Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Denis Whelan Date

AN ADDITIVE SPATIOTEMPORAL COVARIANCE FUNCTION USING STREAM AND RIVER DISTANCE

By

Denis Whelan

Master of Science

Biostatistics

_____ [Advisor's signature]

Howard Chang, Ph.D.

Advisor

_____ [Member's signature]

Lance Waller, Ph.D.

Committee Member

_____ [Member's signature]

Zhaohui Qin, Ph.D.

Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

AN ADDITIVE SPATIOTEMPORAL COVARIANCE FUNCTION USING STREAM AND RIVER DISTANCE

By

Denis Whelan

B.S., Villanova University, 2014

Advisor: Howard Chang, Ph.D.

An abstract of

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of
Master Science
in Biostatistics

2017

Abstract

AN ADDITIVE SPATIOTEMPORAL COVARIANCE FUNCTION USING STREAM AND RIVER DISTANCE

By Denis Whelan

BACKGROUND: Modeling all geo-referenced phenomenon using strictly Euclidean distance is restrictive and often implausible, especially when measuring the movement of organisms in infectious disease research and ecology. Studying the transmission of diarrheal disease incidence in developing countries involves studying the movement of waterborne pathogens. In order to effectively estimate the spread of diarrheal disease, spatiotemporal heterogeneity must be considered explicitly.
OBJECTIVE: This paper presents a covariance function that incorporates a weighted combination of multiple distance metrics to estimate spatiotemporal dependence of a Gaussian outcome using a Bayesian framework.
DATA: Twenty-one communities in northwestern Ecuador were randomly selected from data collected as a part of the ECODESS study (Ecología, Desarollo, Salud y Sociedad), which was geared towards achieving a better understanding of community-level risk factors of diarrheal disease by examining environmental and ecological factors.
METHODS: An additive covariance function incorporating both Euclidean and river distance is proposed to estimate spatiotemporal dependence. Metropolis-Hastings and Gibbs sampling are used for MCMC parameter estimation. We use three fit measures designed for Bayesian models to compare the model fit of the additive covariance function to a model with Euclidean distance only and a model with river distance only.
RESULTS: The additive covariance function incorporating both Euclidean and river distance was the best performing model in terms of all three Bayesian model fit criteria only when the simulated Gaussian was generated using a combination of those simulated distance matrices. Results were mixed when this method was applied to observed data.
CONCLUSIONS: This paper lays the foundation for estimation of covariance functions using multiple distance matrices with wide applications in infectious disease and ecology research and can motivate a range important methodological extensions.

AN ADDITIVE SPATIOTEMPORAL COVARIANCE FUNCTION USING STREAM AND RIVER DISTANCE

By

Denis Whelan

B.S., Villanova University, 2014

Advisor: Howard Chang, Ph.D.

A thesis submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science
in Biostatistics

2017

Acknowledgments:

To Howard, my advisor, for his patience and his guidance, to my committee, for their helpful suggestions, to my classmates for all that I have learned through them and alongside them.

Table of Contents:

INTRODUCTION

Modeling all geo-referenced phenomenon using strictly Euclidean distance is restrictive and often implausible. In infectious disease research and ecology, many species of interest or waterborne pathogens may travel only along a stream network. In operations research or military applications, stakeholders may be interested in modeling the movement of goods, people, or illicit drugs along transportation networks such as roads. Especially in rural, mountainous, or otherwise restricted-movement areas, using a spatial autocovariance function that only depends on Euclidean (as the crow flies) distance on a projected map could lead to very unreliable estimates.

For situations such as these, stream distance and road distance may be more appropriate metrics for modeling autocovariance. Stream or road distance is defined as the shortest distance between two locations, where distance is calculated only within the stream or road network *(Ver Hoef et al. 2006*). However, stream or road distance alone may not be enough to accurately characterize spatial autocovariance, given that it would be unrealistic to assume that these processes are isotropic. For example, stream and transportation networks may have correlations defined by direction and volume of flow. *Ver Hoef et al.* (2006) have used moving average constructions to develop valid spatial autocovariance models using stream distance and incorporating directional flow.

For modeling the spread of infectious diseases transmitted by both humans and waterborne pathogens, it is important to not only consider transmission along streams but to simultaneously consider multiple routes of transmission. In this paper, we propose a novel method to incorporate combinations of Euclidean, river, and road distances to identify the most appropriate spatial autocorrelation model. River distances have been used for geostatistics in

ecology (*Money et al. 2009*, *Peterson et al. 2010, 2007, 2006*, *Cressie et al. 2006*), however they have not yet been incorporated into infectious disease modeling. Despite their practical relevance, road networks have very rarely (*Saby et al. 2006*) been used in geospatial models, and, to our knowledge, combinations of distance matrices have never been used together to model spatial autocorrelation, in public health settings, or otherwise.

In addition to the spatial challenges of modeling the transmission of waterborne infectious diseases, varying climatological patterns present additional temporal complications. To account for this, *Reich et al.* (2011) proposed a class of spatiotemporal covariance functions that allow for meteorological covariates to affect a covariance function. Based on the work of [*Schmidt, Guttorp et al. (2011)* and *Schmidt, Rodríguez et al. (2011)*], Reich et al. developed a framework for nonstationary covariance, such that the correlation between pairs of points separated by the same distance may have different correlations depending on precipitation or other meteorological factors.

In this paper, we present a covariance function that incorporates multiple distance metrics to estimate spatiotemporal dependence of a Gaussian outcome. We first use simulated spatiotemporal data to evaluate the method's performance based on different model fit criteria for Bayesian models. We then apply this model to estimate spatiotemporal dependence of village-level incidence rates of diarrheal disease in the Esmeraldes region on the northwestern coast of Ecuador.

Diarrhea is the second leading cause of death in children under five, with 1.7 billion cases of childhood diarrheal disease each year (*WHO*). Worldwide, nearly 1.9 million children die from diarrheal disease annually, which is 19% of all deaths in this age group (*WHO*). This global burden of diarrheal disease disproportionately affects developing regions like northwestern

Ecuador, largely because of the lack of safe water and basic sanitation (*Boschi-Pinto et al., 2008, Kotloff et al., 2013*).

It has been shown (*Patz et. al, 2005, Jones et. al 2008, Lipp et al. 2001*) that pathogens are influenced by rainfall and temperature. *Checkley et al.* (2000) found that diarrhea increased by 8% for each one-degree Celsius increase in mean ambient temperature in Peru during the period before El Niño. In Canada, *Thomas et al.* (2006) found that waterborne outbreaks (1975 – 2001) were significantly associated with total maximum degree-day above 0^0C and accumulated rainfall. In the U.S., waterborne diseases outbreaks (1948-1994) due to surface water contamination showed a strong association with extreme precipitation *(Curriero et al., 2001)*.

However, not all results have been consistent. In Bangladesh, the number of cholera cases increased with both high and low rainfall in the weeks preceding hospital visits (*Hashizume et al., 2008)*. In Botswana, *Alexander et al.* (2013) found that forecasted increases in temperature and decreases in precipitation may lead to a prolongation of the present dry season peak and a subsequent increase in diarrheal disease. However, the same study predicted that incidence of diarrheal disease in the wet season would decline.

Research focused on modeling associations between rainfall and waterborne disease incidence (*Singh et al 2001, Peng et al. 1999, Auld et. al 2004, Gordon et al. 2009,* and many others) does not account for spatiotemporal heterogeneity of the effect of rainfall on diarrheal disease. Though this paper's focus is on the heterogeneity of the covariance, better understanding of the heterogeneity of the covariance is an important step to an enhanced understanding of the spatial heterogeneity of the effect of rainfall on the disease. Ignoring this crucial characteristic of the data may limit investigators' ability to detect and estimate accurate associations between precipitation and disease, predict patterns of disease based on precipitation, and interpret the

relative importance of other community-level risk factors. The covariance function proposed in this paper should allow us to better examine what geographical factors (i.e. distance measures) and environmental factors (i.e. precipitation) will influence the spatiotemporal dependence between relative risks of diarrheal disease between villages.

DATA

Data for this analysis were collected as a part of the ECODESS study (Ecologia, Desarrollo, Salud, y Sociedad), which was designed to contribute to a better understanding of the community-level risk factors of diarrheal disease by examining how environmental and ecological factors such as social networks and sanitation impact transmission (*Eisenberg et al., 2006*). Initial findings and study design have previously been described in detail (*Eisenberg et al., 2006*) and further discussion is included elsewhere (*Ahn et al., 2014*).

Twenty-one communities were randomly selected from the 158 communities in the Esmeraldas province of northwestern Ecuador based on a randomized-block design which used location, size, population and relative distance to Borbón, the main population center of the region (Figure 1a). (Actual village names have been replaced by numbers for confidentiality throughout.) This region forms a part of the Chocó rainforest, a tropical rainforest that cuts through parts of Panamá, Colombia, and Ecuador, and known to be one of the most biologically-diverse regions in the world. These communities are found along the three river system in the region (Río Onzole, Río Cayapas, and Río Santiago), with 18 villages located beside one of the three major rivers (Figure 1b). Due to the recent construction of roads connecting villages fostering increased movement between communities, this region is an ideal place to study

community-level risk factors of disease. (Ten of the 21 villages are now connected by the new road system.)

Research team members recruited and surveyed households from February 2004 to July 2007 for 177 consecutive weeks to identify diarrhea cases, defined as three or more loose stools in a 24-hour period. Other community-level factors such as remoteness, social cohesion, travel patterns, and sanitation were collected, but are not yet available for analysis at this time.

METHODS

Model Formulation

In order to estimate spatiotemporal dependence of village-level incidence rates of diarrheal disease, we propose an additive covariance function that incorporates stream or road distance and Euclidean distance. First, let's consider the following spatial-temporal regression model:

$$\boldsymbol{Y}(s_i, t_i) = \boldsymbol{W}(s_i, t_i) + \varepsilon_i$$

where s_i and t_i denote the spatial location and week for observation i. The independent mean-zero Gaussian residual error term ε_i is assumed to be *iid* with mean zero and variance σ^2. At each time point t_i, $\boldsymbol{W}(s_i, t_i)$ is a mean-zero Gaussian process with stationary and isotropic covariance function $\sum$. Hence, the joint distribution $\mathbf{Y}= [Y(s_1)\ldots Y(s_n)]'$ is Gaussian with mean zero and variance $\sum + \sigma^2 I_{n\times n}$ For the application of our model to the Ecuador data, we let $Y(s_{ij},t_{ij})$ to represent the log of proportion of of the local population at risk that are incident cases per week in village s during week t. The element $\sum_{ij}$ is determined by a parametric covariance function Cov (d_{1ij}, d_{2ij}; θ, τ^2, ρ_1, ρ_2) where d_{1ij} is the Euclidean distance between s_i and s_j and d_{2ij} is the stream or road distance between s_i and s_j:

$$\boldsymbol{S} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\theta * \exp\left(\frac{-d_{1ij}}{\rho_1}\right) + (1-\theta) * \exp\left(\frac{-d_{2ij}}{\rho_2}\right)]$$

When only one distance metric is used, θ is assumed to be zero or one *a priori*.

