose the following hierarchical models for different countries to set up the experiments.

2.4.1 Nigeria

Model 1 (National Model)

The first model for Nigeria ignores the area variable, and uses data from all the areas to fit a national model. In addition, it has a fixed effect for the constructed regression splines of time t with 3 degrees of freedom, denoted as $f_1(t), f_2(t), f_3(t)$, and a random intercept for sites, denoted as b_s, which follows a normal distribution, as follows:

$$\eta_{ast} = \beta_0 + \beta_1 f_1(t) + \beta_2 f_2(t) + \beta_3 f_3(t) + b_s,$$

where $b_s \sim N(0, \gamma)$ and β_i for $i = \{0, \cdots, 3\}$ are coefficients of the predictors.

Model 2 (By Area)

Model 2 treats each area independently. To simplify this, we add a area-specific intercept and use area as the slope of the constructed regression splines of time t with three degrees of freedom. We use the same notation as model 1 except that the subscript a represents each area. The model is as follows:

$$\eta_{ast} = \beta_{0a} + \beta_{1a} f_1(t) + \beta_{2a} f_2(t) + \beta_{3a} f_3(t) + b_s,$$

where $b_s \sim N(0, \gamma)$ as in model 1.

Model 3

Model 3 is a combination of model 1 and model 2 above. We use data from all the areas, but allow coefficients of the constructed spline bases variate randomly across areas. In other words, area is a random slope of the regression splines of time. Also, a fixed intercept of time and a fixed intercept of site are contained in the model. This model is the most complicated one we have so far.

$$\eta_{ast} = \beta_0 + \beta_1 f_1(t) + \beta_2 f_2(t) + \beta_3 f_3(t) + b_{0a} + b_{1a} f_1(t) + b_{2a} f_2(t) + b_{3a} f_3(t) + b_s,$$

where each b follow a different normal distribution as follows:

$$b_{0a} \sim N(0, \delta_0^2)$$

$$b_{1a} \sim N(0, \delta_1^2)$$

$$b_{2a} \sim N(0, \delta_2^2)$$

$$b_{3a} \sim N(0, \delta_3^2)$$

$$b_s \sim N(0, \gamma)$$

Note that when $\delta_0^2 = \delta_1^2 = \delta_2^2 = \delta_3^2 = 0$, this model is the same as model 1. However, when $\delta_0^2 = \delta_1^2 = \delta_2^2 = \delta_3^2 \to +\infty$, this model becomes model 2.

2.4.2 Papua New Guinea

We propose exactly the same models for Papua New Guinea as the models for Nigeria. To be concise, we don't include the details of the models in this section. Please use the models for Nigeria as a reference.

2.4.3 Vietnam

Model 1 (National Model)

Model 1 for Vietnam is a national model, constructed in the same way as model 1 for Nigeria. This model fits the time trend for all high risk groups and all areas, using the entire data set. To briefly sum up, this model has a fixed effect of the constructed regression splines of time t with 3 degrees of freedom, and random intercept for sites that follows a normal distribution, as the following:

$$\eta_{astr} = \beta_0 + \beta_1 f_1(t) + \beta_2 f_2(t) + \beta_3 f_3(t) + b_s,$$

where $b_s \sim N(0, \gamma)$ and β_i for $i = \{0, \cdots, 3\}$ are coefficients of the predictors.

Model 2 (By Risk Group)

Model 2 treats each risk group independently, equivalently, we add a risk-group-specific intercept and use risk group as the slope of constructed regression splines of time t. We use the same notation as model 1 except that the subscript r represents each risk group. The model is as follows:

$$\eta_{astr} = \beta_{0r} + \beta_{1r} f_1(t) + \beta_{2r} f_2(t) + \beta_{3r} f_3(t) + b_s,$$

where $b_s \sim N(0, \gamma)$ as in model 1 and the coefficients β are dependent on each risk group.

Model 3

Model 3 is almost the same as model 4, substituting the random effect of area as an intercept and as a slope for the constructed regression splines of time t, with the fixed effect instead. Please see the details of model 4 for a reference to model 3.

Model 4

Model 4 uses risk group as the slope for the regression splines of time, risk group and site as fixed intercepts, and area as a random slope for the constructed regression splines of time t. This is the most complicated model for Vietname in this paper.

$$\eta_{astr} = \beta_{0r} + \beta_{1r} f_1(t) + \beta_{2r} f_2(t) + \beta_{3r} f_3(t) + b_s + b_s + b_{0a} + b_{1a} f_1(t) + b_{2a} f_2(t) + b_{3a} f_3(t),$$

where the coefficients β are dependent on the risk group and b follows different normal distributions as in model 3 for Nigeria.

Chapter 3

Results

3.1 Example of generalized epidemic – Nigeria

Nigeria is an example country with generalized epidemic we use for experiments. This section presents the outputs, including tables and plots, from model 1, 2, and 3 for Nigeria. Note that we don't include all the coefficients for the model to simplify the visualization.

	Intercept	β_1	β_2	β_3	DIC
Model 1	-4.33327172	2.055973606	0.730303578	1.086334386	74837.73125
Model 2	1.585937922	-8.056664571	-4.241963527	-4.442548726	74825.51
Model 3	-4.330350151	2.140001954	0.722415294	1.060765772	74823.23

Table 3.1: Models Results for Nigeria

We use deviance information criterion (DIC) as a model evaluation method in this paper. As seen from the table above, model 3 has the lowest DIC, that is, this model fits best on the HIV prevalence data for Nigeria. This model is the most complicated one but it contains many other considerations that the first two models do not offer.

The remaining of this section are the plots of different model fits by different areas of Nigeria.

