HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2021 May 13.

Published in final edited form as:

Stat Med. 2012 November 30; 31(27): 3278–3284. doi:10.1002/sim.5343.

A Bayesian analysis of the 2009 decline in tuberculosis morbidity in the United States

Michael P. Chen[*,†], **Nong Shang**, **Carla A. Winston**, **Jose E. Becerra**

Division of Tuberculosis Elimination, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, GA, 30329 U.S.A.

Abstract

Although annual data are commonly used to model linear trends and changes in trends of disease incidence, monthly data could provide additional resolution for statistical inferences. Because monthly data may exhibit seasonal patterns, we need to consider seasonally adjusted models, which can be theoretically complex and computationally intensive. We propose a combination of methods to reduce the complexity of modeling seasonal data and to provide estimates for a change in trend when the timing and magnitude of the change are unknown. To assess potential changes in trend, we first used autoregressive integrated moving average (ARIMA) models to analyze the residuals and forecast errors, followed by multiple ARIMA intervention models to estimate the timing and magnitude of the change. Because the variable corresponding to time of change is not a statistical parameter, its confidence bounds cannot be estimated by intervention models. To model timing of change and its credible interval, we developed a Bayesian technique. We avoided the need for computationally intensive simulations by deriving a closed form for the posterior distribution of the time of change. Using a combination of ARIMA and Bayesian methods, we estimated the timing and magnitude of change in trend for tuberculosis cases in the United States. Published 2012. This article is a US Government work and is in the public domain in the USA.

Keywords

tuberculosis trends; time series analysis; seasonality; ARIMA; intervention models; change point; Bayesian analysis

1. Introduction

The incidence rate of tuberculosis (TB) in the United States declined from 52.6 cases per 100,000 persons in 1953 to 5.8/100,000 in 2000 [1, 2]. Since 2000, TB incidence rates steadily declined at an average annual percent change of −3.8% [3]. In 2009, however, a sharp and unexpected incidence decline of −11.4% (3.8/100,000) was observed in the National Tuberculosis Surveillance System (NTSS) [4], followed by a decline of −3.9% in 2010 (3.6/100,000) [2]. Moreover, the observed annual TB case count in 2009 and 2010 fell

[*]Correspondence to: Michael Chen, Division of TB Elimination, CDC Atlanta, GA 30329, U.S.A. [†] mchen1@cdc.gov.

outside of 95% prediction bounds of the log transformed linear trend of 2000–2008 (Figure 1).

Considering the steady decline in TB trend in previous years, we were interested to determine what statistical inference about the change in trend, and the timing and magnitude of the change, could be drawn. However, with only two newly observed annual data points outside of the prediction bounds of the linear trend, limited statistical inference was possible. Therefore, we used detailed monthly case count data [5] from the NTSS to increase the resolution for analysis and allow for better inference.

Tuberculosis cases reported to the NTSS from January 1, 2000 through December 31, 2010 were included for analysis if TB treatment had started within the same reporting year. The use of the treatment start date and the report date for the selection of the TB cases ensured comparable monthly data throughout the study period, particularly for the last months of 2010 in comparison to earlier years. Cases were aggregated by month of treatment start date. As shown in Figure 2, the monthly TB data exhibited a seasonal pattern.

Because monthly time series data were seasonal, linear trend analysis without seasonal adjustment would not have been appropriate and more sophisticated models were required. To estimate the trend and change in trends in a time series data, autoregressive integrated moving average (ARIMA) models can be explored [5–8]; intervention analysis [8–10] can be used to estimate the magnitude of the change when the intervention timing is known. When the timing of change is uncertain, modeling using ARIMA and Joinpoint methods [5, 11] or a dynamic model combining time series modeling with Bayesian methods could be helpful.

Because it is theoretically complex and computationally intensive to build a dynamic model combining the time series and Bayesian components, we approached the modeling in several stages. First, we used ARIMA models to analyze the residuals and forecast errors. Next, we used several ARIMA intervention models to estimate the timing and magnitude of the drop, specifying for each model that the intervention occurred at a different month in the vicinity of late 2008/early 2009. Finally, Bayesian analysis [12–14] was constructed to estimate the timing of the drop and its credible interval. While different Bayesian methods have been used to model the change points in other studies [13, 14], the Bayesian technique developed in this study provides a closed form for the posterior distribution for the timing of the change point.