Parameter estimation is accomplished under a Bayesian Framework. We further specify the above regression model as a Bayesian model with parameters $\Omega = \{\theta, \sigma^2, \beta, \boldsymbol{W}, \tau^2, \rho_1, \rho_2\}$, allowing hierarchical specification of model elements. For Bayesian inference, each element in Ω is treated as a random variable and requires a prior distribution. The joint posterior distribution is given by

$$\pi(\Omega|\mathrm{Y}) = \frac{L(Y|\Omega) * \pi(\Omega)}{f(Y)}$$

Where $L(\mathbf{Y}|\Omega)$ is the data likelihood, which describes a probabilistic description of how the data arise, $\pi(\Omega)$, which describe prior information in unknown parameters, and $f(Y)$, a normalizing constant ensuring a proper posterior distribution.

The data likelihood and prior distributions for the above model are given below

$$L(\boldsymbol{Y}|\theta, \sigma^2, \boldsymbol{W}, \tau^2, \rho_1, \rho_2) \sim N(\boldsymbol{W}, \sigma^2\boldsymbol{I})$$

$$\pi(\theta, \sigma^2, \boldsymbol{W}, \tau^2, \rho_1, \rho_2) = \pi(\boldsymbol{W}|\tau^2, \rho_1, \rho_2,) * \pi(\theta) * \pi(\sigma^2)$$

$$\pi(\boldsymbol{W}|\tau^2, \rho_1, \rho_2, \theta) \sim N\left(0, \sum(\tau^2, \rho_1, \rho_2, \theta)\right)$$

$$\pi(\sigma^2) \sim InvGamma(a, b)$$

$$\pi(\tau^2) \sim InvGamma(a, b)$$

$$\pi(\rho_1) \sim Gamma(\mathrm{c}, d)$$

$$\pi(\rho_2) \sim Gamma(\mathrm{c}, d)$$

$$logit(\theta) = \log(\frac{\theta}{1-\theta}) \sim 1$$

where hyper-parameters are given by the values a=0.0001, b=0.0001, c=0.6, and d=0.1.

Estimation

In order to simulate samples from the joint posterior distribution $\pi(\theta, \sigma^2, \beta, \boldsymbol{W}, \tau^2, \rho_1, \rho_2)$, we use Markov Chain Monte Carlo (MCMC). MCMC consists of Monte Carlo integration using Markov chains. Monte Carlo integration draws samples from the specified posterior distribution and uses sample averages to approximate expectations. Estimation uncertainty as measured by posterior standard deviation and posterior intervals can also be calculated from quantiles of the posterior samples. MCMC draws these samples by updating model parameters via Markov chains and has the useful property of gradually "forgetting" its initial state—subject to certain regularity conditions—and eventually converging to a set of samples from a unique stationary distribution that is equivalent to the joint posterior distribution of interest. Thus after "burning in" over sufficient iterations, subsequent iterations will represent dependent samples from the posterior distribution based on the ergodic theorem (*Eckman & Ruelle, 1985*). For our models, we run 5,000 iterations and discard the first 1,000 as burn-in samples. We use a combination of Gibbs sampling and the Metropolis-Hastings algorithm to define our Markov chain samples.

Gibbs-Sampling Steps

The first MCMC technique we use is Gibbs sampling. Gibbs sampling generates posterior samples by sampling from the full conditional distribution of each parameter with all other parameters remaining fixed at their current values, resulting in a Markov chain whose stationary

distribution is the posterior distribution. Thus, instead of sampling directly from the posterior distribution, samples are simulated by passing through all posterior conditionals and updating each parameter, one at a time. The following parameters in are model are sampled using Gibbs samplers $\boldsymbol{W}$, τ^2, and σ^2.

Metropolis-Hastings Algorithm

For certain parameters, however, it is not possible to specify the full conditional distribution in a convenient closed form and we must use another technique called Metropolis-Hastings. The Metropolis-Hastings algorithm is an approach to simulate a distribution where we only know the density up to a proportional constant and can be used when the full conditional distribution is not available. At each iteration *k*, for a parameter ω_2:

1.) A proposal value ω_2^* is drawn from the distribution $\pi(\omega_2^* | \omega_2^{(k-1)})$

2.) ω_2^* is accepted as $\omega_2^{(k)}$ with probability:

$$\min\left\{1, \frac{\pi(\boldsymbol{W}|\varpi_1, \varpi_2^*) * \pi(\varpi_2^*) * \pi\left(\varpi_2^{(k-1)}\middle|\varpi_2^*\right)}{\pi\left(\boldsymbol{W}\middle|\varpi_1^k, \varpi_2^{k-1}\right) * \pi\left(\varpi_2^{k-1}\right) * \pi(\varpi_2^*|\varpi_2^{(k-1)})}\right\}$$

Otherwise, if ω_2^* is not accepted, then $\omega_2^{(k)} = \omega_2^{(k-1)}$. For our method, for each $\omega \in \Omega$ we must take care to ensure to propose a value in the support of $\pi(\omega)$.

Details of the MCMC algorithm for the proposed model is given below.

1.) Define initial values for the unknown parameters: σ^2, τ^2, ρ_1, and ρ_2. Use those initial values to initialize **R**, **S**, and $\boldsymbol{W}$.

$$\boldsymbol{R} = \theta * \exp\left(\frac{-d_{1ij}}{\rho_1}\right) + (1-\theta) * \exp\left(\frac{-d_{2ij}}{\rho_2}\right)$$

2.) Update $\boldsymbol{W}$:

$$\pi(\boldsymbol{W} \mid \theta, \sigma^2, \tau^2, \rho_1, \rho_2) \ \alpha \ L(\boldsymbol{Y}|\boldsymbol{W}, \sigma^2) * \pi(\boldsymbol{W}|\tau^2, \rho_1, \rho_2, \theta)$$

$$\pi(\boldsymbol{W}|\tau^2, \rho_1, \rho_2, \theta) \sim N(\widehat{\boldsymbol{\mu}}, \widehat{\sum})$$

$$\boldsymbol{VVV} = \frac{1}{\sigma^2}\boldsymbol{I}_{n*t\ x\ n*t} + \frac{1}{\tau^2}R^{-1}$$

$$\boldsymbol{XXX} = \boldsymbol{I}_{n*t\ x\ n*t}\,\boldsymbol{Y}$$

$$\widehat{\boldsymbol{\mu}} = \frac{1}{\sigma^2}VVV^{-1}XXX$$

$$\widehat{\sum} = VVV^{-1}$$

3.) Update σ^2:

$$\pi(\sigma^2 \mid \theta, \boldsymbol{W}, \tau^2, \rho_1, \rho_2) \ \alpha \ L(\boldsymbol{Y}|\theta, \sigma^2, \boldsymbol{W}) * \pi(\sigma^2)$$

$$\pi(\sigma^2 \mid \theta, \boldsymbol{W}, \tau^2, \rho_1, \rho_2) \sim InvGamma\left(a + \frac{n*t}{2}, b + \frac{(\boldsymbol{Y}-\boldsymbol{W})'(\boldsymbol{Y}-\boldsymbol{W})}{2}\right)$$

4.) Update τ^2:

$$\pi(\tau^2 \mid \theta, \boldsymbol{W}, \rho_1, \rho_2)\ \alpha\ \pi(\boldsymbol{W} \mid \tau^2, \rho_1, \rho_2)\pi(\tau^2)$$

$$\pi(\tau^2 \mid \theta, \boldsymbol{W}, \tau^2, \rho_1, \rho_2) \sim InvGamma\left(a + \frac{n*t}{2}, b + \frac{\boldsymbol{W'R\,W}}{2}\right)$$

5.) Update ρ_1:

a.) Propose ρ_1*| ρ_1 where ρ_1*~ log-Normal(log(ρ_1), ρ_{1tune})

b.) Calculate **R***= ɵ $*$ exp $\left(\frac{-d_{1ij}}{\rho_{1^*}}\right)$ + (1 − ɵ) $*$ exp $\left(\frac{-d_{2ij}}{\rho_2}\right)$

c.) Decide whether to accept proposed value ρ_1* as $\rho_1^{(k)}$ with probability α.

$$\alpha = \frac{L(\boldsymbol{W}|\, \rho_1^*, \boldsymbol{R}^*, \theta, \boldsymbol{W}, \tau^2, \rho_2) * \pi(\rho_1|\, \rho_1^*, \theta, \boldsymbol{W}, \tau^2, \rho_2)}{L(\boldsymbol{W}|\, \rho_1, \boldsymbol{R}, \theta, \boldsymbol{W}, \tau^2, \rho_2) * \pi(\rho_1^*|\rho_1\, , \theta, \boldsymbol{W}, \tau^2, \rho_2)}$$

6.) Update ρ_2:

a.) Propose ρ_2*| ρ_2 where ρ_2*~ log-Normal(log(ρ_2), ρ_{2tune})

b.) Calculate **R***= ɵ $*$ exp $\left(\frac{-d_{1ij}}{\rho_1}\right) + (1 - ɵ) * \exp\left(\frac{-d_{2ij}}{\rho_2^*}\right)$

c.) Decide whether to accept proposed value ρ_2* as $\rho_2^{(k)}$ with probability α.

$$\alpha = \frac{L(\boldsymbol{W}|\, \rho_2^*, \theta, \boldsymbol{R}^*, \boldsymbol{W}, \tau^2, \rho_1) * \pi(\rho_2|\, \rho_2^*, \theta, \boldsymbol{W}, \tau^2, \rho_1)}{L(\boldsymbol{W}|\, \rho_2, \theta, \boldsymbol{R}\, \boldsymbol{W}, \tau^2, \rho_1) * \pi(\rho_2^*|\rho_2\, , \theta, \boldsymbol{W}, \tau^2, \rho_1)}$$

7.) Update θ by reparametrizing it as $\gamma = \text{logit}\,(\theta)$:

a.) Propose γ*| γ where γ *~ N (γ, γ_{tune}) and $\gamma^* = \text{logit}\,(\theta) = \log(\frac{\theta}{1-\theta})$

b.) Back-transform ɵ* $= \frac{\exp(\gamma*)}{1+ \exp(\gamma*)}$

c.) Calculate **R***= ɵ* $*$ exp $\left(\frac{-d_{1ij}}{\rho_1}\right) + (1 - ɵ^*) * \exp\left(\frac{-d_{2ij}}{\rho_2}\right)$

d.) Decide whether to accept proposed value θ* as $\theta^{(k)}$ with probability α.

$$\alpha = \frac{L(\boldsymbol{W}|\, \gamma^*, \theta^*, \boldsymbol{R}^*, \tau^2, \rho_1, \rho_2) * \pi(\gamma|\, \gamma^*, \theta^*, \boldsymbol{W}, \tau^2, \rho_1, \rho_2)}{L(\boldsymbol{W}|\, \gamma, \theta, \boldsymbol{R}, \tau^2, \rho_1, \rho_2) * \pi(\gamma^*|\gamma\, , \theta, \boldsymbol{W}, \tau^2, \rho_1, \rho_2)}$$

Model Comparison

To evaluate the use of different distance metrics in the covariance function, we use three model fit measures for Bayesian models: DIC, wAIC, and PPL. All three metrics can be thought of as the sum of a goodness of fit term and a penalty term.