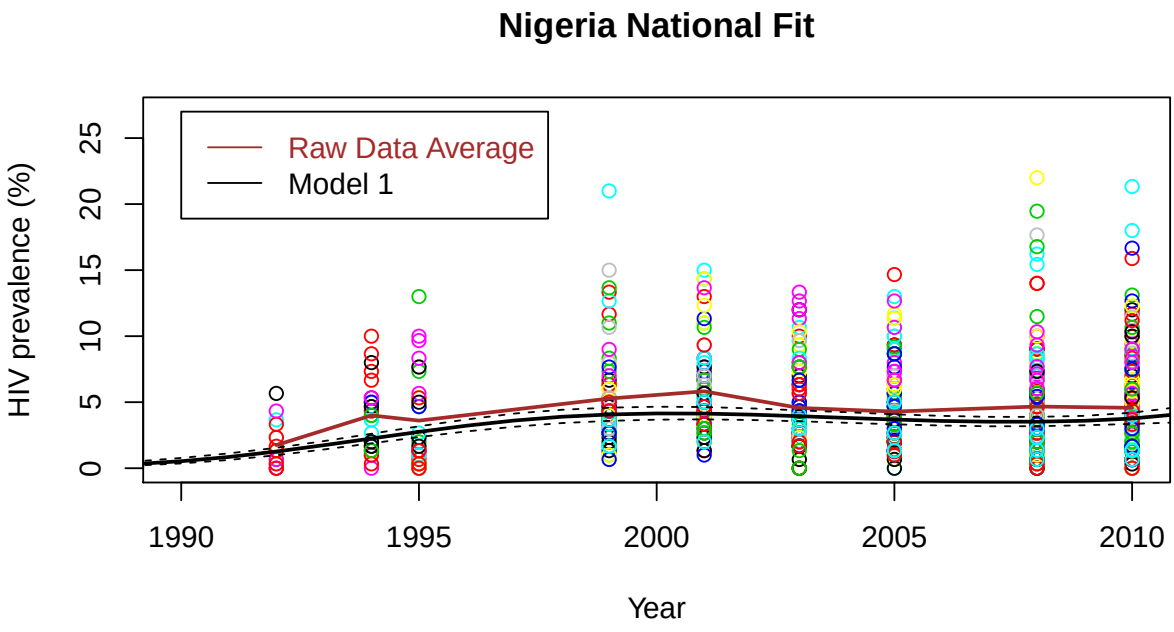

Figure 3.1: National Model for Nigeria

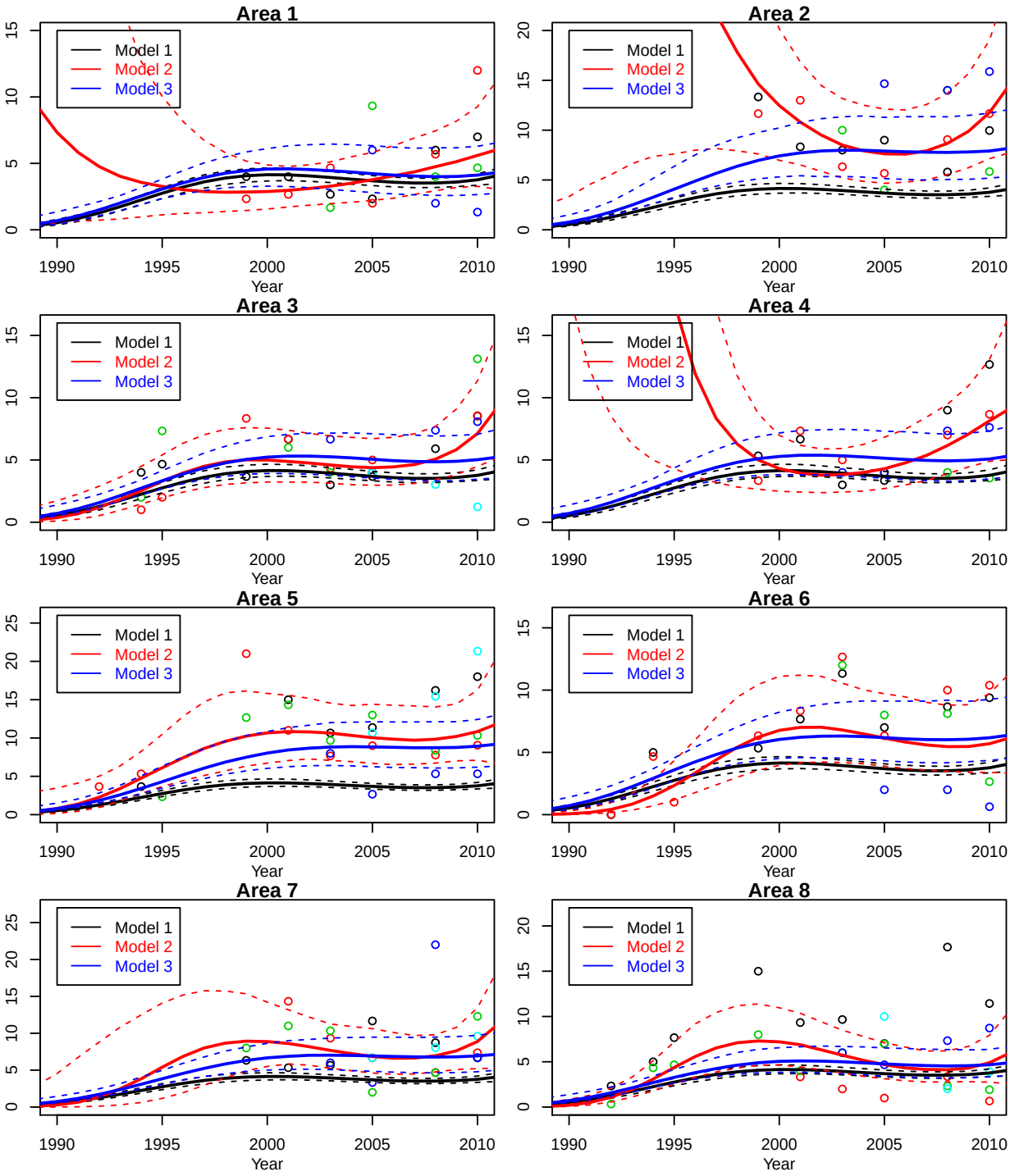

Figure 3.2: Model Fits for Nigeria by Area (1)

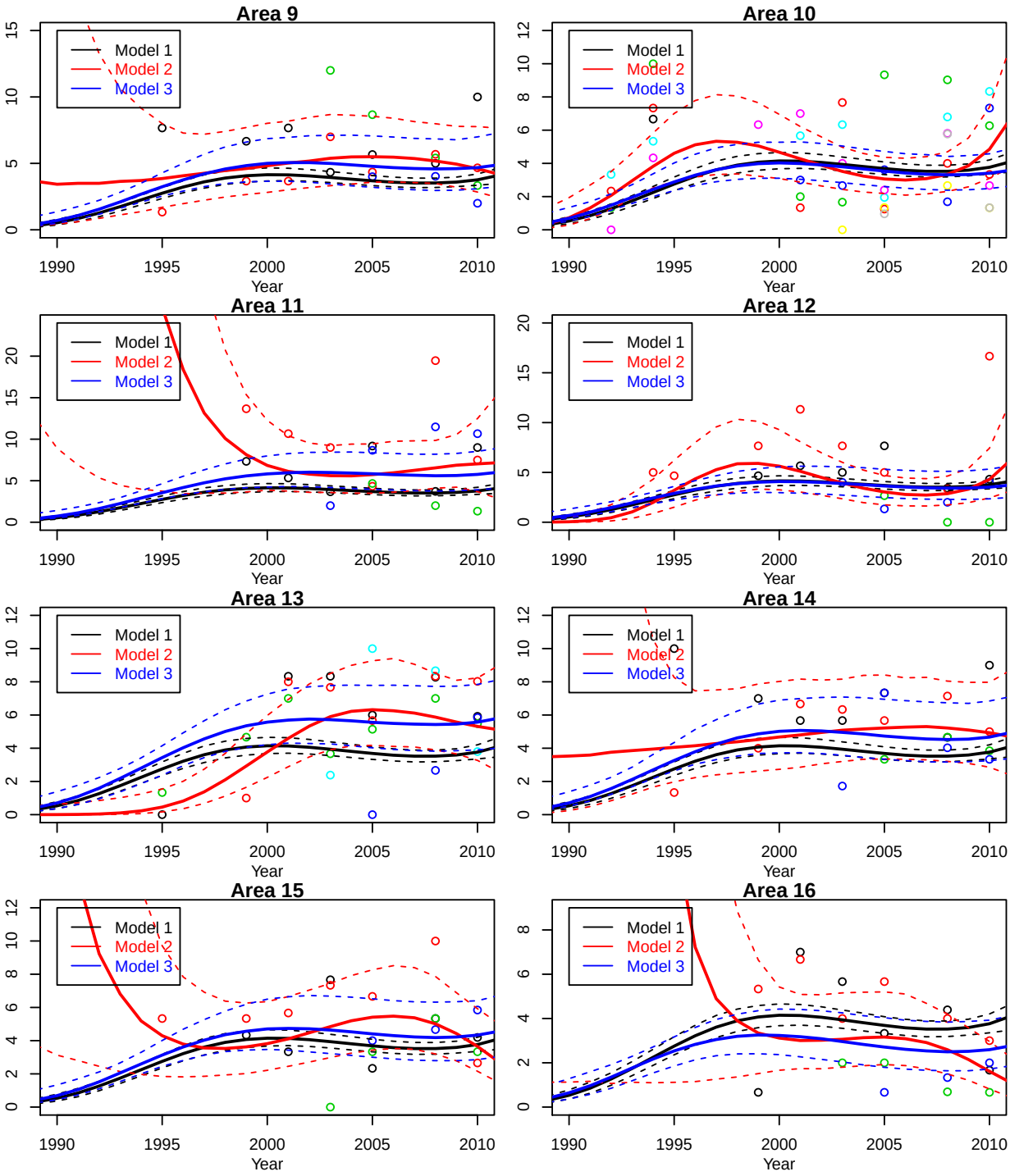

Figure 3.3: Model Fits for Nigeria by Area (2)

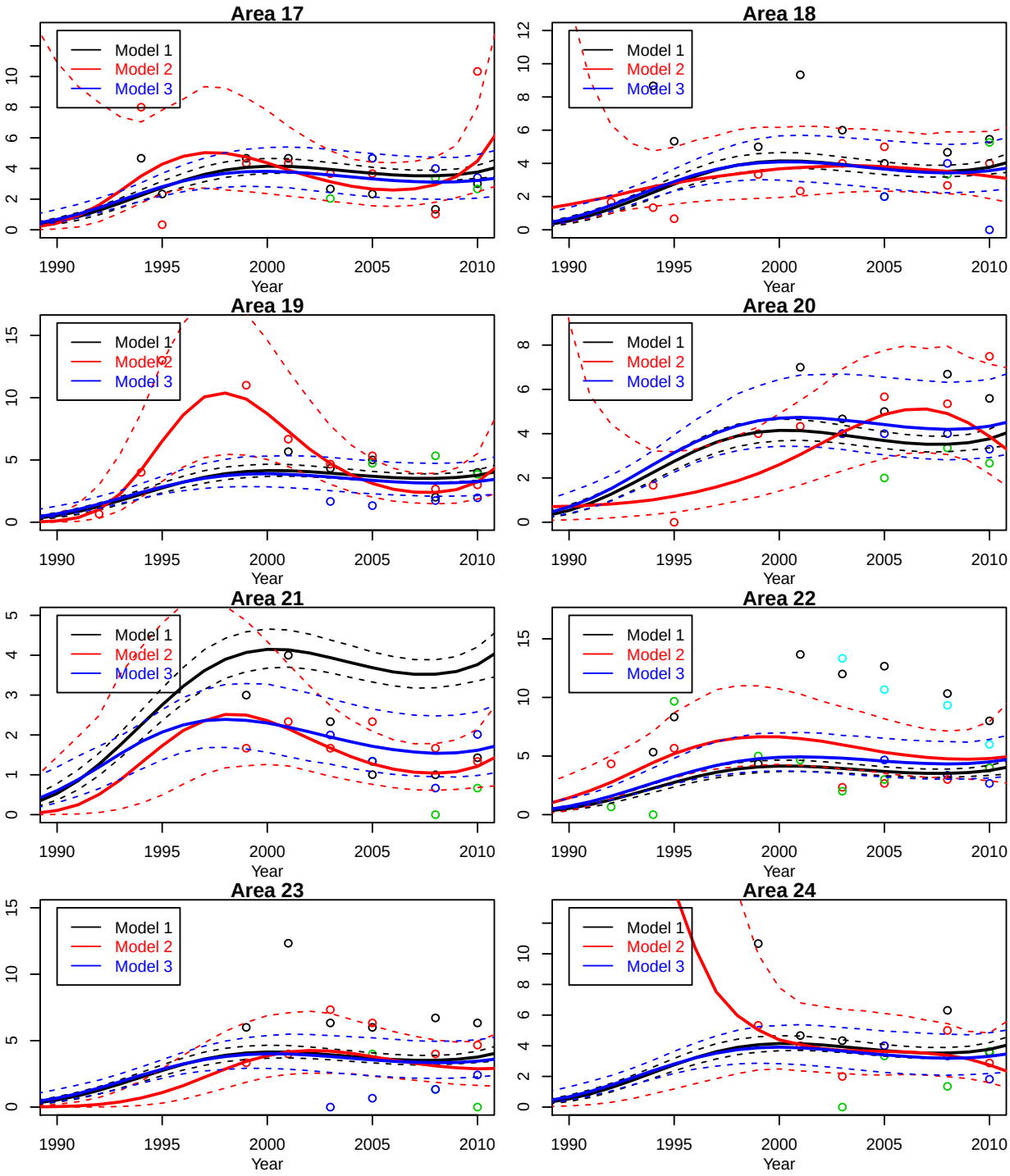

Figure 3.4: Model Fits for Nigeria by Area (3)