2. Methods

2.1. Time series

Time series methodology may be used to assess seasonal patterns in the monthly case data. We employed classical seasonal ARIMA models [8] for the 2000–2007 log-transformed data. These take the form

$$\left(1-\sum_{i=1}^{p}\alpha_i B^i\right)\left(1-\sum_{i=1}^{P}\beta_i B^i\right)(1-B)^d\left(1-B^{12}\right)^D y_t=\left(1-\sum_{i=1}^{q}\theta_i B^i\right)\left(1-\sum_{i=1}^{Q}\gamma_i B^i\right)e_t$$

where y_t is the expected number of cases at time t, e_t is the white noise error term at time t, and B is the backward shift operator, $By_t = y_{t-1}$. These models include: autoregressive parameters α_i, β_i (seasonal) with orders p, P; moving average parameters θ_i, γ_i (seasonal) with orders q, Q; differences d, D (seasonal). The seasonal lag is 12 months.

To select the best model for monthly TB case counts with associated orders (p, d, q), (P, D, Q), we calculated the autocorrelation, partial autocorrelation, inverse autocorrelation, and white noise probability functions. We also examined the distribution of residuals, quantile-residual plots, and forecasting values. Goodness of fit was assessed by the Akaike information criterion (AIC) and Schwarz criterion. R^2 was also computed. We selected the best model and used it to forecast the monthly data for 2008 to 2010. Observed data from 2008 were compared with the model forecast for validation. The goal was to establish a model that fit 2000–2007 data and reliably forecast 2008 data. The model was used to predict the monthly case counts for 2009–2010 assuming no change in trend.

To evaluate the model and make statistical inferences from the monthly data, we examined residuals from the 2001–2007 data and differences of observed and expected cases for 2008 to 2010. We used t-tests for the null hypotheses of no differences between 2001–2007 residuals and observed minus expected cases for each individual year 2008, 2009, 2010. We calculated p-values and interpreted $p < 0.05$ to indicate a significant difference.

2.2. Intervention models

Intervention analysis has been used in economics and other disciplines [9, 10]. It can be applied to time series data when a predetermined intervention time m is specified, corresponding to the drop in cases

$$Z_t = \beta S_t + Y_t \quad S_t = \begin{cases} 0 & (t < m) \\ 1 & (t \geqslant m) \end{cases}$$

Here, Y_t is an ARIMA model, β is the magnitude of the drop, and S_t is a step function.

In the monthly TB data, however, the time of the drop in cases was unknown. Therefore, we constructed a series of intervention models, one model for each month from July 2007 to June 2010, with the month corresponding to the intervention time, m. These were used to fit log-transformed monthly TB data. In each instance, we modeled the magnitude of the drop β, representing the sharp decline of TB cases. For all intervention models, identical ARIMA parameters (p, d, q), (P, D, Q) were selected to ensure consistency when comparing the AIC values for model fitting. We retained the model with smallest AIC as the best fit. After identifying the best model and the time of the drop in cases, we calculated the magnitude of the drop and its confidence interval.

In the intervention model, confidence bounds for the time of the drop are inestimable because m is not a statistical parameter. We adopted a Bayesian approach to estimate the time of change and its associated credible interval.

2.3. Bayesian analysis

To estimate the timing of the drop in TB cases and its confidence bounds, a Bayesian approach was developed. In this approach, the residuals of 2001–2007 and the differences of observed and expected data of 2008–2010 in the first ARIMA model were used for analysis. The monthly data of observed minus expected (x_1, x_2, …, x_n) from 2001 to 2010 were assumed to follow two normal distributions: the first distribution before month m, the second after and including month m, where the change point $m(1 < m \leqslant n)$ is a parameter of interest.

$$x_1, x_2, \cdots, x_{m-1} \mid \mu_1, \sigma_1^2 \sim N\left(\mu_1, \sigma_1^2\right)$$

$$x_m, \cdots, x_n \mid \mu_2, \sigma_2^2 \sim N\left(\mu_2, \sigma_2^2\right)$$

To facilitate calculation, we chose the conjugate prior distribution of σ_i^2 as scaled inverse chi-squared and the conditional distribution of μ_i given σ_i^2 as normal [12]

$$\mu_i \mid \sigma_i^2 \sim N\left(\theta_i, \sigma_i^2/\kappa_i\right), \quad \sigma_i^2 \sim \text{Inv} - \chi^2\left(\upsilon_i, \tau_i^2\right), \quad i = 1, 2$$

We assumed a noninformative uniform prior distribution for m : $p(m) \propto 1$.