Deviance Information Criterion (DIC) (*Spiegelhalter et al. 2002*) is essentially a Bayesian version of Akaike Information Criterion (AIC) with two slight, but important modifications: first,

the maximum likelihood estimate is replaced with the posterior mean $\hat{\theta}_{\text{Bayes}}$ and *k*, the effective number of parameters, is replaced with a bias correction based on the data:

$$DIC = -2 * \log[P(y|\hat{\theta}_{Bayes})] + 2 * p_{DIC}$$

$$p_{DIC} = 2 * \{\log(P(y|\hat{\theta}_{Bayes}) - \frac{1}{S}\sum_{s=1}^{S} \log(P(y_i|\theta_s)\ )\}$$

The Watanabe-Akaike information criterion (wAIC) (Watanabe 2010) is considered a more Bayesian approach. It consists of the difference between the log pointwise predictive density and the posterior variance of the log predictive density for each data point y_i.

$$wAIC = \sum_{i=1}^{n} \log(\frac{1}{S}\sum_{s=1}^{S} P(y_i|\theta_s)\ ) - p_{wAIC}$$

$$p_{wAIC} = 2 * \sum_{i=1}^{n} Var_{post}\{\log(P(y_i|\theta)\}$$

Finally, we compute the Posterior Predictive Loss (PPL) (*Gelfand et al. 1998*), which is calculated by first minimizing the posterior loss for a given model, and then, for models being considered, choosing the one that minimizes this criterion. Below it is expressed as the sum of its goodness of fit term *G* and its penalty term *P*:

$$PPL = G + P$$

$$G = \left|\left|(y - E[Y_{rep}|y])\right|\right|^2$$

$$P = tr(Var(Y_{rep\backslash y}))$$

SIMULATION

Formulation

Simulated Euclidean and river distance matrices were created to test our method. Given that the actual distances from the Esmeraldes data ranged from 0 to 90 km for river distances, and 0 to 47 km for Euclidean distances, simulated distance matrices were created from a uniform distribution to somewhat closely mimic these values, with a range 0 to 100 km. To account for the windiness and turns of river/road distances compared to Euclidean distances, non-Euclidean distances were given some additional noise. This was performed using a log normal distribution to reduce correlation between the two distance matrices. Actual data for the Esmeraldes analysis consisted of 18 spatial points *s* (villages) and 177 time points *t* (weeks). In order to examine the performance of our method with varying availability of spatial and temporal data, we tested a range of possible combinations varying *s* from 18 to 500 and *t* from 1 to 200. Because the actual data are more temporal than spatial, here we focus on results that more closely resemble small *s* and high *t*. Furthermore, we tested non-Euclidean distance matrices with a varying percentage of connected points to observe how the methods would perform when not all points were located on a river or road. We do not present much results of these explorations, however, as they are beyond the scope of these analyses. It should be noted, nonetheless, that as expected, convergence was better with increased spatial data and a higher percentage of points located on river.

Simulated Distance Matrices: Results

Parameters were estimated using MCMC with three covariance functions described below. $\mathbf{S_1}$ is the "correct" combined covariance function which consists of a weighted combination of river

distance and Euclidean distance, and which was used to simulate the outcome. $\mathbf{S_2}$ represents the incorrect or incomplete covariance function as it only considers Euclidean distance and $\mathbf{S_3}$ is similar with river distance only instead.

$$\boldsymbol{S_1} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\text{ɵ} * \exp\left(\frac{-d_{1ij}}{\rho_1}\right) + (1 - \text{ɵ}) * \exp\left(\frac{-d_{2ij}}{\rho_2}\right)].$$

$$\boldsymbol{S_2} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\exp\left(\frac{-d_{1ij}}{\rho_1}\right)].$$

$$\boldsymbol{S_3} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\exp\left(\frac{-d_{2ij}}{\rho_2}\right)].$$

Because the data were simulated with $\mathbf{S_1}$, the model fit with this covariance function should be the best model, and thus minimize all three Bayesian model fit measures: DIC, wAIC, and PPL and should converge slightly better than the other two. In Table 1, we notice that for the two selected combinations of spatial and temporal points (s=20, t=75 and s=200, t=10), the combined model performs best for all three fit measures in the both scenarios suggesting that model performance is as expected for data that are both more temporal than spatial and *vice versa*.

In the corresponding figures (Figures 2a & 2b), we notice that parameters θ, ρ_1, ρ_2, and τ^2 converge fairly well to the pre-specified true value for the "correct" additive covariance function ($\mathbf{S_1}$). Though the residual variance (independent of space) σ^2 has more difficulty converging, we note that the total variability ($\tau^2 + \sigma^2$) is estimated well. As expected, for the Euclidean only covariance function (Figure 2c), τ^2, ρ_1, and σ^2 more or less converge to the pre-specified true value but have more difficulty than in the additive model. We observe similar results for the river distance only model (Figure 2d).

Esmeraldes Data Distance Matrices: Results

After performing MCMC using the above covariance matrices with simulated distance matrices and a simulated outcome ***Y***, the same procedure was performed using a simulated outcome and the observed river distance and Euclidean distance matrices. These analyses are concerned with the 18 villages located along one of the three rivers, and thus without any missing pairwise river distance data (Figure 1b). When using the Esmeraldes data distance matrices (Table 1), we notice that the river only $\mathbf{S_3}$ is the best-performing model when the three covariance functions are compared for model fit, even though the $\mathbf{S_1}$ matrix was used to generate the outcome. In the combined model (Figures 3a & 3b), the posterior distributions for all parameters converge well to the pre-specified true mean with the exception of θ, which is overestimated.

ANALYSIS

Pre-processing

For the 18 villages over 176 weeks, 161 incidence rates were missing (5% of total data), after excluding one week without collected data which was removed. The remaining missing values were imputed in the following way. If an incidence rate was missing for week *t*, but weeks *t-1* and *t+1* were not missing, the incidence rate for week *t* was set to the mean of the incidence rates for weeks *t-1* and *t+1*. If an incidence rate was missing for week *t* and week *t+1…t+n*, the incidence rate for week *t-1* was iteratively set to the incidence rate to incidence rate *t* until the first scenario was achieved and the mean was imputed for the last value. The first and last weeks did not contain any missing values. By imputing the missing weeks chronologically based on information from prior weeks, or averaged information from prior and subsequent weeks, this

method provides a simple way of reasonably estimating incidence rates for weeks with missing cases or population denominators.

Data Description

Incidence rates of diarrheal disease were calculated as follows:

$$Incidence\ Rate = \ 1000 * \frac{\#\ of\ cases}{\#\ surveyed}$$

Thus, denominators could not be missing, and when denominators were small or relatively close to the number of cases, the incidence rate became extremely large. These outliers are noteworthy in Figure 5, which displays the mean number of incident cases for each week during the study period. One can see that weeks 77, 78, and 79 have extraordinarily high mean weekly incidence rates. This actually does not reflect a large increase in the *number* of cases per week but merely the number surveyed on a given week, as there was a great deal of variation between number of persons surveyed per week in a particular village. The standard deviation of the denominator ranged from 6.5 to 111.8 across the 18 villages (median=28.0) whereas the standard deviation of the numerator ranged from 0.36 to 1.86 across the 18 villages (median=0.90).

Results

When fitting the models to the real data, a range of starting values first had to be used in order to achieve convergence. In figures 4a-4d, we observe estimates for the posterior means. We notice that τ^2, the temporal variance, is about 22 and σ^2 is about 25, providing evidence to suggest that the incidence of diarrheal disease in Esmeraldes may vary slightly more spatially than temporally. Model performance was best for lower *t* and became unstable for a large number (t > 40) of weeks at a time, thus results are displayed for 30 weeks at a time. As Table 1 shows, there

is no clear pattern as to which of the three covariance functions performs best for the estimation of the incidence rate of diarrheal disease using the inter-village data in the Esmeraldes region of Ecuador.

DISCUSSION

This paper develops a framework for a valid covariance function that incorporates multiple distance metrics to estimate spatiotemporal dependence of a Gaussian outcome. This work lays the foundation for a variety of useful extensions that may be applied to statistical research in ecology, infectious diseases, and other applications.

In these analyses, we observed that the additive covariance function incorporating both Euclidean and river or road distance was the best performing model in terms of all three Bayesian model fit criteria only when the simulated Gaussian covariance was generated using both those simulated distance matrices (Table 1). Nonetheless, we believe it remains likely that using river distance or Euclidean distance alone is insufficient to estimate spatiotemporal covariance for data located along river networks. There are a variety of additional strategies to build different valid covariance functions that are motivated by this same interplay between estimating Euclidean-driven correlation, and another distance metric. The first may be to consider the model presented here with different prior distributions or hyper-parameter specifications. One may also consider different covariance functions, such as a model that estimates two different τ^2 parameters instead of the weight parameter θ (see $\mathbf{S_{1a}}$), a multiplicative model instead of an additive model ($\mathbf{S_{1b}}$), a common ρ ($\mathbf{S_{1c}}$) or other variations ($\mathbf{S_{1d}}$ & $\mathbf{S_{1d}}$).

$$\boldsymbol{S_{1a}} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau_1^2 * \exp\left(\frac{-d_{1ij}}{\rho_1}\right) + \tau_1^2 * \exp\left(\frac{-d_{2ij}}{\rho_2}\right)$$

$$\boldsymbol{S_{1b}} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\exp\left(\frac{-d_{1ij}}{\rho_1} + \frac{-d_{2ij}}{\rho_2}\right)]$$

$$\boldsymbol{S_{1c}} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\theta * \exp\left(\frac{-d_{1ij}}{\rho}\right) + (1-\theta) * \exp\left(\frac{-d_{2ij}}{\rho}\right)]$$

$$\boldsymbol{S_{1d}} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2 * \exp\left(\frac{-d_{1ij}}{\rho_1}\right)[1 - \exp\left(\frac{-d_{2ij}}{\rho_2}\right)]$$

$$\boldsymbol{S_{1e}} = Cov\left[d_{1ij}, d_{2ij}\right] = \tau^2[\exp\left(\frac{-d_{1ij}}{\rho_1}\right) + \exp\left(\frac{-d_{2ij}}{\rho_2}\right)]$$

Simulated outcome and distance data were chosen to fairly closely resemble the Esmeraldes diarrhea data. However, differences between the distribution of simulated versus real data remain the most likely reason for the differences witnessed in model performance. For example, one might reconsider the log-Gaussian assumption on the outcome. Given that the incidence rate could be considered count data with a significant number of zeros, possible alternatives might be a Zero-Inflated Poisson model or another mixture model.

These analyses focused on estimating spatiotemporal dependence at locations that are connected for all distance metrics being examined. It is more practical, however, to build a weighted multi-distance covariance function that allows for locations that are not connected by road or river. We have seen that valid covariance functions without this assumption have similar convergence issues (Appendix Figures 1a-2b), but recognize that there a variety of options for developing similar models capable of handling this nuance. This would allow the incorporation

of road distance matrices, and consequently weighted combinations of the three matrices together.

Furthermore, for simplicity, missing incidence values were imputed based solely on using observed information close in time. However, it may be useful to consider a more sophisticated approach that imputes missing data using both observed information close in time and space, using a classification method such as k-nearest-neighbor for nearby spatial points (*Cover 1967*).

Another important extension of this paper would be the inclusion of community-level covariates. For estimating diarrheal disease incidence between different villages, for example, it would be useful to consider village-level factors such as remoteness, social cohesion, and travel patterns. It would also be useful to average over selected individual-level factors associated with diarrheal disease such as the proportion of the population with more than six years of education and the percentage of the population less than five years of age. With regard to temporal estimation, precipitation data would be an important determinant of village-level diarrheal disease that could be incorporated into the model.