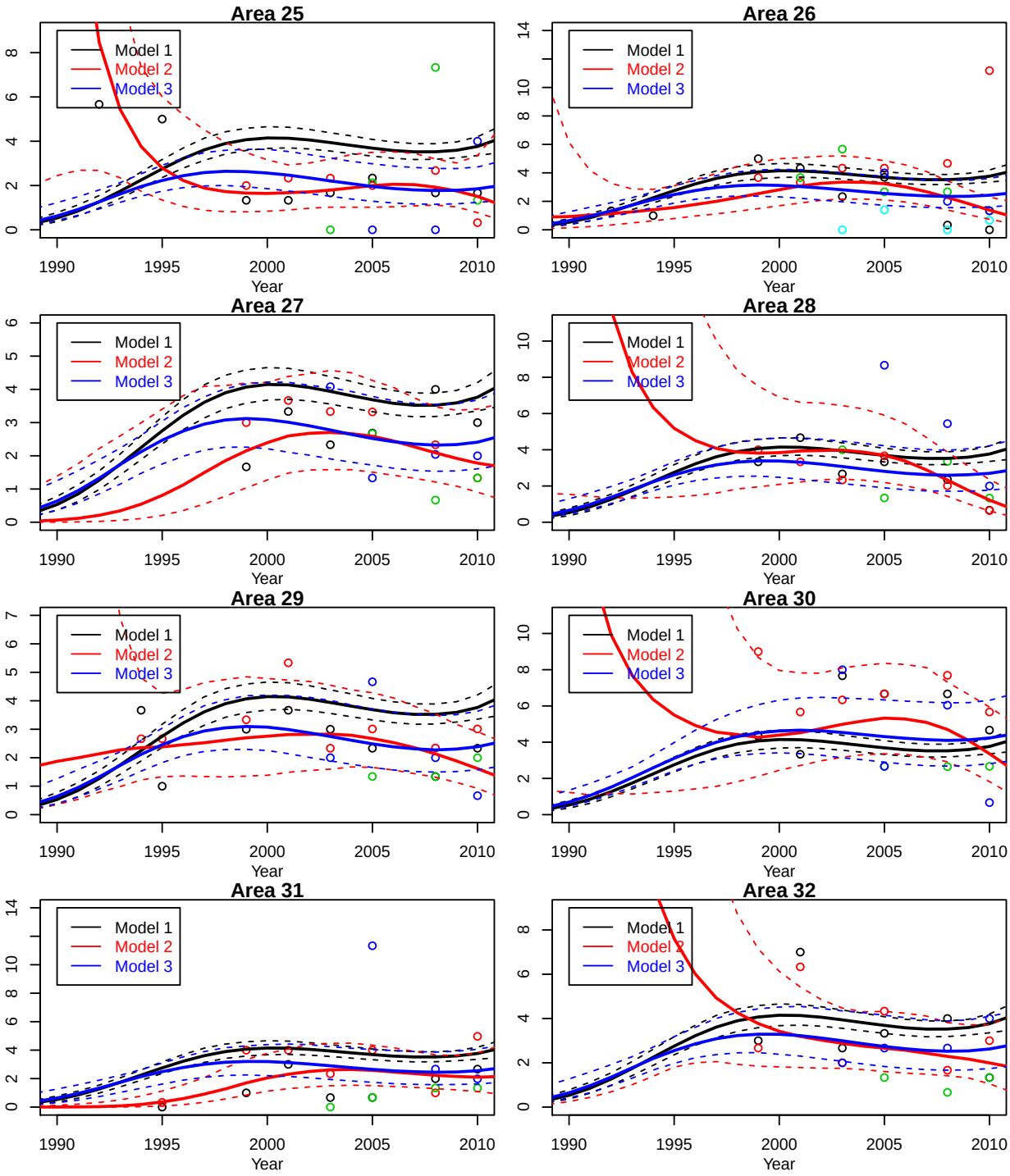

Figure 3.5: Model Fits for Nigeria by Area (4)

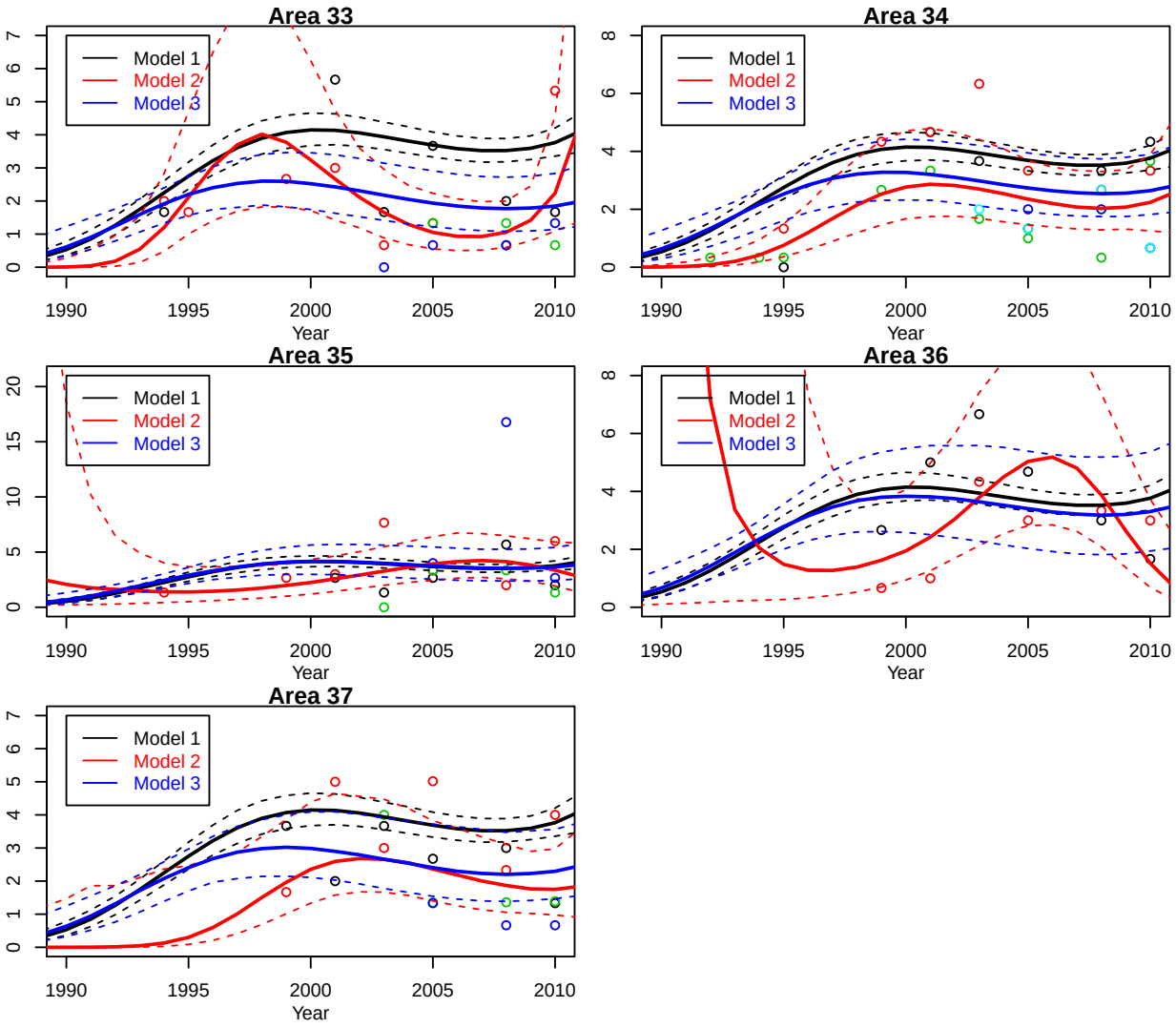

Figure 3.6: Model Fits for Nigeria by Area (5)

3.2 Examples of concentrated epidemics – Papua New Guinea and Vietnam

Papua New Guinea and Vietnam are the two example countries with concentrated epidemic we use for experiments. This section begins with the model outputs and plots from Model 1, 2, and 3 for Papua New Guinea. After that the model outputs for Vietnam will also be presented. Note that we do not include all the coefficients for the model to simplify the visualization.

	Intercept	β_1	β_2	β_3	DIC
Model 1	-7.09178948	4.2207927	-0.545826533	0.864442356	20001.73
Model 2	-0.676220464	-3.294938989	-7.97323281	-5.087105565	19998.86
Model 3	-7.1919637	4.3584340	-0.4484656	0.9878673	19984.83

Table 3.2: Models Results for Papua New Guinea

From Table 3.2, we can see that model 3 has the lowest DIC and thus this model fits best on the HIV prevalence data for Papua New Guinea. This model is the most complicated model among other models we proposed in this paper.

The remaining of this section are the plots of different model fits by different areas of Papua New Guinea.