The corresponding joint prior densities, assumed to be independent of each other, are

$$p\left(\mu_i, \sigma_i^2\right) = p\left(\mu_i \mid \sigma_i^2\right) \cdot p\left(\sigma_i^2\right) \propto \left(\sigma_i^2\right)^{-[(\upsilon_i+1)/2+1]} \cdot \exp\left\{-\left[\upsilon_i\tau_i^2 + \kappa_i(\theta_i - \mu_i)^2\right]/\left(2\sigma_i^2\right)\right\}, \quad i = 1, 2$$

The joint posterior density is given by

$$\begin{aligned} f\left(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n, m, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) &\propto \sigma_1^{-(m-1)} \exp\left\{-\left[\sum_{i=1}^{m-1}(x_i - \mu_1)^2\right]/\left(2\sigma_1^2\right)\right\} \\ &\left(\sigma_1^2\right)^{-[(\upsilon_1+1)/2+1]} \exp\left\{-\left[\upsilon_1\tau_1^2 + \kappa_1(\theta_1 - \mu_1)^2\right]/\left(2\sigma_1^2\right)\right\} \\ &\cdot \sigma_2^{-(n-m+1)} \exp\left\{-\left[\sum_{i=m}^{n}(x_i - \mu_2)^2\right]/\left(2\sigma_2^2\right)\right\}\left(\sigma_2^2\right)^{-[(\upsilon_2+1)/2+1]} \exp\left\{-\left[\upsilon_2\tau_2^2 + \kappa_2(\theta_2 - \mu_2)^2\right]/\left(2\sigma_2^2\right)\right\} \end{aligned}$$

We integrated over the nuisance parameters (μ_1, σ_1^2, μ_2, σ_2^2) to obtain the closed form of the marginal distributions

$$\begin{aligned} f(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n, m) &= \int f\left(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n, m, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) \mathrm{d}\mu_1 \mathrm{d}\sigma_1^2 \mathrm{d}\mu_2 \mathrm{d}\sigma_2^2 \\ &\propto (m - 1 + \kappa_1)^{-1/2} G_1^{-[(\upsilon_1+m-1)/2]} \Gamma[(\upsilon_1 + m - 1)/2] \\ &\cdot (n - m + 1 + \kappa_2)^{-1/2} G_2^{-[(\upsilon_2+n-m+1)/2]} \Gamma[(\upsilon_2 + n - m + 1)/2] \end{aligned}$$

where

$$G_1 = \sum_{i=1}^{m-1} x_i^2 + \upsilon_1\tau_1^2 + \kappa_1\theta_1^2 - \left(\sum_{i=1}^{m-1} x_i + \kappa_1\theta_1\right)^2 /(m-1+\kappa_1)$$

$$G_2 = \sum_{i=m}^{n} x_i^2 + \upsilon_2\tau_2^2 + \kappa_2\theta_2^2 - \left(\sum_{i=m}^{n} x_i + \kappa_2\theta_2\right)^2 /(n-m+1+\kappa_2)$$

Hence, the posterior probability distribution of m is given by

$$f(m \mid x_1, \cdots, x_{m-1}, x_m, \cdots, x_n) = f(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n, m)/f(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n)$$

where

$$f(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n) = \sum_{m=2}^{n} f(x_1, \cdots, x_{m-1}, x_m, \cdots, x_n, m)$$

Using this posterior density, the timing of the drop and its credible intervals can then be obtained.

All the computations were carried out by using the software SAS 9.2 (SAS Institute, NC, USA) and the R 2.13.1 software (www.r-project.org).

3. Results

A seasonally adjusted model for the monthly 2000–2010 TB data is shown in Figure 2. After considering several alternatives, an ARIMA (0, 1, 1) $(0, 1, 1)_{12}$ model produced the best fit to the 2000–2007 log transformed data ($R^2 = 0.96$). Because of the first-order difference of the ARIMA model, the residuals start at 2001 (Figure 3).

To assess the validity of model forecast, the differences of observed data and expected values from the model for 2008 were tested against the residuals of 2001–2007. No significant difference was found by t-test ($p = 0.90$), suggesting that the model was valid for forecasting future data that follow the same trend or pattern of previous years.