Finally, rivers and roads not only consist of different paths that individuals, pathogens, or goods may take, but they are also governed by important characteristics such as direction and flow. As mentioned previously, *Ver Hoef et al.* (2006) have developed a moving average approach to this estimation that provides a useful framework for incorporating these factors into a covariance function. This discussion mentions a few of the many important extensions that this paper may motivate, with wide applications in infectious disease research, ecology, and a variety of other fields.

TABLES & FIGURES

Table 1. Model comparison results with the analysis of simulated and observed data: deviance information criterion (DIC), Watanabe-Akaike information criterion (wAIC), and posterior predictive loss (PPL)

Outcome[1]	Distance Matrix		s[1]	t[2]	DIC	wAIC	PPL	Trace Plots
Simulated	Combined	Simulated	18	75	-233[4]	2890	464	2a & 2b
Simulated	Euclidean Only	Simulated	18	75	2184	3765	846	2c
Simulated	River Only	Simulated	18	75	4408	5554	3298	2d
Simulated	Combined	Simulated	200	10	-1[4]	240	31	Appendix 3a & 3b
Simulated	Euclidean Only	Simulated	200	10	-231	415	48	Appendix 3c
Simulated	River Only	Simulated	200	10	24	567	94	Appendix 3d
Simulated	Combined	Esmeraldes Data	18	75	5229	5404	3302	3a & 3b
Simulated	Euclidean Only	Esmeraldes Data	18	75	4875	5436	3363	3c
Simulated	River Only	Esmeraldes Data	18	75	3051	4417	1422	3d
Incidence Rate	Combined	Esmeraldes Data	18	30	3448	3551	27305	4a & 4b
Incidence Rate	Euclidean Only	Esmeraldes Data	18	30	3452	3531	31490	4c
Incidence Rate	River Only	Esmeraldes Data	18	30	3437	3546	27238	4d

[1]For simulated outcomes, the outcome was generated based on the model with the combined distance matrix.
[2]Number of villages
[3]Number of weeks
[4]Note: Negative DIC values are, in fact, possible as DIC is defined as the sum of a negative goodness of fit term and a positive penalty term

Table 2. Parameter estimates and (95%) posterior intervals using either simulated or observed distance matrices with simulated Gaussian outcomes. All outcomes were simulated assuming a combined covariance distance matrix of both river and Euclidean distance.

		Simulated Distance Matrices			Observed Distance Matrices		
Model	**Parameter**	**True Value**	**Estimate**	**95% PI**	**True Value**	**Estimate**	**95% PI**
Combined	τ^2	10	10.1	(9.1, 11.1)	10	8.6	(6.6, 10)
Combined	θ	0.7	0.7	(0.5, 0.9)	0.6	0.8	(0.4, 1)
Combined	ρ_1 (Euclidean)	10	9.9	(6.9, 12.9)	8	10.4	(7.9, 15.1)
Combined	ρ_2 (River)	10	9.3	(6.4, 13.5)	10	7.9	(6.3, 10.4)
Combined	σ^2	0.5	0	(0, 0.2)	1	1.2	(0.7, 2.1)
Euc. Only	τ^2	10	9.9	(9.1, 10.8)	10	8.8	(7.4, 10.1)
Euc. Only	ρ_1 (Euclidean)	10	10.6	(8.5, 12.9)	10	10	(8.2, 12.9)
Euc. Only	σ^2	0.5	0.2	(0, 0.5)	1	1.3	(0.5, 2.2)
River Only	τ^2	10	9.1	(7.1, 10.6)	10	9.6	(8.1, 10.8)
River Only	ρ_2 (River)	10	7.5	(5.8, 9.7)	8	7.4	(6.3, 8.8)
River Only	σ^2	0.5	1.1	(0.1, 2.9)	1	0.3	(0, 1.4)

Table 3. Parameter estimates and 95% posterior intervals for the analysis diarrheal incidence among 18 villages in Ecuador.

Model	**Parameter**	**Estimate**	**95% PI**
Combined	τ^2	22.1	(16.4, 29.8)
Combined	θ	0.1	(0, 0.2)
Combined	ρ_1 (Euclidean)	36	(16.6, 83.4)
Combined	ρ_2 (River)	25.2	(24, 26.4)
Combined	σ^2	25.5	(21.6, 30.1)
Euc. Only	τ^2	19.4	(13.3, 28.6)
Euc. Only	ρ_1 (Euclidean)	109.9	(58.8, 150.8)
Euc. Only	σ^2	29.6	(25.6, 33.9)
River Only	τ^2	22.2	(16.3, 30.2)
River Only	ρ_2 (River)	25.2	(24, 26.5)
River Only	σ^2	25.3	(21.2, 29.7)

Figure 1a. Map of the Esmeraldes province in Ecuador showing the three major rivers in blue (figure from *Ahn et al.*, 2014)

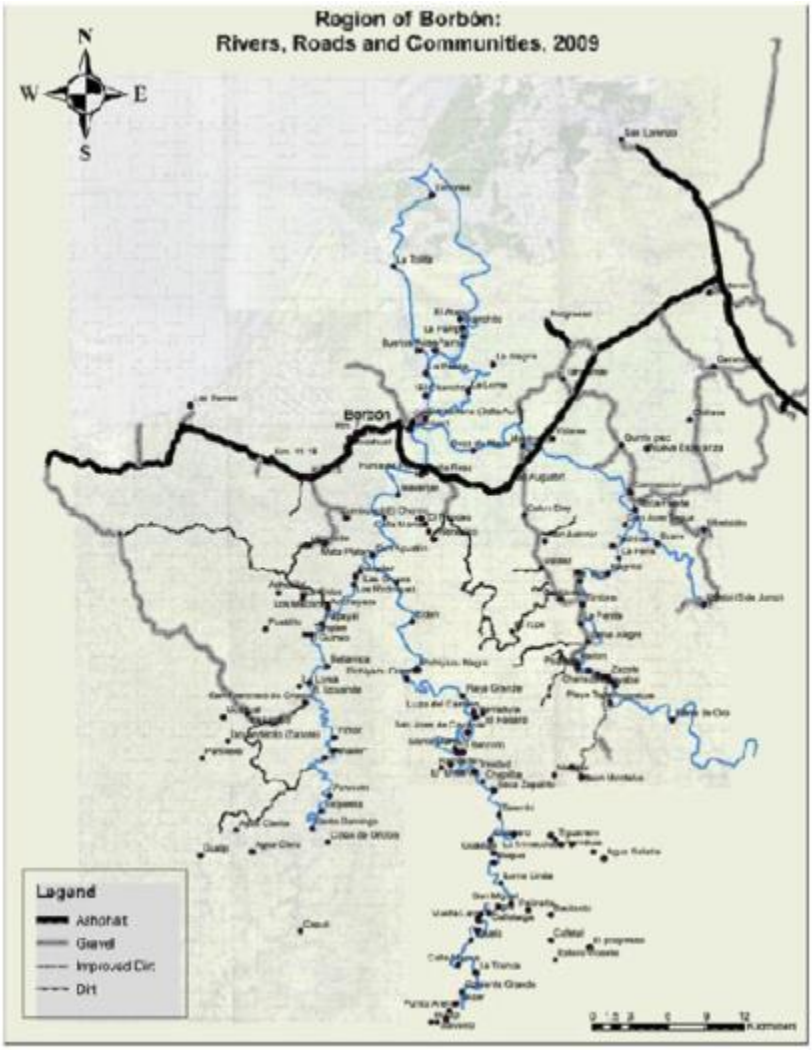


Figure 1b. Map of the 21 villages of Esmeraldes chosen by random-block design (in red), the road network (in black) and the river network (in blue)

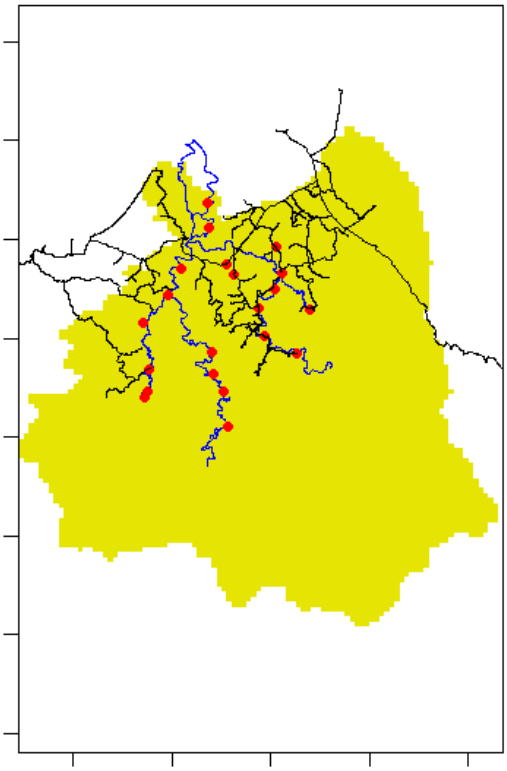


Figure 2a. Trace plots and histograms of the posterior distributions of the posterior distributions of τ^2, and σ^2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices