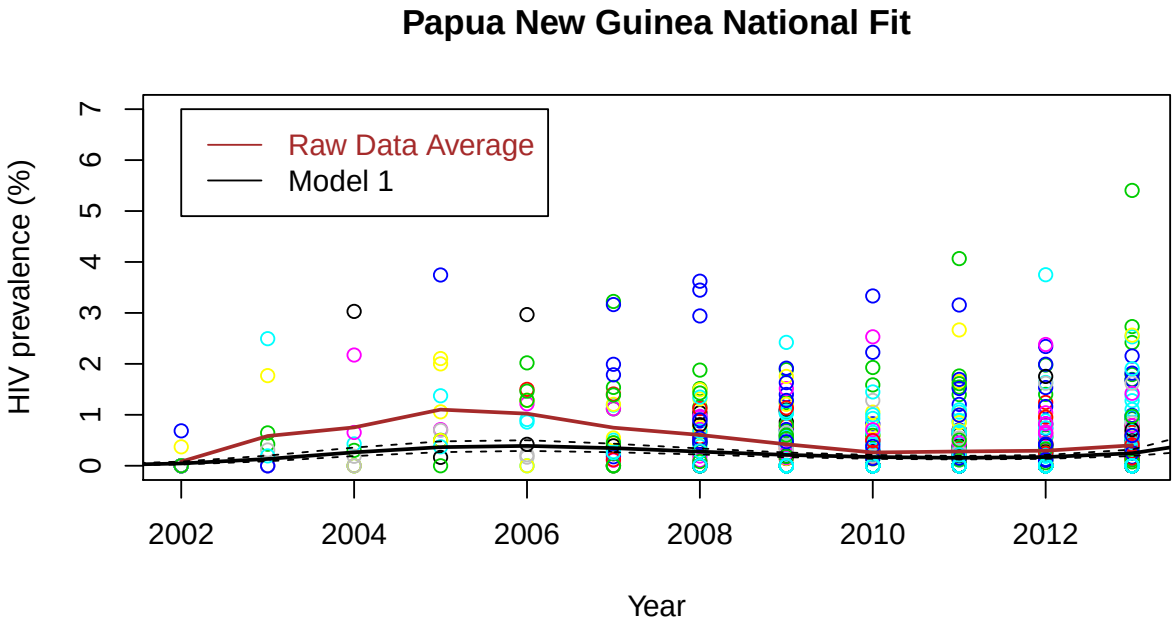

Figure 3.7: National Model for Papua New Guinea

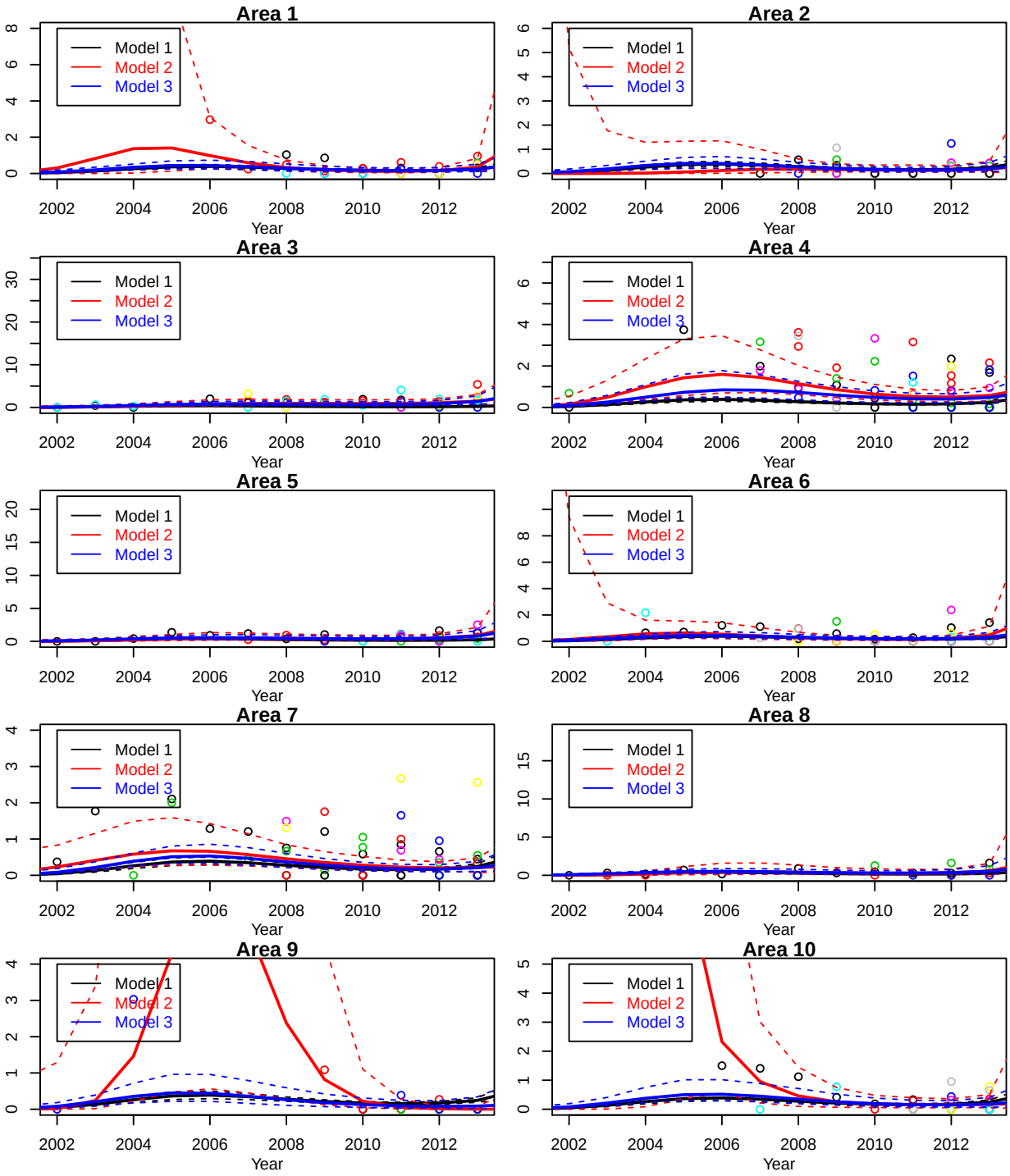

Figure 3.8: Model Fits for Papua New Guinea by Area (1)

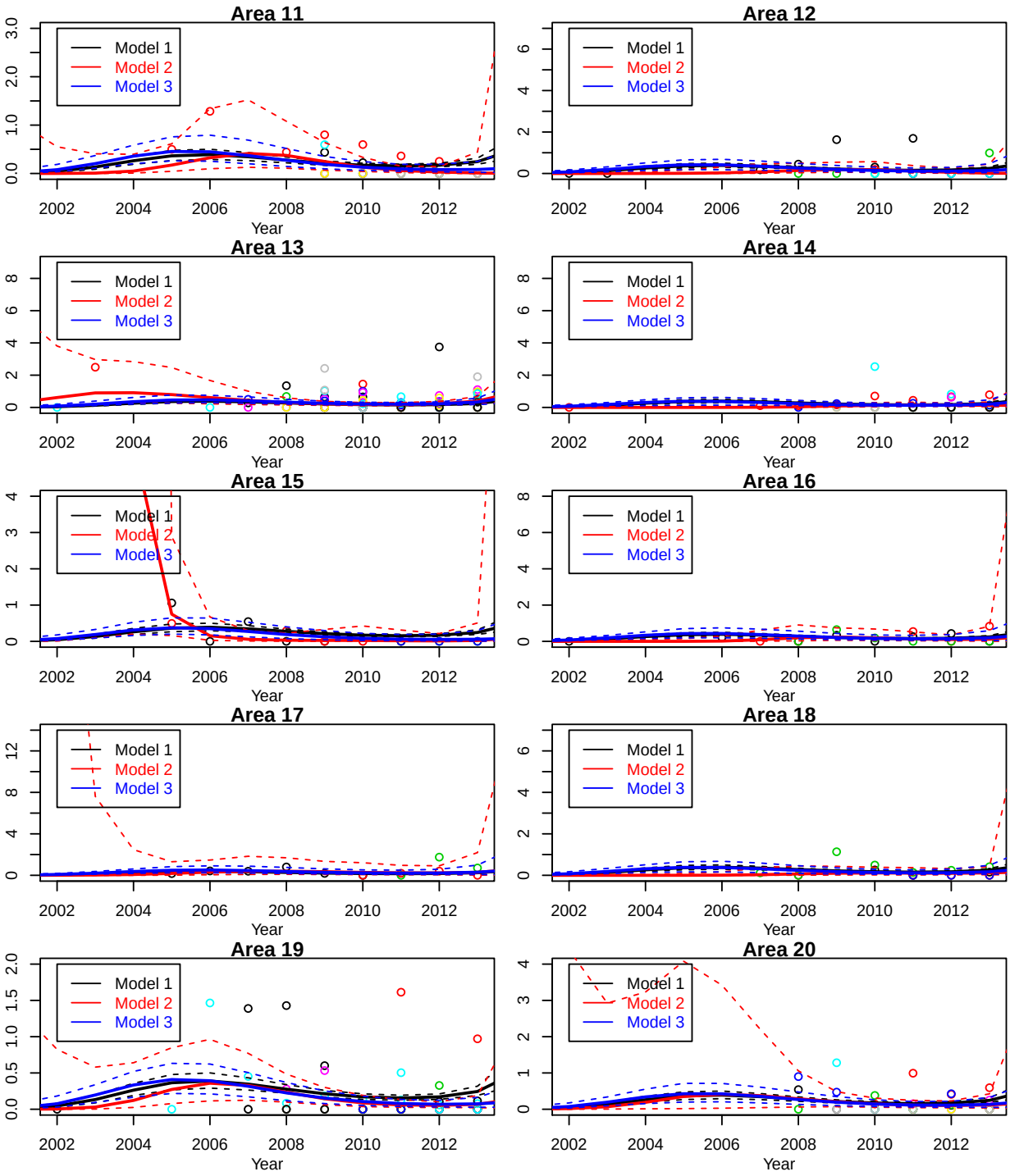

Figure 3.9: Model Fits for Papua New Guinea by Area (2)

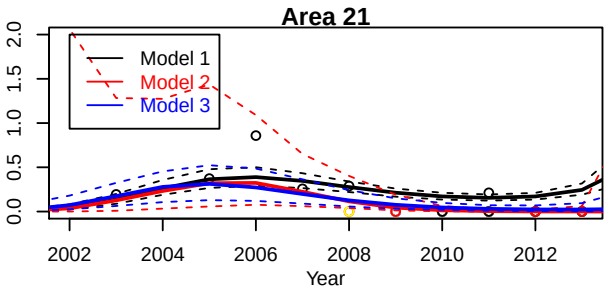

Figure 3.10: Model Fits for Papua New Guinea by Area (3)

The following table results and plots are from the models for Vietnam. From the table, we can see that model 3 has the lowest DIC and it is therefore the best model that fits the original data well for Vietnam.