When the differences of observed data and expected values for 2009 and 2010 were tested against the residuals of 2001–2007, significant differences were found by t-test ($p = 0.005$ for 2009 and $p = 0.001$ for 2010). These results suggested that 2009 and 2010 data did not follow the same trend of previous years.

A substantial drop in TB cases in late 2008 and early 2009 was observed (Figure 3). To further study the drop, multiple intervention models, one for each month from July 2007 to June 2010, were used to model the log transformed monthly data by assuming each model had an intervention location at different months of the year, with unknown magnitude. The

best fit model for each intervention month was determined by the AIC value. Repeating this process, the ARIMA (0, 1, 1) $(0, 1, 1)_{12}$ models with step intervention functions of different magnitudes at each intervention month were compared by computing their corresponding AIC values (Figure 4). The smallest AIC value in January 2009 indicated that the intervention timing was most likely to be in January 2009, with the next best fitting location of the timing in November 2008. The magnitude of the drop and its 95% confidence interval were estimated by the β coefficient for January 2009 as −11.2% (−14.5%, −7.9%).

To estimate the credible intervals (CI) for the timing of the intervention, a Bayesian technique was developed. The residuals of 2001–2007 and the differences of the observed data and expected values from the ARIMA model in first stage, normalized by a factor of 0.01 (Figure 3), were used to build the Bayesian model. The initial values ($\theta_1 = 0$, $\theta_2 = -1$, $\kappa_i = 1$, $\nu_i = 0.5$, and $\tau_i^2 = 3$; $i = 1, 2$) were used for the prior distributions. Using the posterior probability distribution formula for the timing variable (m), we estimated the posterior probability distribution for the timing of change m (Figure 5). The maximum in the posterior probability distribution occurred in January 2009 (95% CI: 07/2008, 04/2009). We further conducted sensitivity analyses for initial values of prior parameters and found no change for the point estimate of m and minimal changes for its credible interval.

4. Discussion

In this paper, we propose a combination of methods to model time series data to achieve higher resolution for statistical inferences. The three-stage modeling for the time series data in this paper provides a practical approach for statistical inferences about the details of the change or early detection of a change in the trend of time series data.

Beyond seasonal monthly data, time series data in public health may include other operational data (e.g., engineering or financial data) with higher resolution such as weekly or daily counts or in smaller units. Usually, time series data with higher resolutions may contain more random noise or have different patterns and therefore be harder to analyze than annual data. Considering the efficiency needed to produce analytical results or preliminary conclusions for public health data, such as rapid analysis of pandemic influenza trends, building a more general theoretical model than the one that we present might not be feasible or practical because of the complexity involved.

The usefulness of this approach is reflected in the modeling of the TB time series data. First, the time series data can be assessed at different stages to learn the patterns of the data and to ensure the validity of the modeling at each stage. Second, the approach by different stages will reduce the complexity of general modeling and produce interpretable results for each stage. Third, the methods for each stage may stand alone to produce estimates for different data series or for difference purposes.

The contributions of the Bayesian analysis include detecting the change in location and its credible interval for data series from another process and for the original data series. Because a closed form posterior distribution was derived, the effects of sampling distributions and values of prior parameters can be analyzed and interpreted in the process of

updating the posterior distribution. This Bayesian approach can also be used to estimate the change in magnitude. Our Bayesian interval estimates are similar to results obtained from a Joinpoint analysis used in another related paper [5].

In summary, the three-stage approach in this paper combining ARIMA, ARIMA intervention, and Bayesian methods provides a practical and effective way for the estimation and interpretation of operational data in public health or other areas. Future work may involve building a single Bayesian model combining the ARIMA and the Bayesian components that we presented.

Acknowledgements

The authors would like to thank Andrew Hill for his proofreading and suggestions and thank Chad Heilig and Dylan Shepardson for their comments and suggestions. We thank the local and state health department personnel who collected national tuberculosis surveillance data used in these analyses. Routine data in these analyses have been determined to be public health surveillance and not human subjects research requiring oversight by an institutional review board.