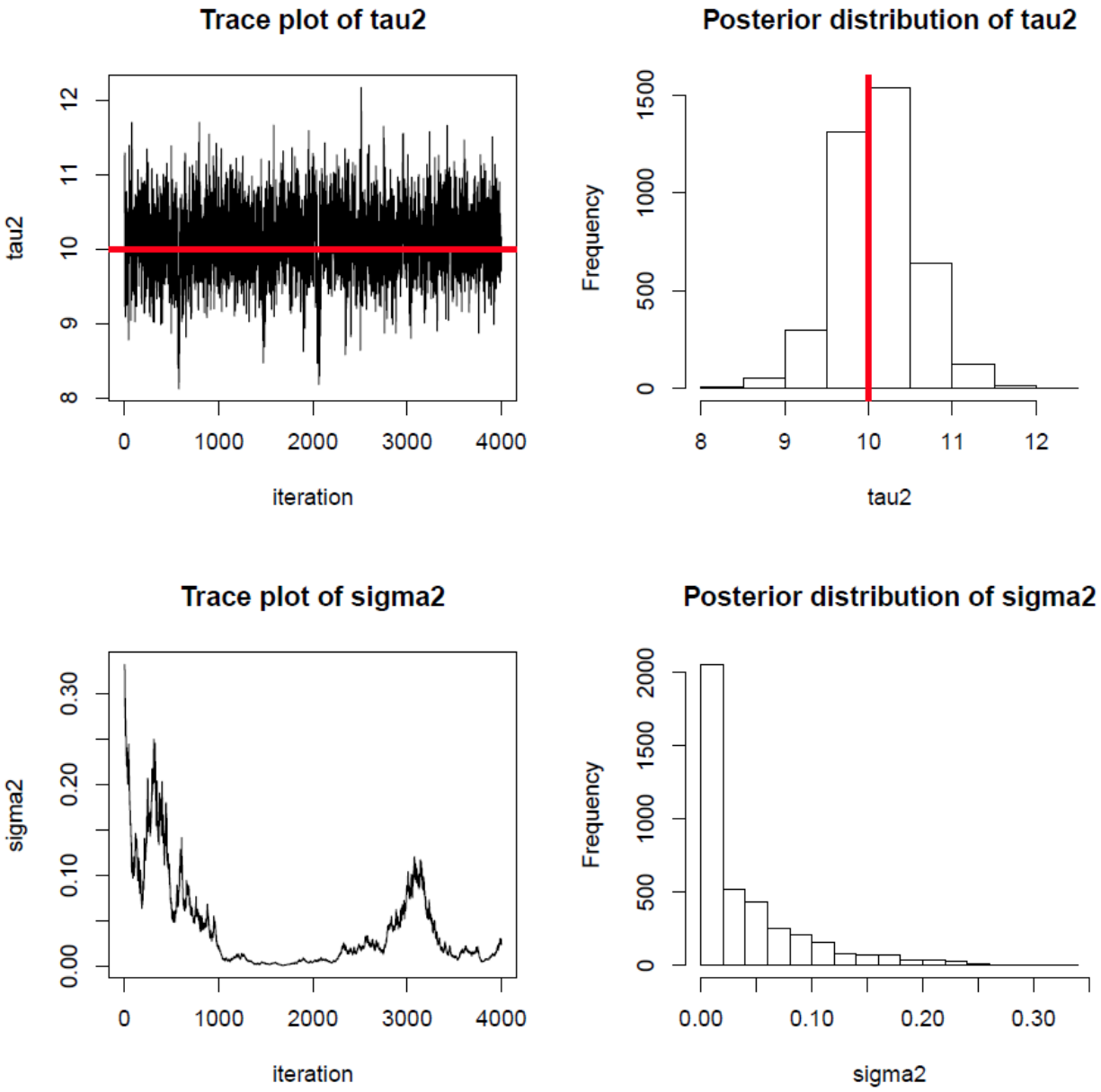


Figure 2b. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices

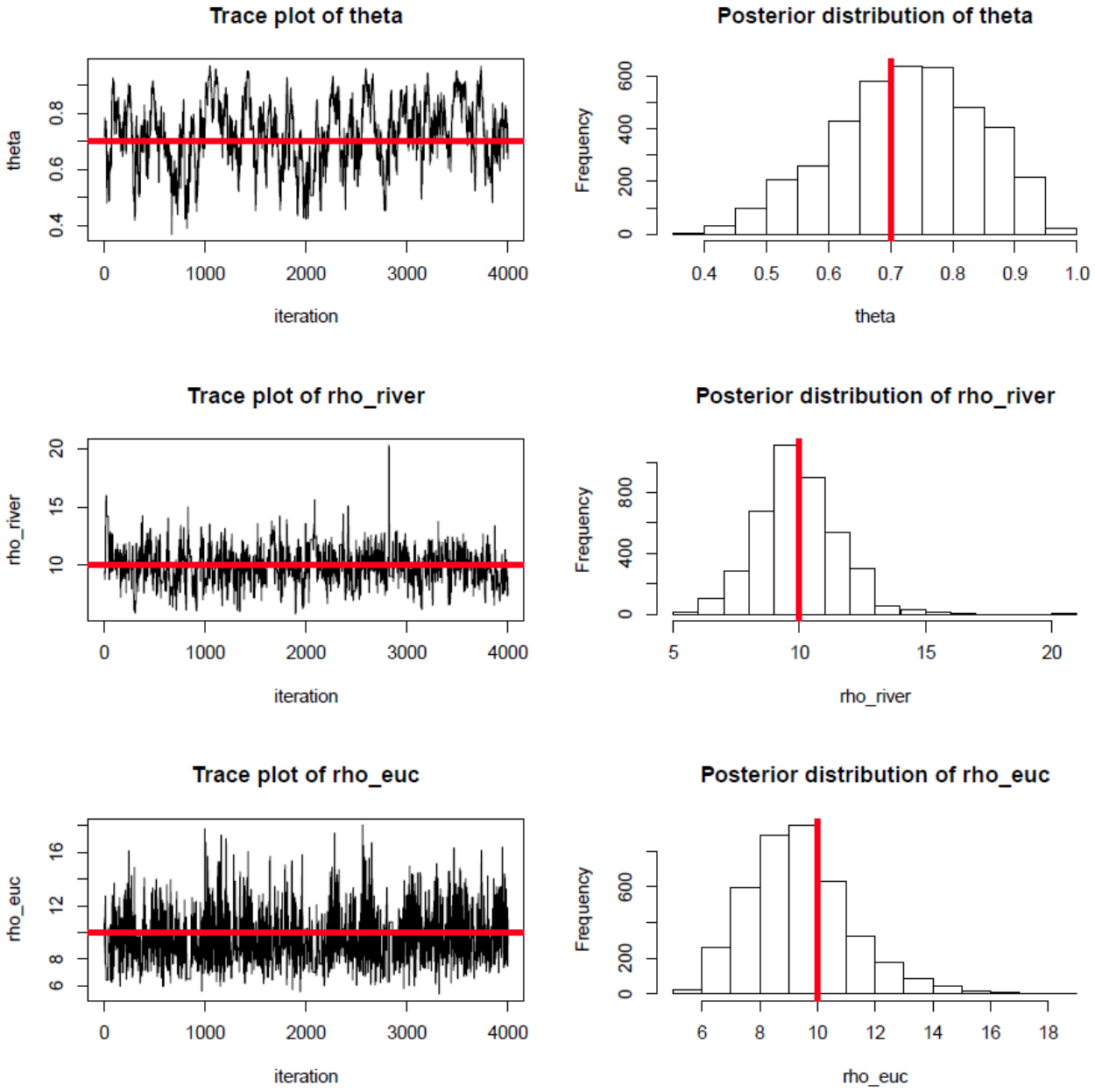


Figure 2c. Trace plots and histograms of the posterior distributions of τ^2, ρ_1, and σ^2 for Euclidean Only covariance function ($\mathbf{S_2}$) for simulated outcome and simulated distance matrices

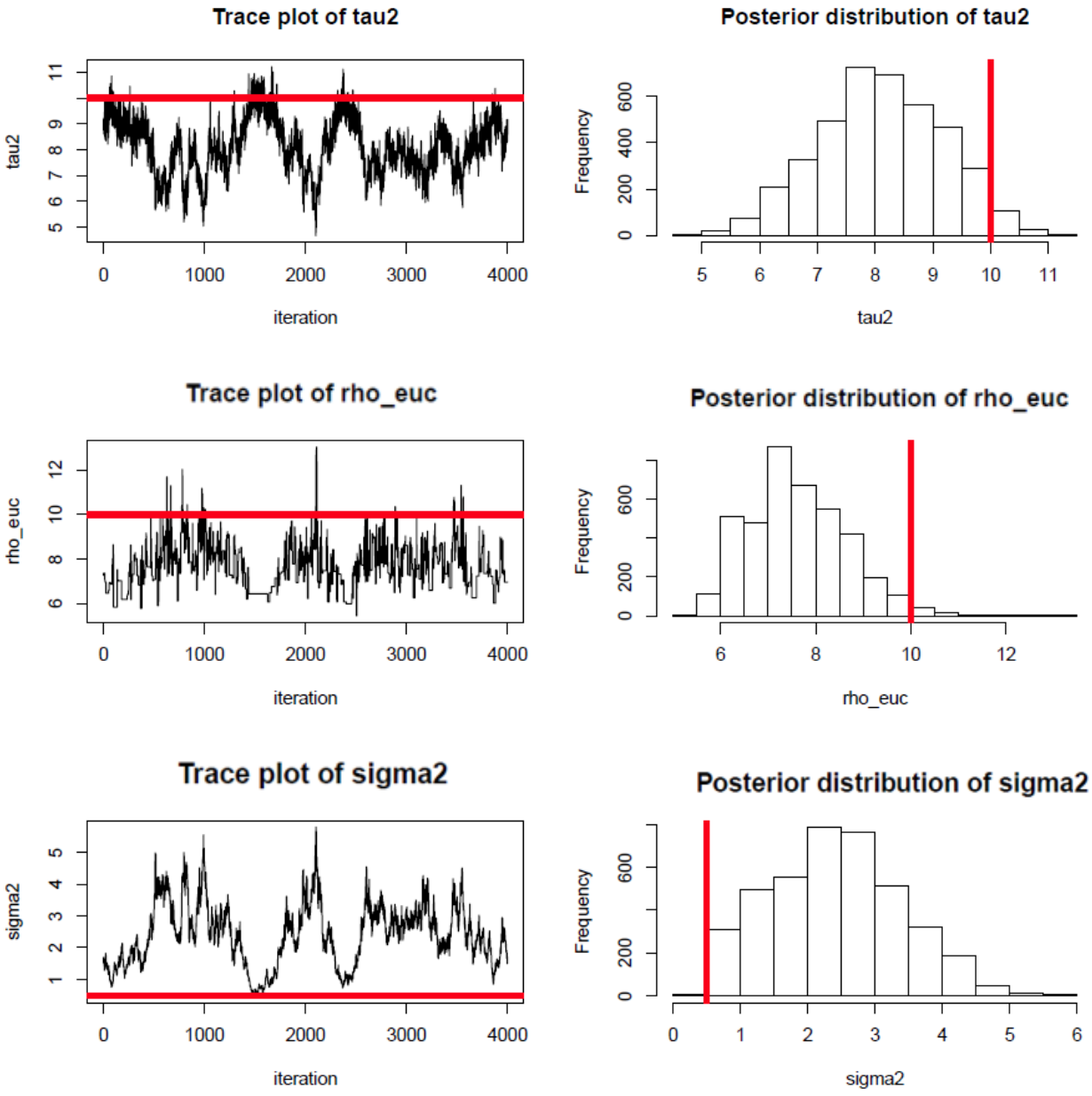
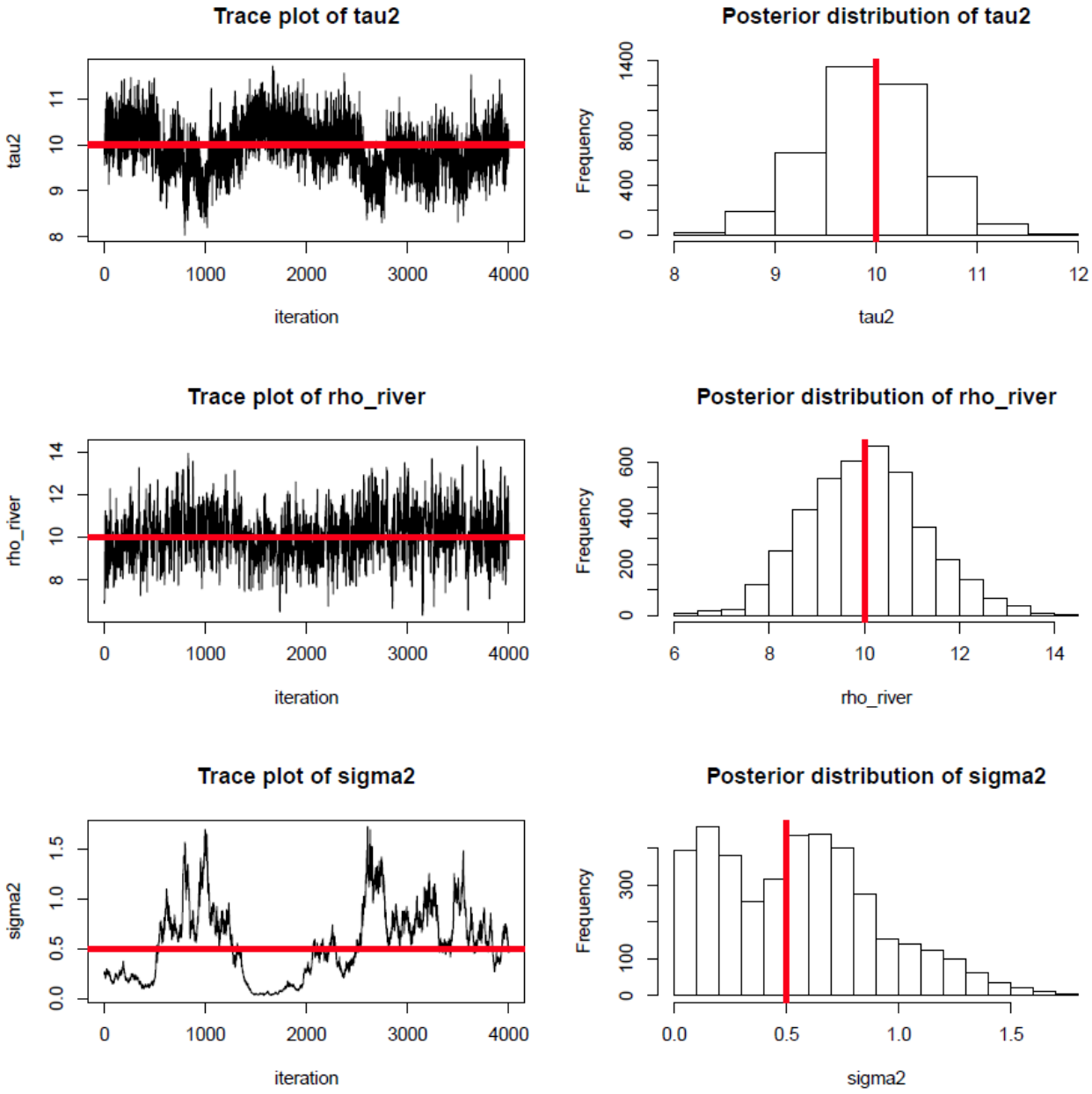