	Intercept	β_1	β_2	β_3	DIC
Model 1	-16.30373	16.27729	14.09275	0.13.10081	258949.8
Model 2	-16.9883735	18.3063631	16.6173750	14.4272571	258946.4
Model 3	-34.7317078	46.0364087	31.5767126	34.4848596	258874.8
Model 4	-16.316815	17.636297	16.488511	13.955919	258901.7

Table 3.3: Models Results for Vietnam

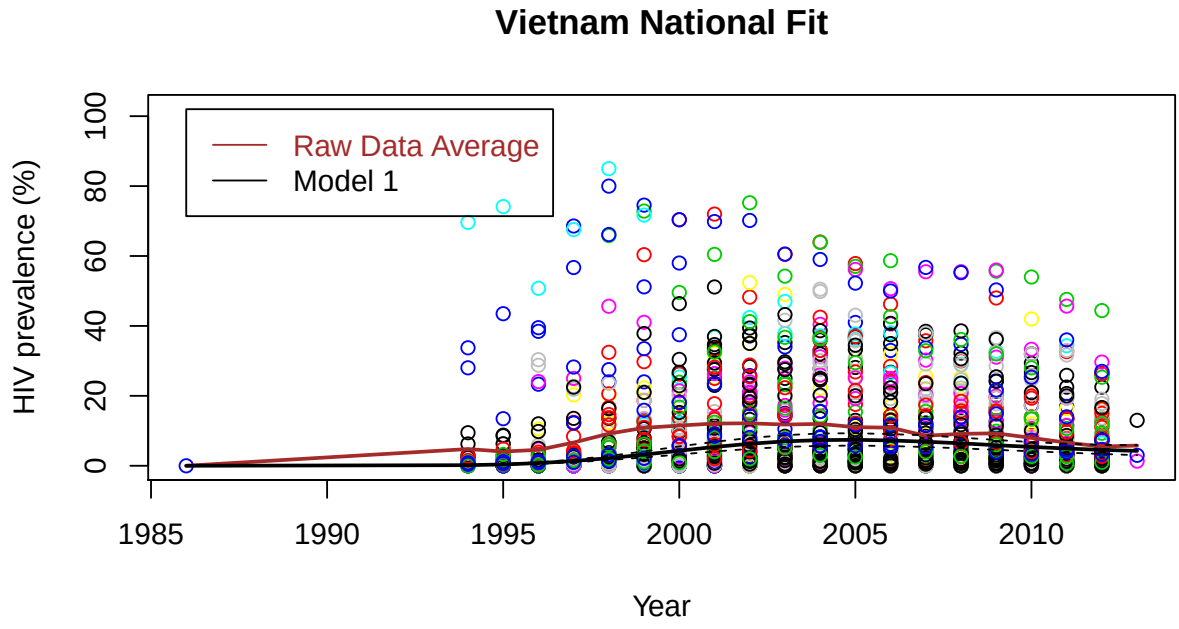

Figure 3.11: National Model for Vietnam

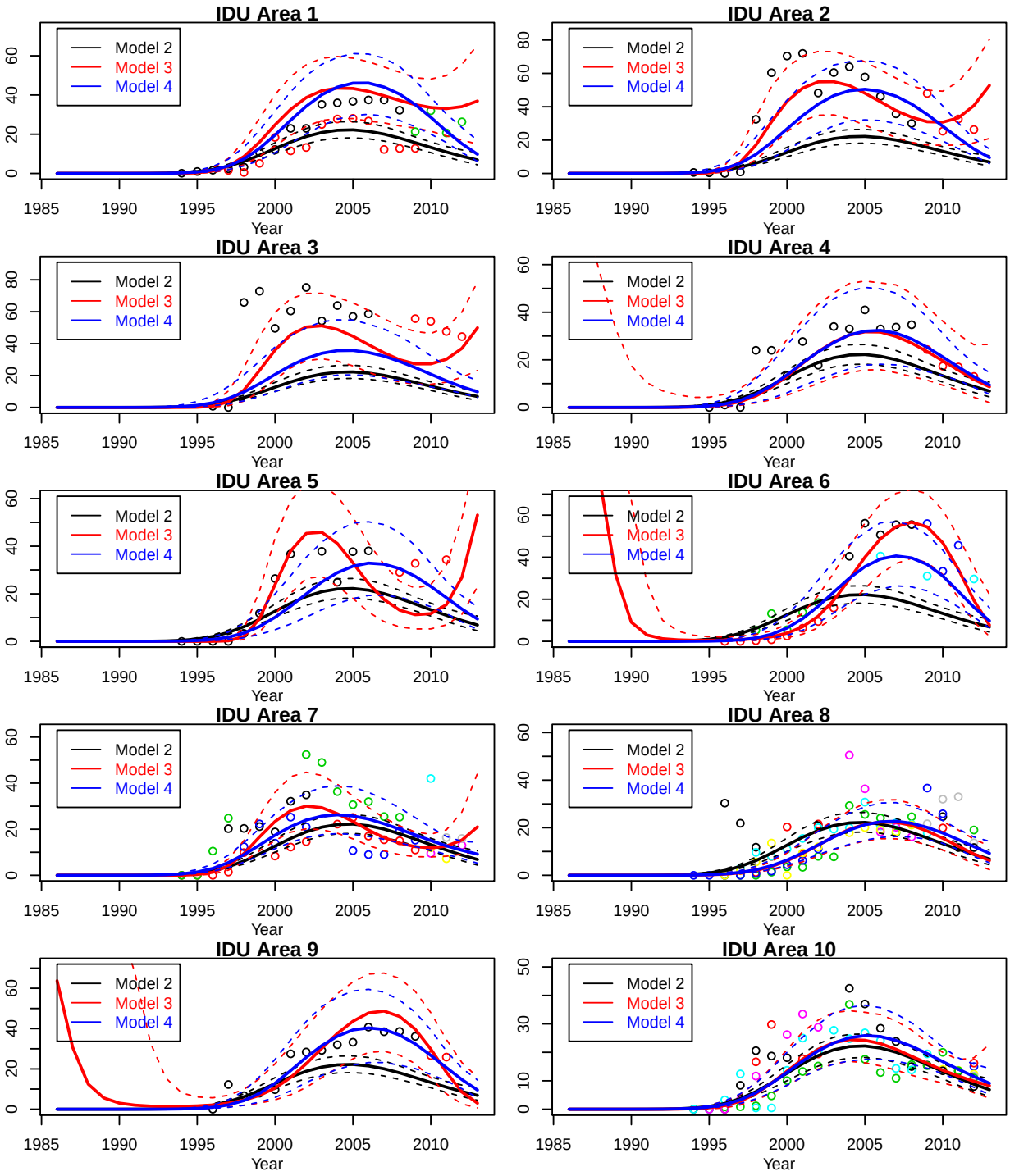

Figure 3.12: Model Fits for Vietnam by Area for Injecting Drug Users (IDU) (1)

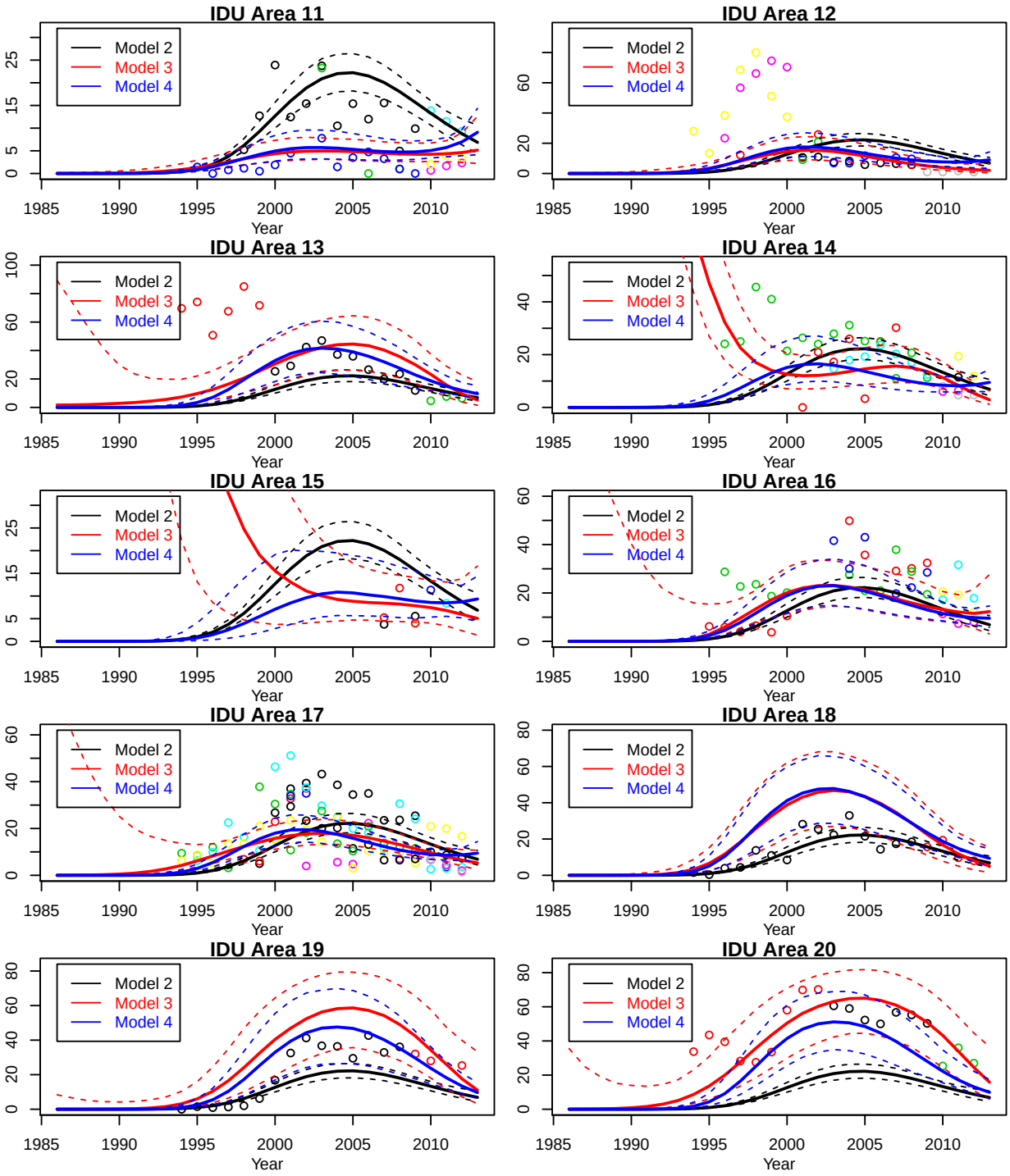

Figure 3.13: Model Fits for Vietnam by Area for Injecting Drug Users (IDU) (2)