The findings and conclusions are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. CDC. Reported Tuberculosis in the United States, 2009. Atlanta, GA: U.S. Department of Health and Human Services, CDC October 2010.
2. Pratt R, Price S, Miramontes R, Navin T, Abraham BK. Trends in tuberculosis –United States, 2010. MMWR 2011; 60:333–337. [PubMed: 21430636]
3. Pratt R, Robison V, Navin TR, Bloss E. Trends in tuberculosis –United States, 2008. MMWR 2009; 58:249–253. [PubMed: 19300406]
4. Winston C, Pratt R, Armstrong L, Navin TR. Decrease in reported tuberculosis cases – United States, 2009. MMWR 2010; 59:289–294. [PubMed: 20300055]
5. Winston CA, Navin TR, Becerra JE, Chen MP, Armstrong LR, Jeffries C, Yelk Woodruff RS, Wing J, Starks AM, Hales CM, Kammerer JS, Mac Kenzie WR, Mitruka K, Miner MC, Price S, Scavotto J, Cronin AM, Griffin P, Lobue PA, Castro KG. Unexpected decline in tuberculosis cases coincident with economic recession – United States, 2009. BMC Public Health 2011; 11:846. DOI: 10.1186/1471-2458-11-846. [PubMed: 22059421]
6. Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. Holden-Day: San Francisco, 1970.
7. Box GEP, Pierce DA. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association 1970; 65:1509–1526.
8. Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting: Methods and Applications. John Wiley & Sons: Hoboken, NJ, 1997.
9. Box GEP, Tiao GC. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 1975; 70:70–79.
10. Box GEP, Tiao GC. Comparison of forecast and actuality. Applied Statistics 1976; 25:195–200.
11. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine 2000; 19:335–351. DOI: 10.1002/(SICI)1097-0258(20000215)19:3<335::AID-SIM336>3.0.CO;2-Z. [PubMed: 10649300]
12. Gelman A, Carlin JB, Stern HS, Rubin D. Bayesian Data Analysis. Chapman and Hall/CRC: Boca Raton, FL, 2003.
13. Akman VE, Raftery AE. Asymptotic inference for a change-point Poisson process. Annals of Statistics 1986; 14:1583–1590.

14. Ghosh P, Ghoshb K, Tiwaric RC. Bayesian approach to cancer-trend analysis using age-stratified Poisson regression models. Statistics in Medicine 2011; 30:127–139. DOI: 10.1002/sim.4077. [PubMed: 20839366]