Figure 2d. Trace plots and histograms of the posterior distributions of τ^2, ρ_2, and σ^2 for River Only covariance function ($\mathbf{S}_3$) for simulated outcome and simulated distance matrices



Figures 3a. Trace plots and histograms of the posterior distributions of τ^2, and σ^2 for combined covariance function ($\mathbf{S}_1$) for simulated outcome and actual distance matrices

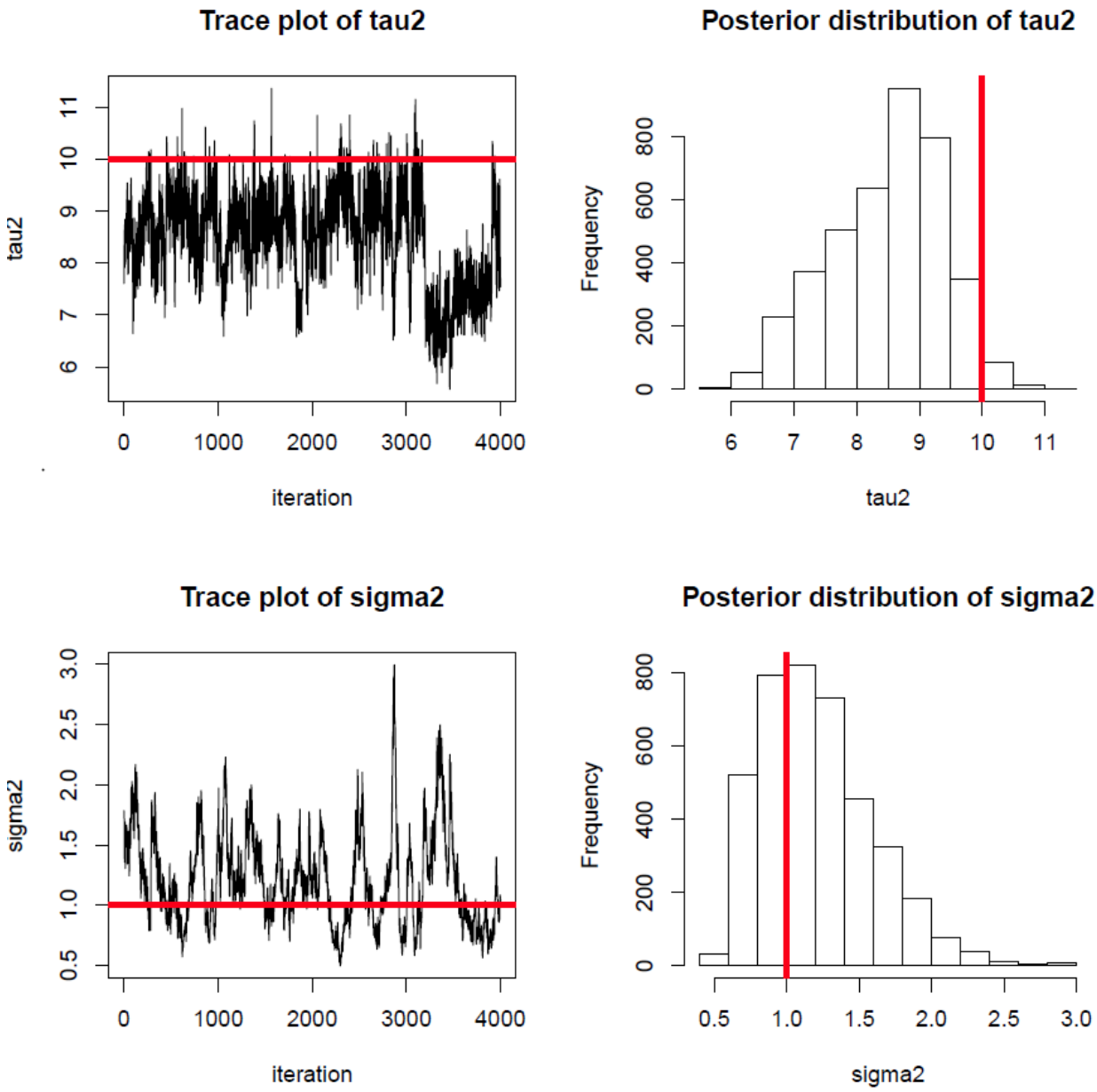


Figure 3b. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and actual distance matrices

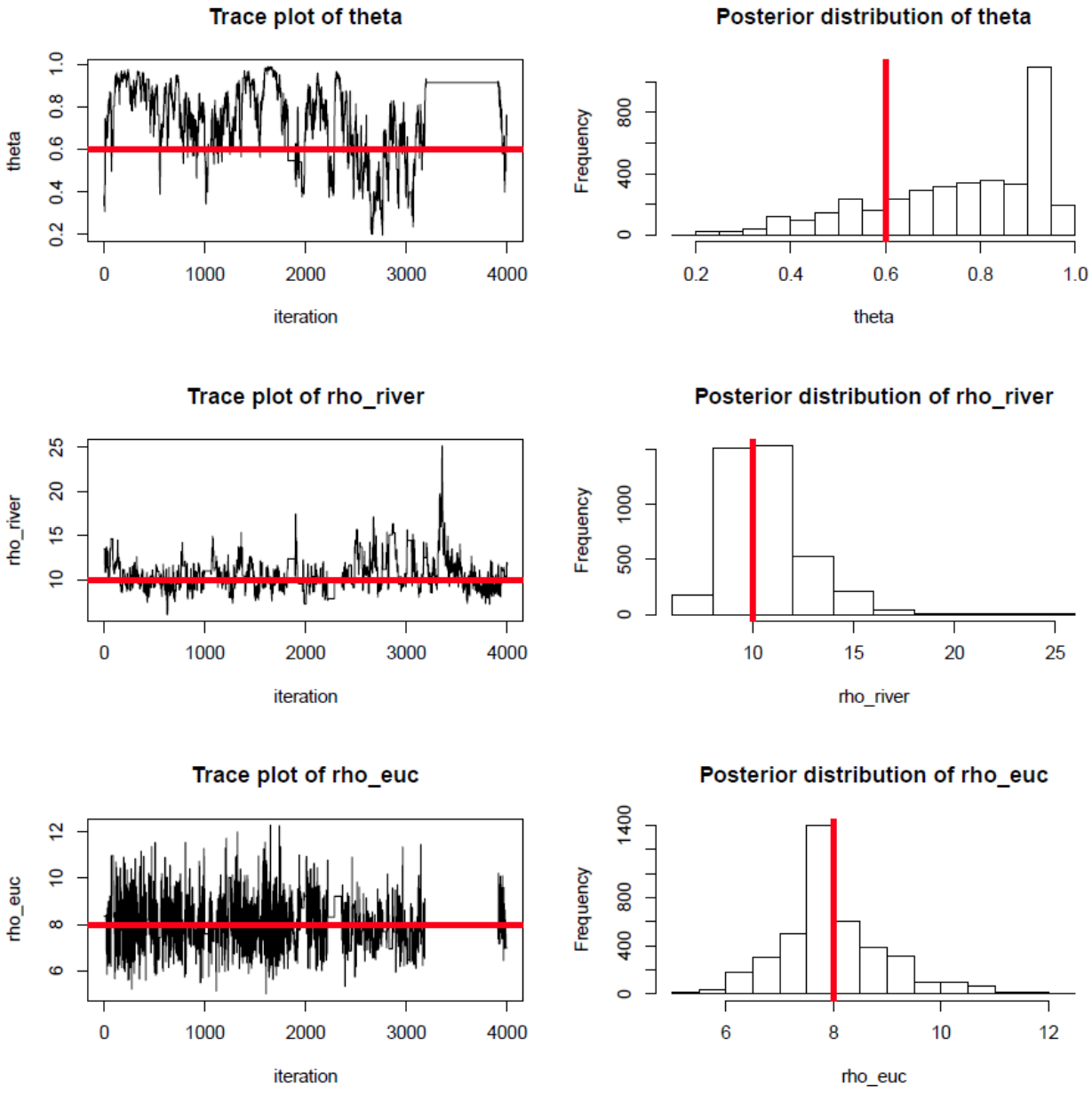


Figure 3c. Trace plots and histograms of the posterior distributions of τ^2, ρ_1, and σ^2 for Euclidean Only covariance function ($\mathbf{S_2}$) for simulated outcome and actual distance matrices