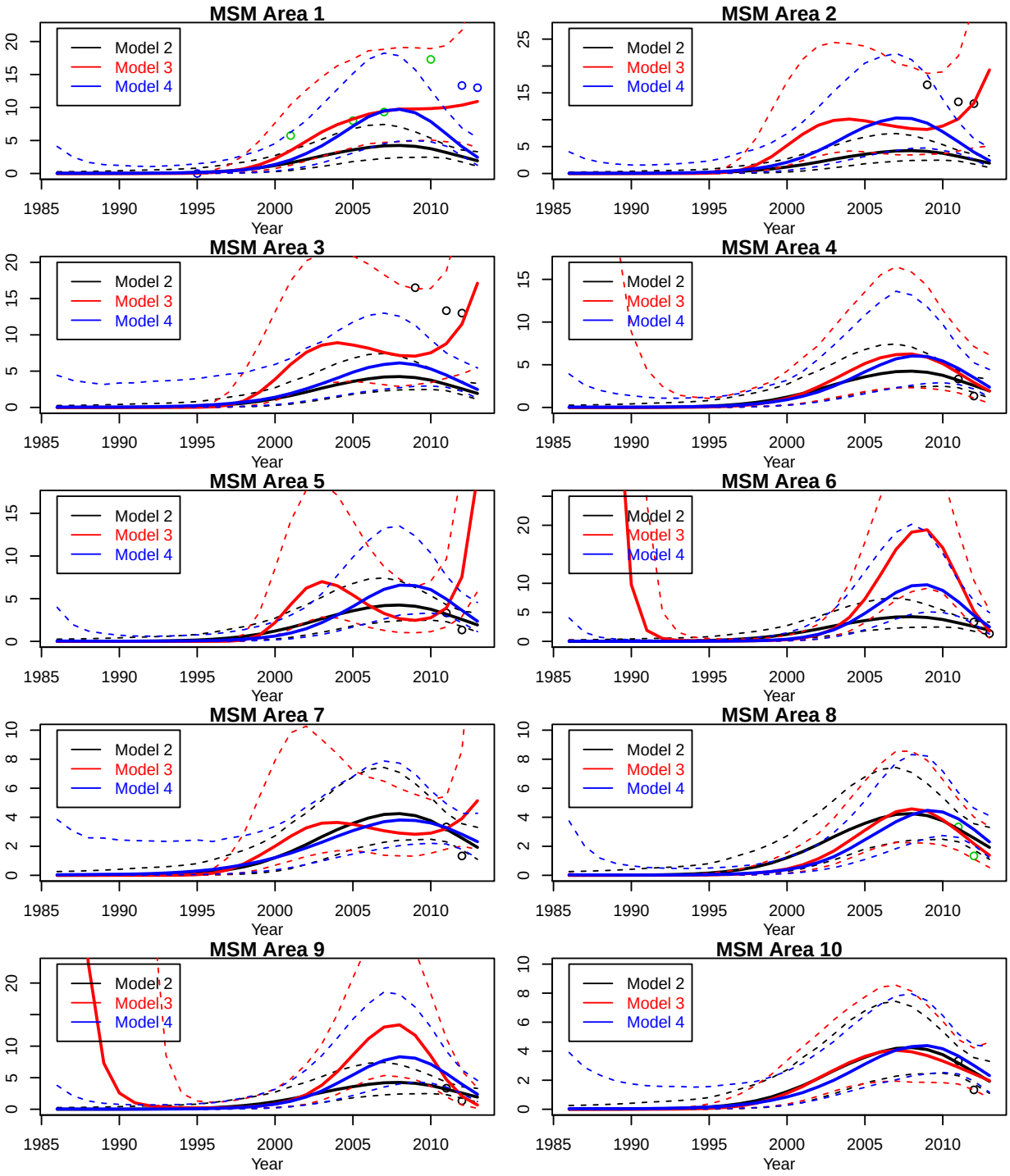

Figure 3.14: Model Fits for Vietnam by Area for Men Who Have Sex with Men (MSM) (1)

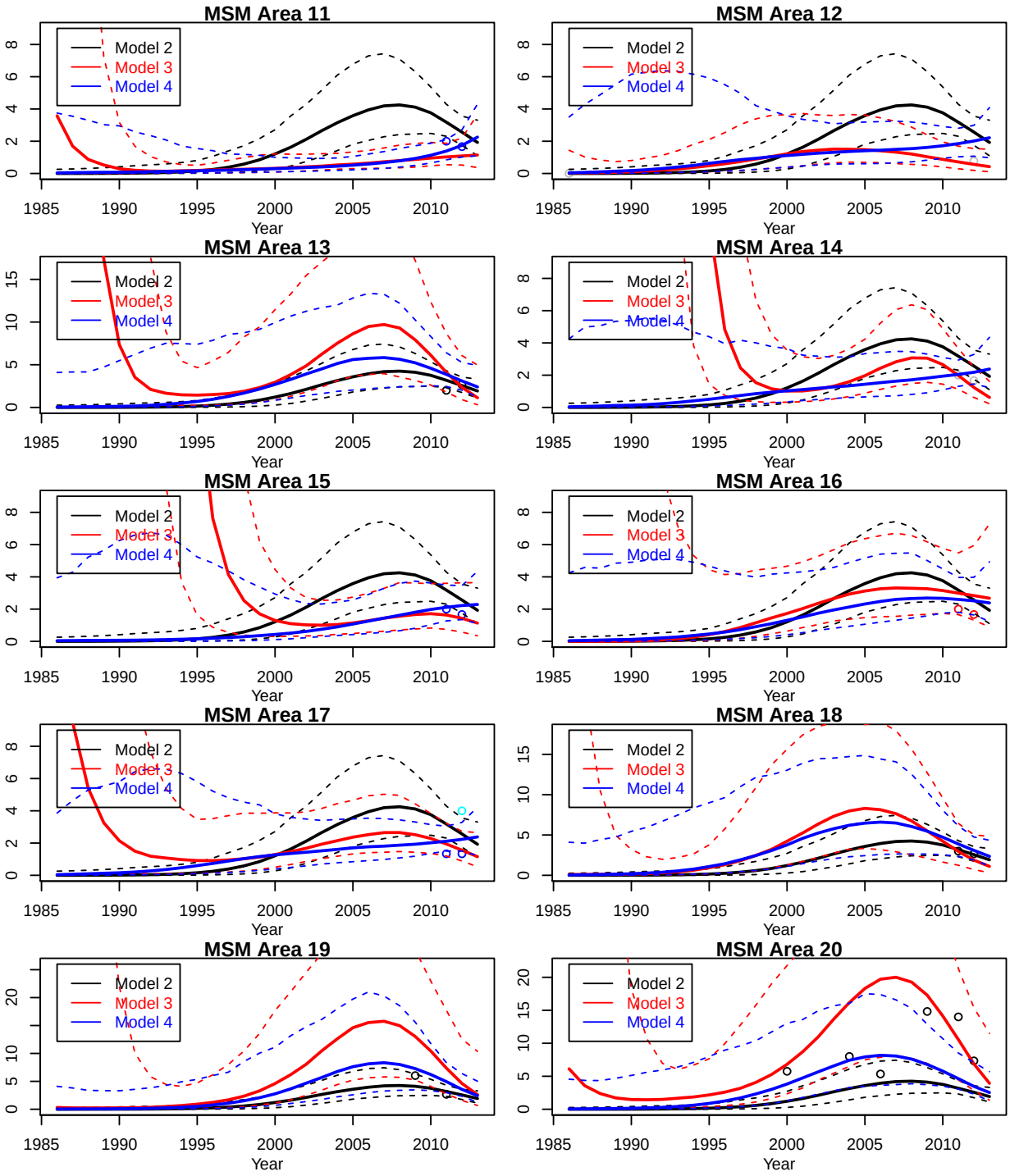

Figure 3.15: Model Fits for Vietnam by Area for Men Who Have Sex with Men (MSM) (2)