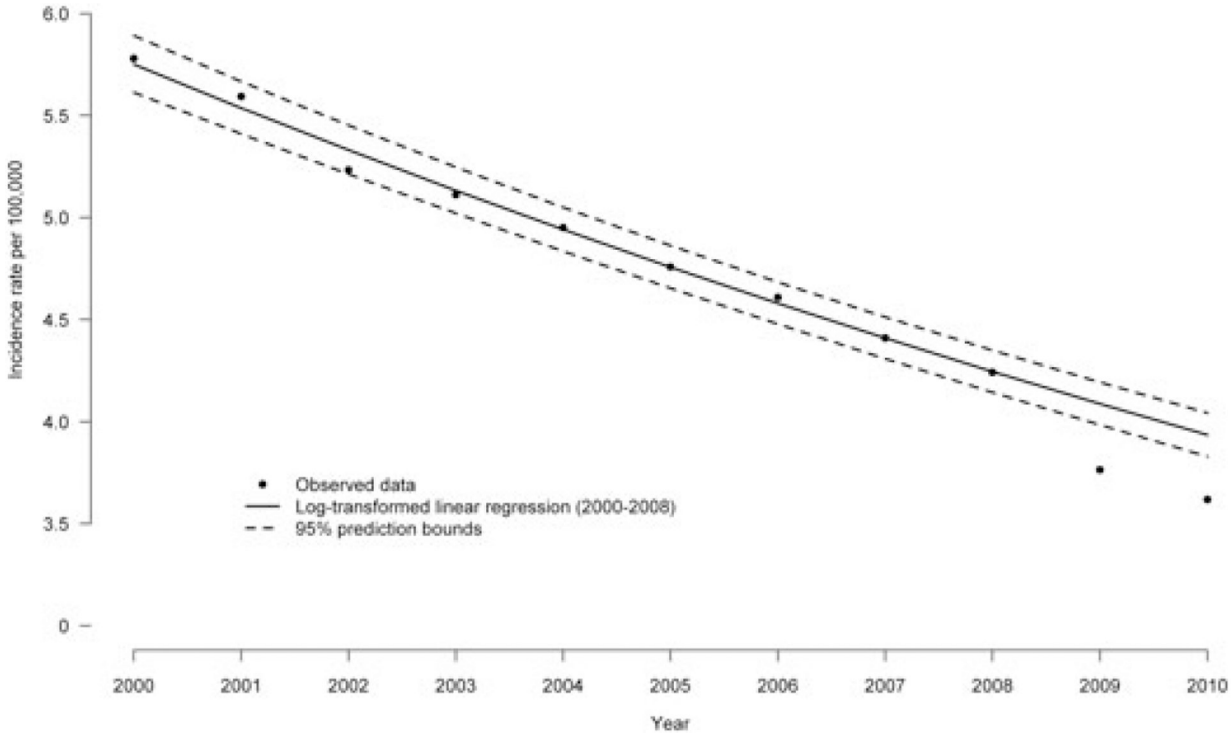

Figure 1.
Annual tuberculosis incidence rates in the United States, 2000–2010.

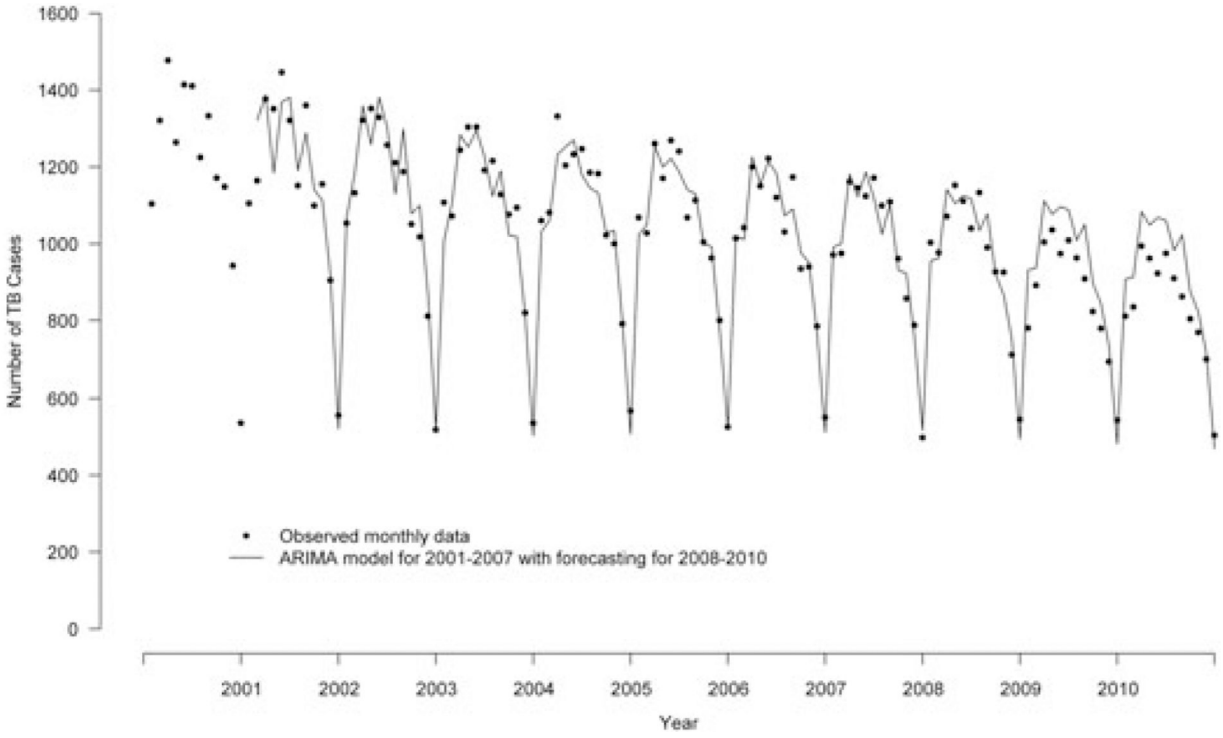

Figure 2.
Monthly TB cases in the United States and ARIMA model, 2000–2010.