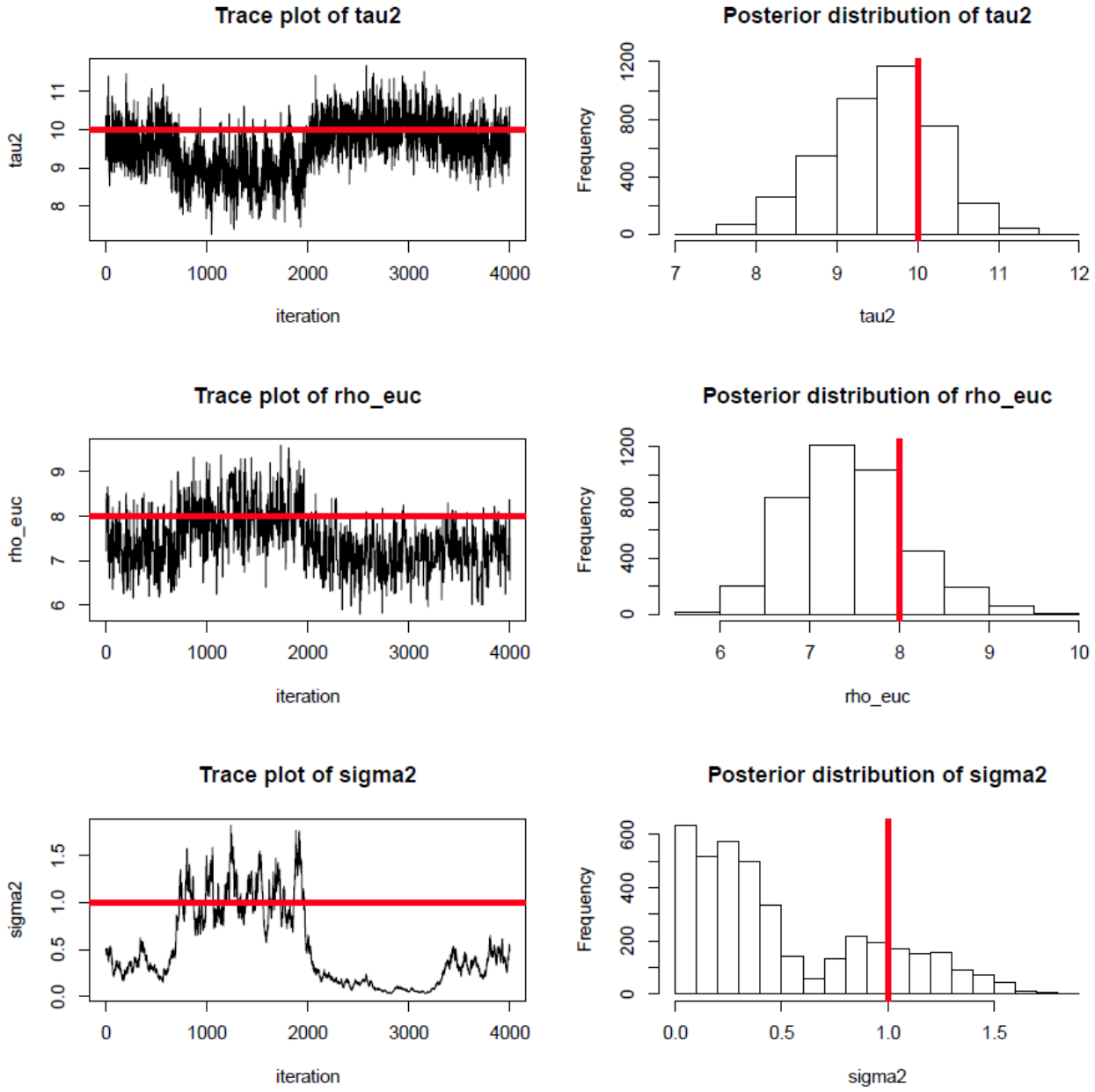


Figure 3d. Trace plots and histograms of the posterior distributions of τ^2, ρ_2, and σ^2 for River Only covariance function ($\mathbf{S}_3$) for simulated outcome and actual distance matrices

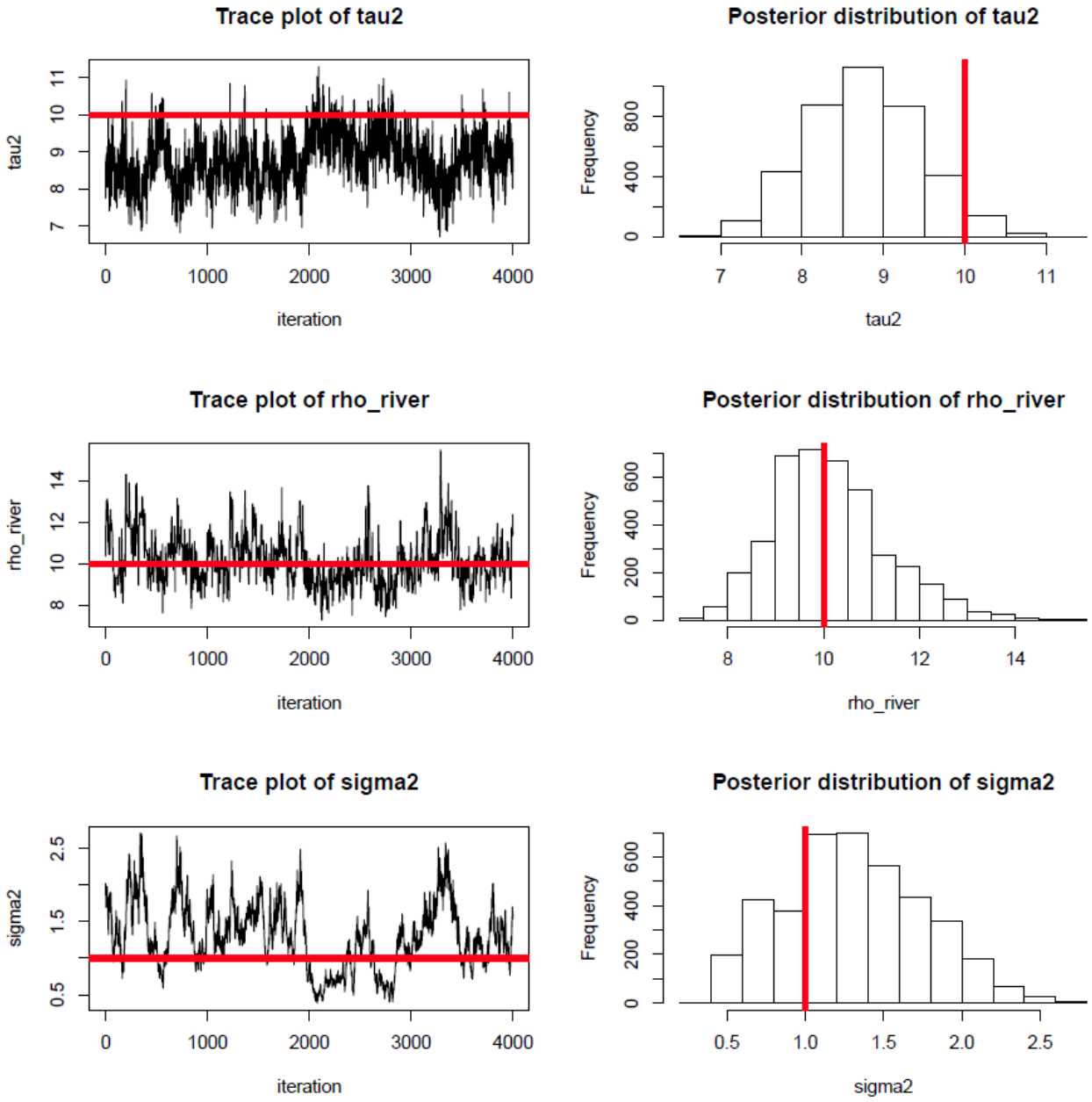


Figure 4a. Trace plots and histograms of the posterior distributions of τ^2, and σ^2 for combined covariance function ($\mathbf{S_1}$) for actual outcome and actual distance matrices

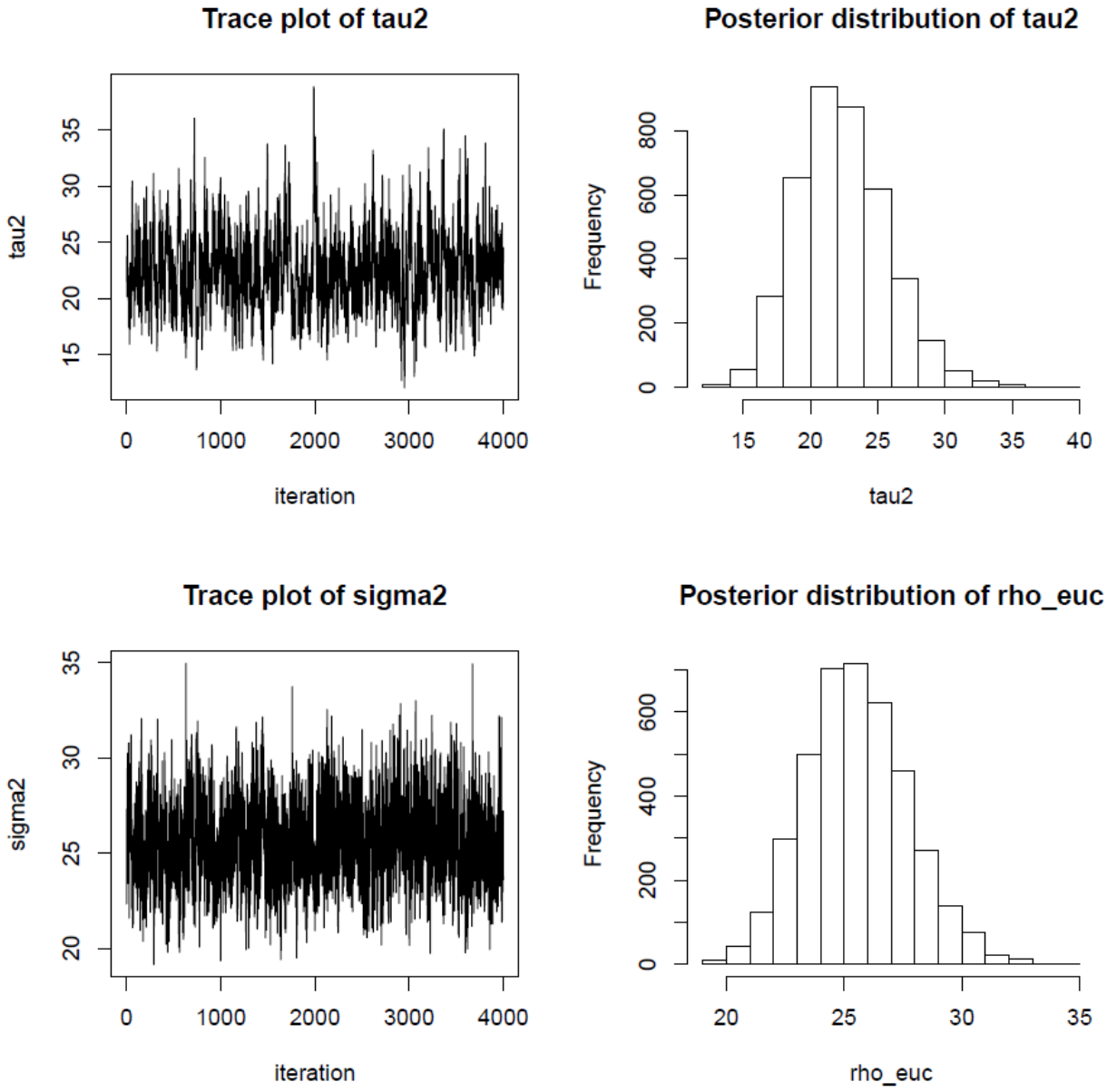


Figure 4b. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for actual outcome and actual distance matrices