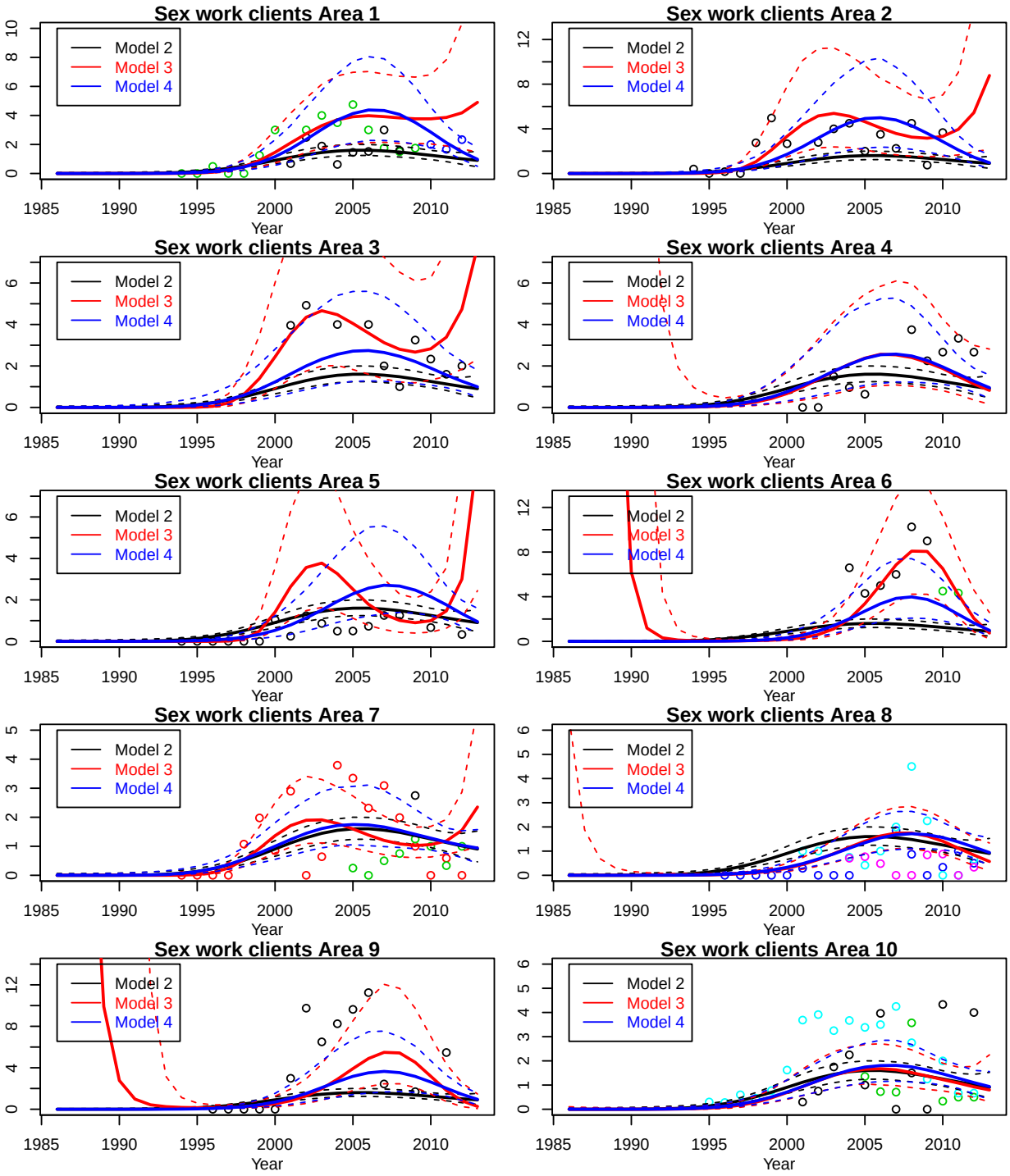

Figure 3.16: Model Fits for Vietnam by Area for Sex Work Clients (1)

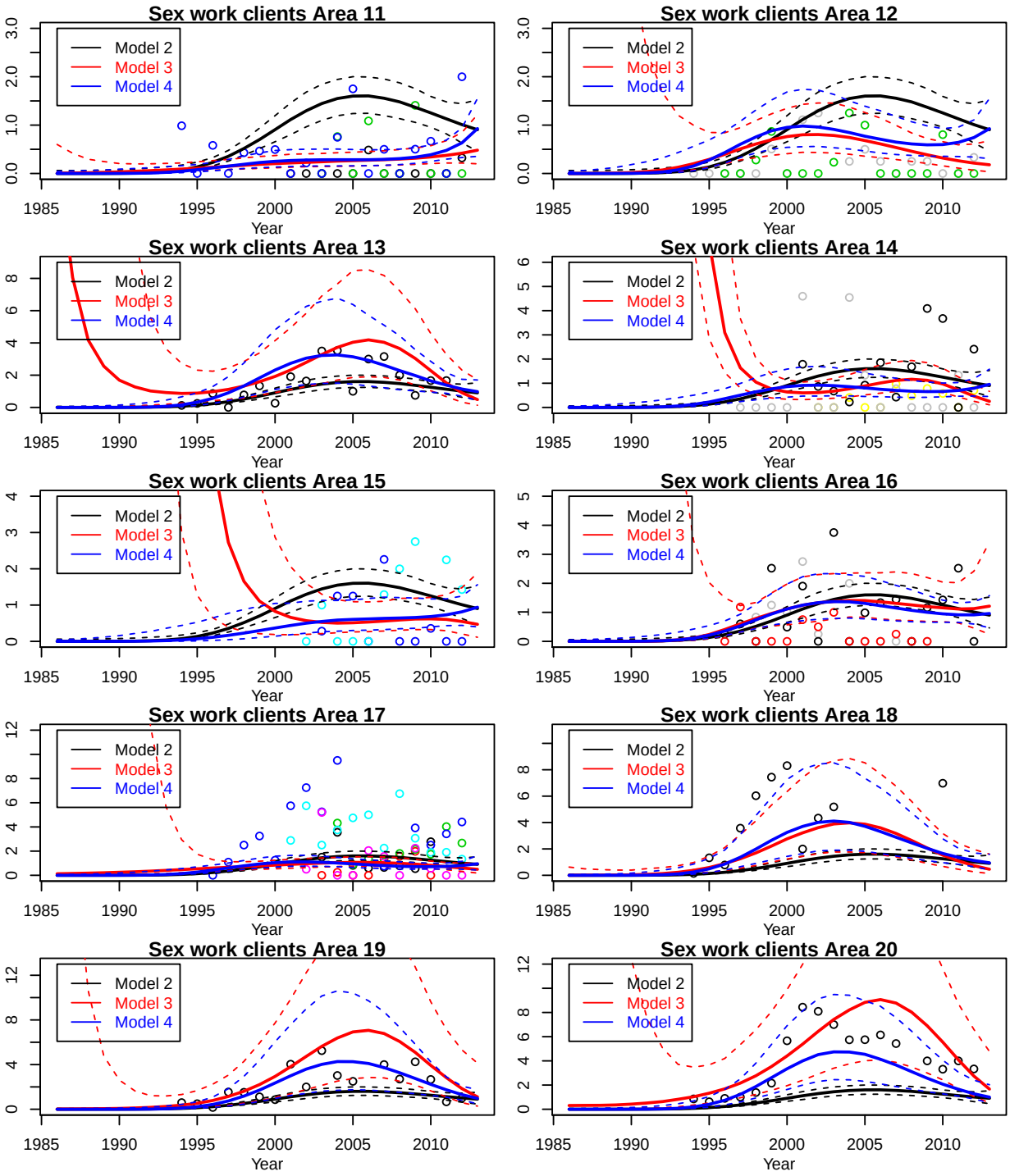

Figure 3.17: Model Fits for Vietnam by Area for Sex Work Clients (2)

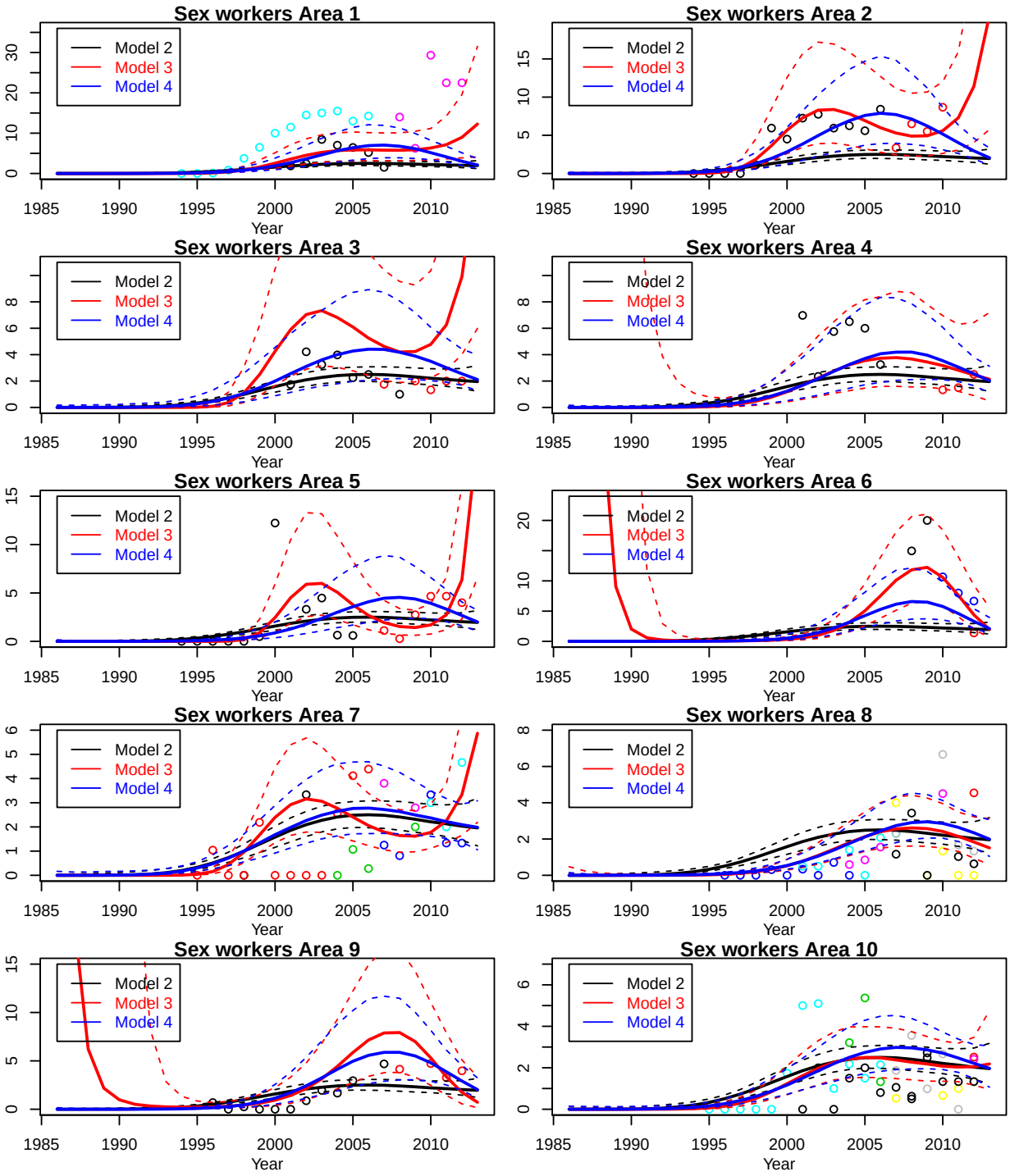

Figure 3.18: Model Fits for Vietnam by Area for Sex Workers (1)