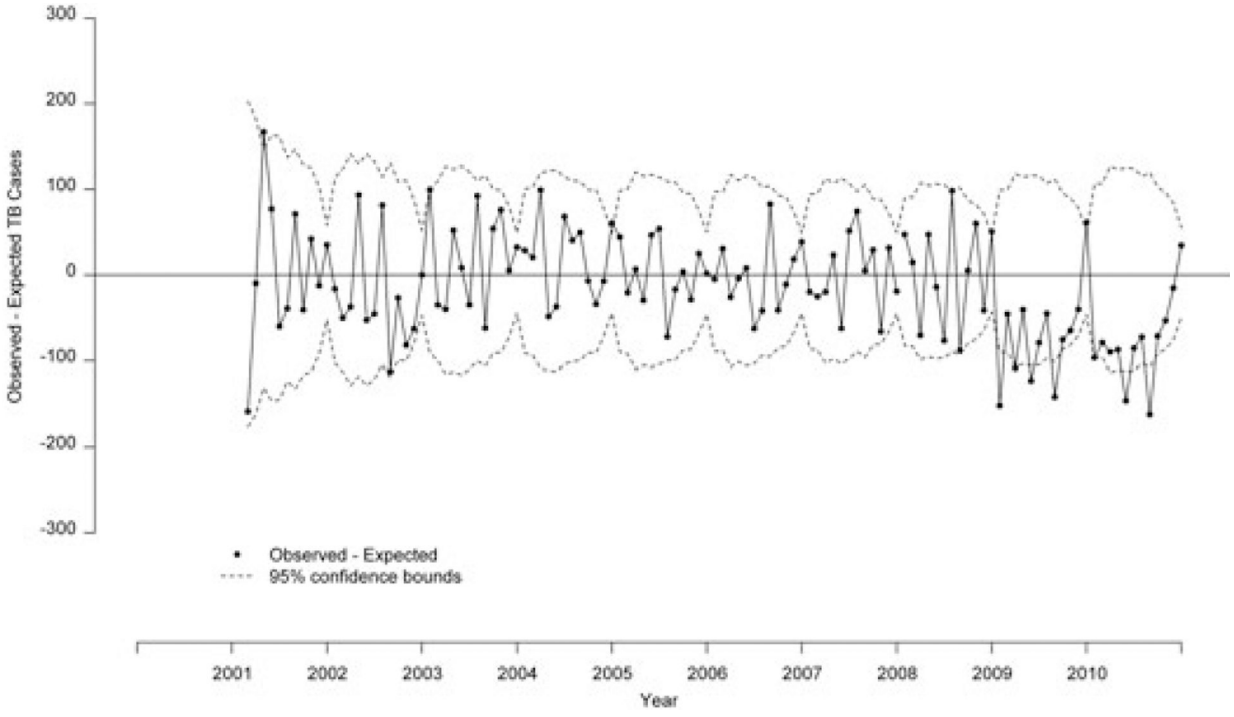

Figure 3.
Differences of observed and expected TB cases from ARIMA model.

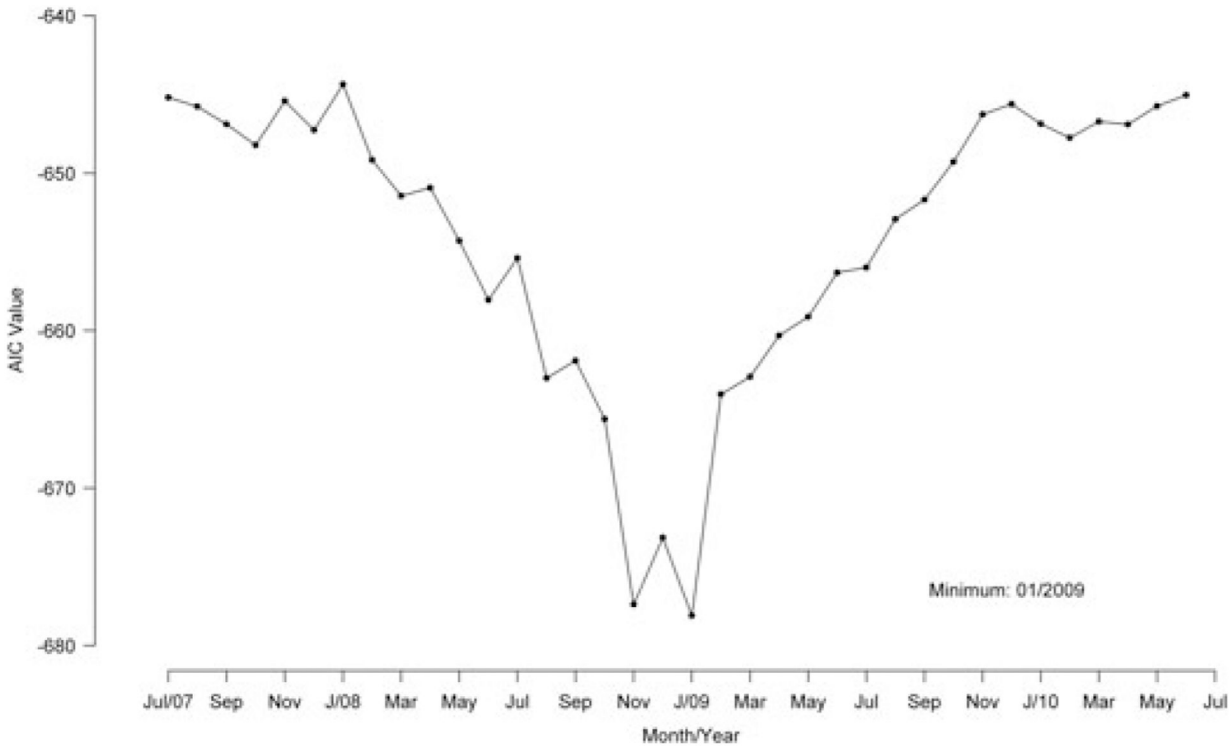

Figure 4.
AIC values of intervention models by timing of intervention.

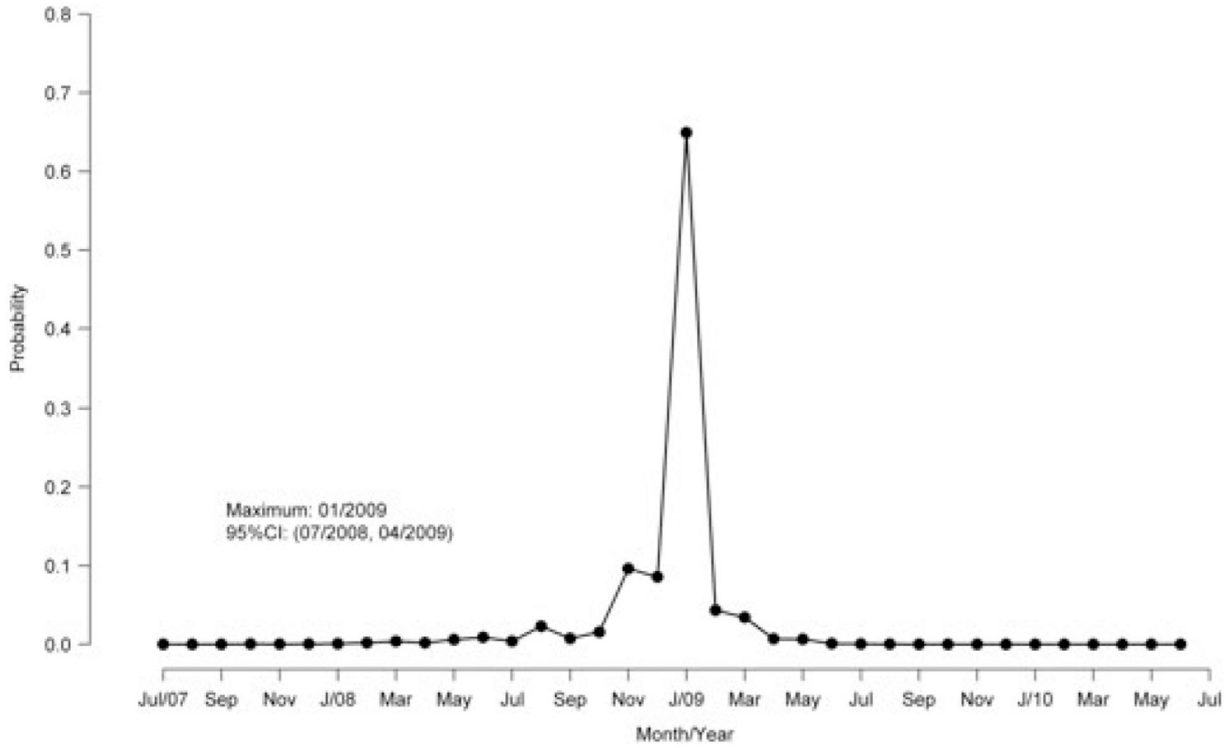

Figure 5.
Posterior probability distribution for timing of change.