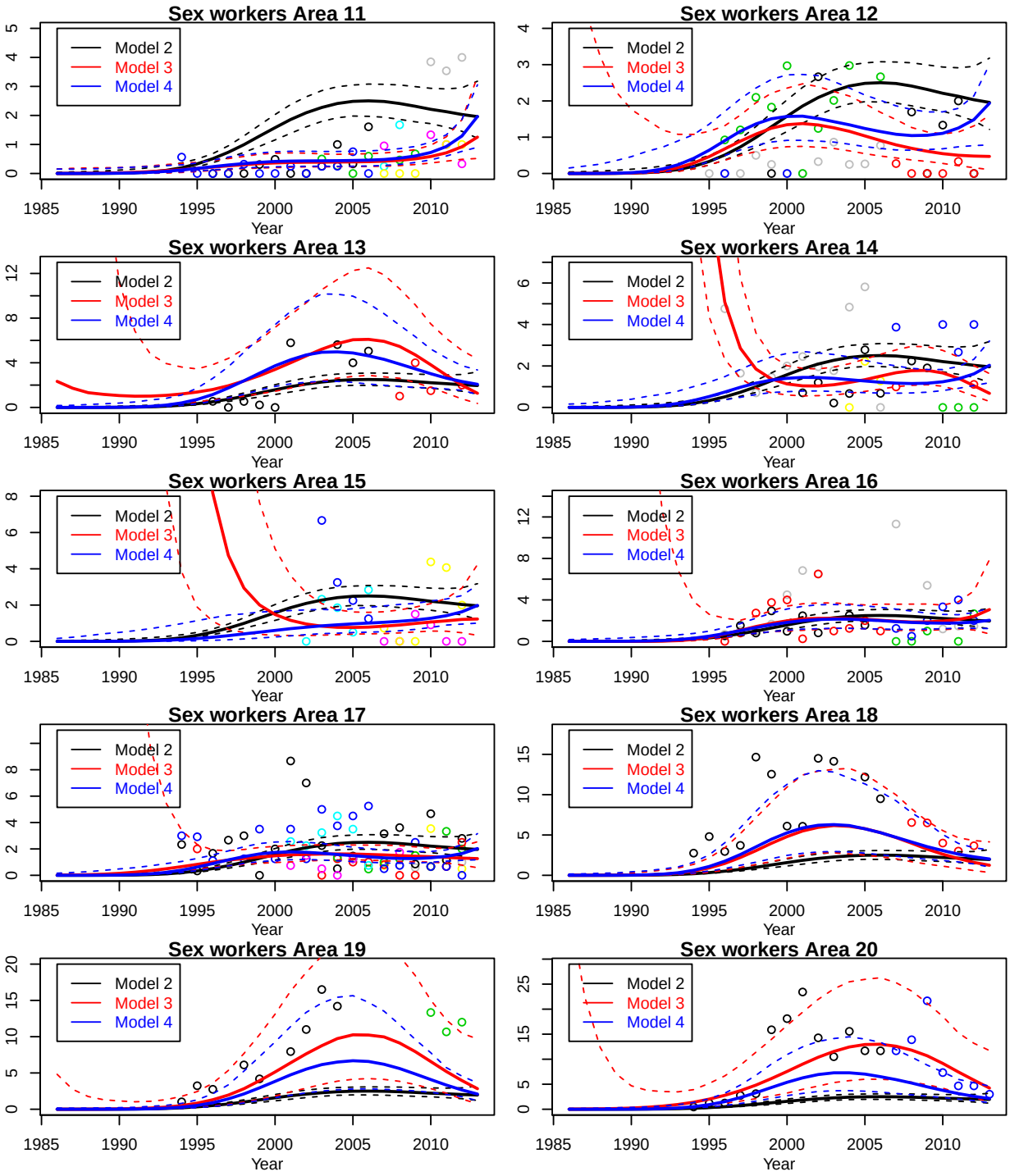

Figure 3.19: Model Fits for Vietnam by Area for Sex Workers (2)

Chapter 4

Discussion

4.1 Future Work

In this paper, we use deviance information criterion (DIC) as the evaluation method of different models fits. However, DIC assumes that the specific parametric family of probability distrbutions that generate future observations encompasses the true model, which does not always hold true. Also, DIC uses the whole data set for both training and testing, which tends to select over-fitted models as the optimal model. There are other evaluation methods that can be used to better evaluate some other perspectives of the models. For instance, *cross-validation* can assess how the results of the statistical model generalize to different independent data set from the original data. In other words, it evaluates how accurately a predictive model performs in practice, which is the type of information that DIC does not infer. Also, some of the models differ only slightly in the DIC, and thus we cannot choose the optimal model with confidence. We will consider other evaluation methods to assess our models in the future.

Also, we noticed from the figures for the model results by area that the boundaries of the fitted curve are not stable or smooth. In other words, extremely high variance occurs at the boundary regions in the models. In the future, we will consider additional boundary constraints so that the models are linear at the boundaries, such as fitting natural splines, which produce a more stable results at the boundaries.

Bibliography

Brown, T., L. Bao, J. W. Eaton, D. R. Hogan, M. Mahy, K. Marsh, B. M. Mathers, and R. Puckett (2014). Improvements in prevalence trend fitting and incidence estimation in epp 2013. AIDS 28(Suppl 4), S415S426.

Brumback, Babette A., and John A. Rice. "Smoothing spline models for the analysis of nested and crossed samples of curves." Journal of the American Statistical Association 93.443 (1998): 961-976.

Eilers, Paul HC, and Brian D. Marx. "Flexible smoothing with B-splines and penalties." Statistical science (1996): 89-102.

Forsey, David R., and Richard H. Bartels. "Hierarchical B-spline refinement." ACM SIGGRAPH Computer Graphics. Vol. 22. No. 4. ACM, 1988.

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: springer, 2009.

Hoff, Peter D. A first course in Bayesian statistical methods. New York: Springer, 2009.

Huang, Yangxin, Dacheng Liu, and Hulin Wu. "Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system." Biometrics62.2 (2006): 413-423.

"National HIV Estimates File." UNAIDS. Web. 30 Mar. 2015. <http://apps.unaids.org/spectrum/>.

Raudenbush, Stephen, and Anthony S. Bryk. "A hierarchical model for studying school effects." Sociology of education (1986): 1-17.

Shire, Norah J., Jeffrey A. Welge, and Kenneth E. Sherman. "Efficacy of inactivated hepatitis A vaccine in HIV-infected patients: a hierarchical bayesian meta-analysis." Vaccine 24.3 (2006): 272-279.

Silverman, Bernhard W. "Some aspects of the spline smoothing approach to non-parametric regression curve fitting." Journal of the Royal Statistical Society. Series B (Methodological) (1985): 1-52.

ACADEMIC VITA

Yuan Tang
terrytangyuan@gmail.com
Unit 11, 915 Southgate Dr.,
State College, PA 16801

Education:

Schreyer Honors College, Pennsylvania State University, US
B.S. in Mathematics with honors (Sep, 2012 - May, 2015)

Work Experience:

DataNovo, San Francisco, CA — Summer 2014-present
Data Scientist/Consultant

Network and Data Mining lab, Penn State — Spring 2014
Undergraduate Research Assistant

Nittany Success Center, Penn State York — Fall 2013 (3 semesters)
Tutor for Multivariate Calculus and Introductory Statistics

Hedge's Biology Lab, Penn State — Fall 2013
Image Processing Programmer

Awards/Honors/Scholarships:

Penn State Dean's list (5 consecutive semesters)
Best Virtual Reality Hack, HackRPI — Fall, 2014
Mu Sigma Rho National Statistics Honorary Society Inductee — Fall, 2013
Penn State PMASS fellowship, by Department of Mathematics — Spring, 2014
MindSumo Programming Challenge Winner — Summer, 2014
& Summit for Software Engineers hosted by Capital One

Awarded by Schreyer Honors College:
John K. Tsui Honors Scholarship — Spring 2014
Pre-Eminence in Honors Education Fund — Summer, 2014
Summer Research Grants — Summer, 2014

Leadership/Activities:

Mentor, HackPSU, Penn State University — Spring, 2015
President/founder, Penn State York Paddlers' Table Tennis Club — Spring, 2013
Mentor, Penn State York Mentoring Program — Spring 2013
Campus Ambassador, MindSumo, San Francisco, CA — Summer, 2014
Summer Program in Quantitative Methods of Social Research @University of Michigan — Summer, 2014

Miscellaneous Affiliations and Involvements: (since college began)
Association for Computing Machinery; Distinguished Honors Faculty Program Society of Distinguished Alumni Mentoring Program; Innovation Interdisciplinary Collaboration Community; First Music Association; Badminton Club; Innoblue Accelerator; New Leaf Initiative; PNC Leadership Assessment Center