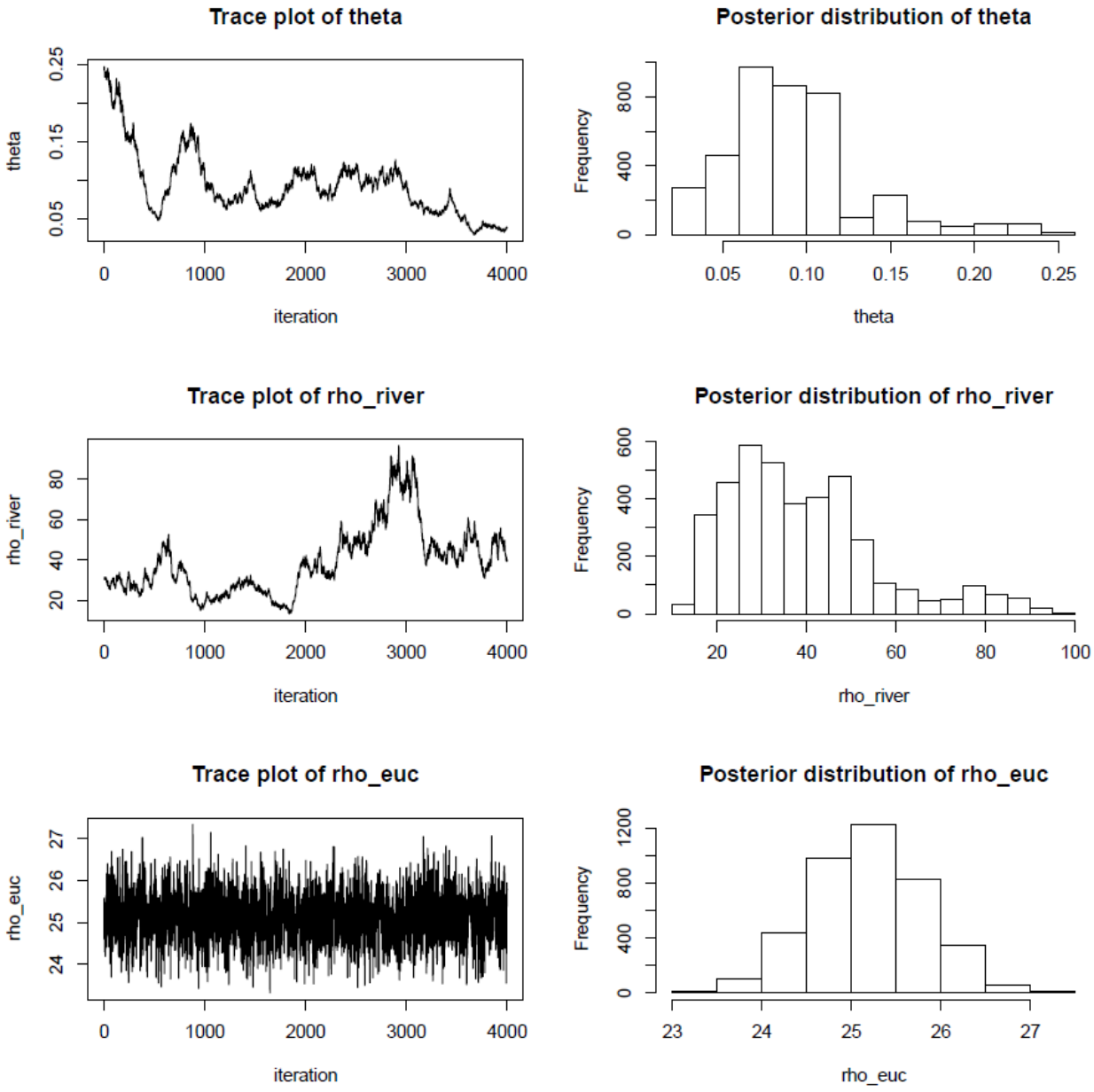


Figure 4c. Trace plots and histograms of the posterior distributions of τ^2, ρ_1, and σ^2 for Euclidean Only covariance function ($\mathbf{S_2}$) for actual outcome and actual distance matrices

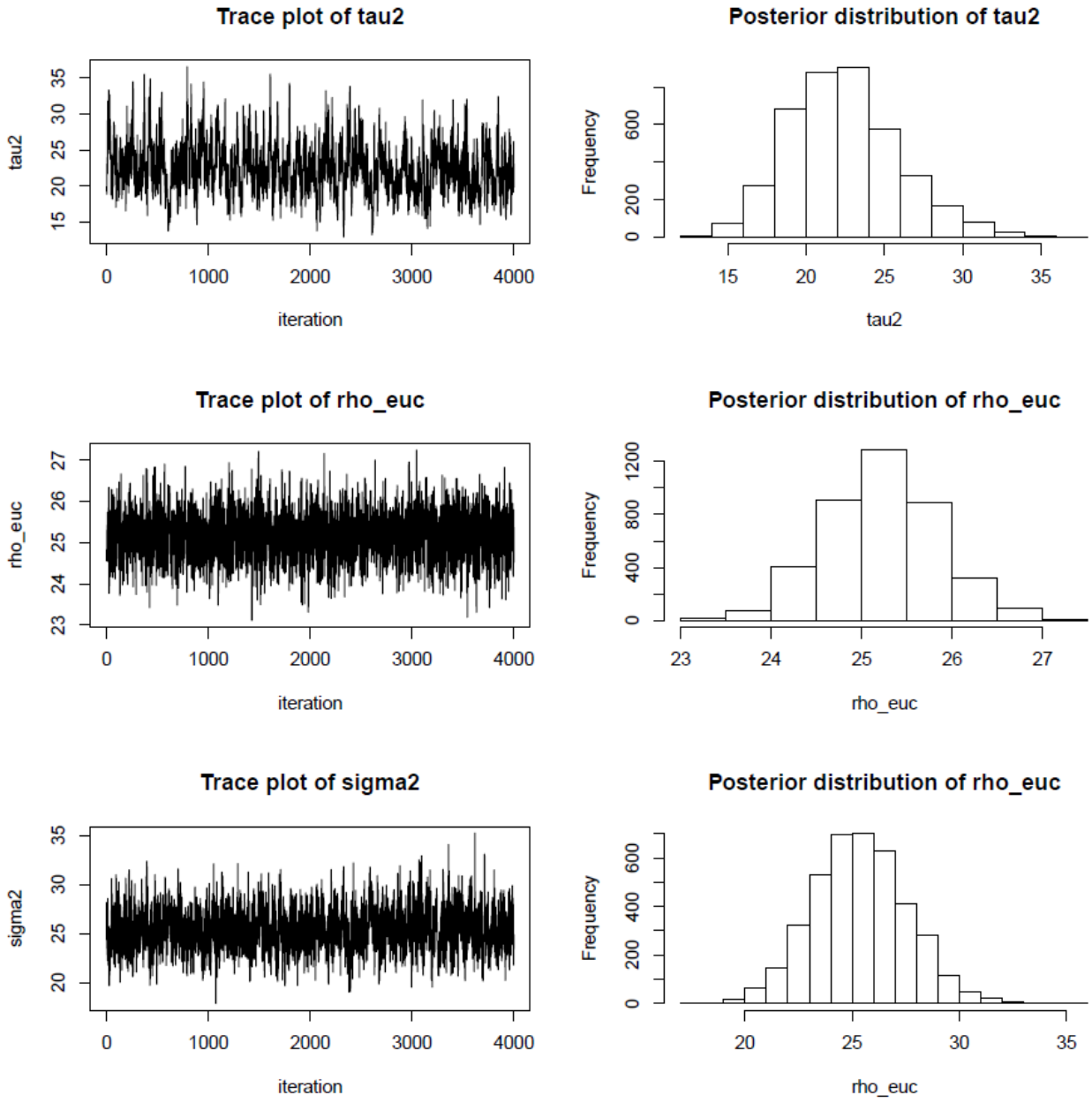


Figure 4d. Trace plots and histograms of the posterior distributions of τ^2, ρ_2, and σ^2 for River Only covariance function ($\mathbf{S_3}$) for actual outcome and actual distance matrices

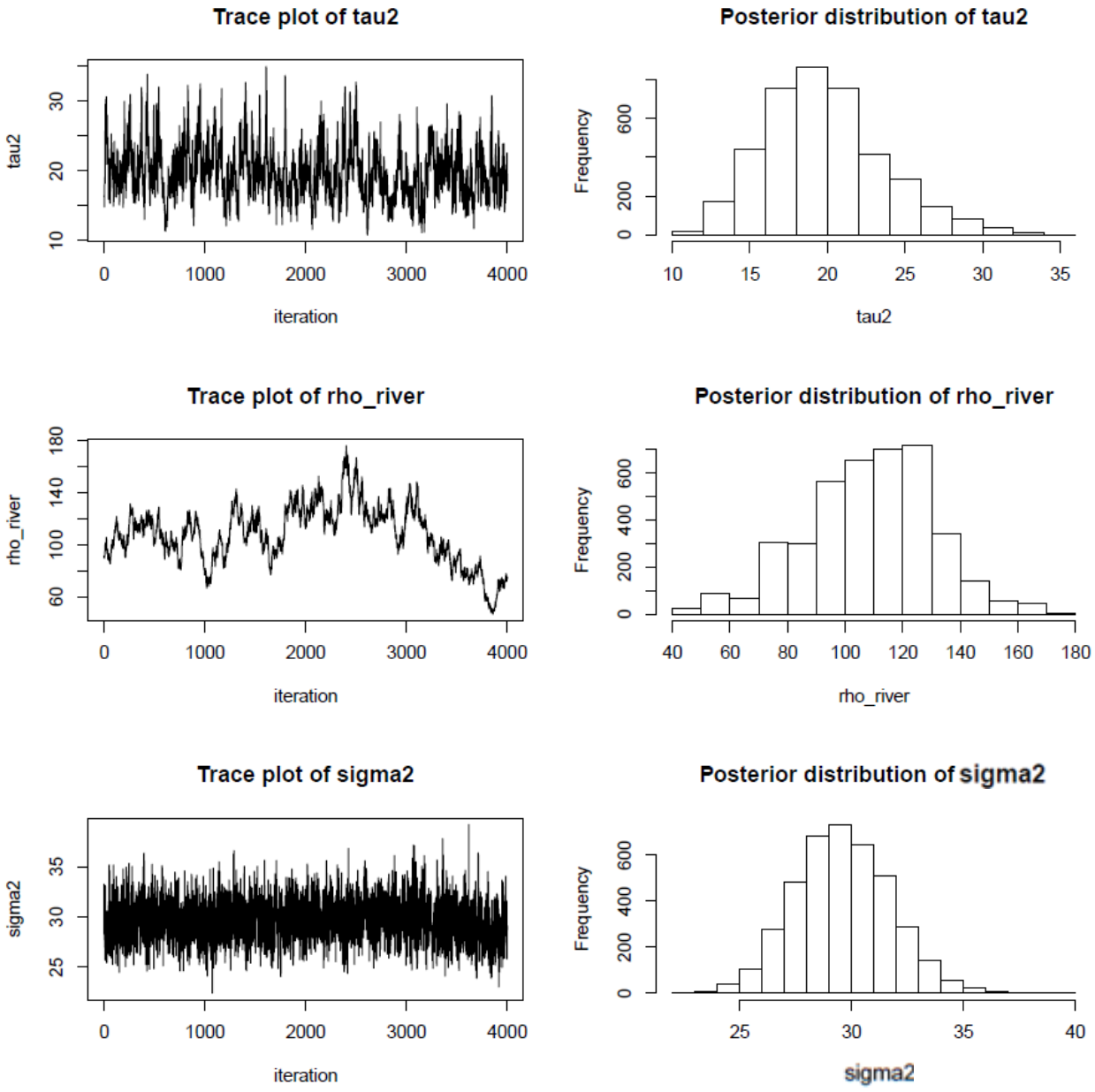


Figure 5. Weekly Average Incidence Rates of Diarrheal Disease

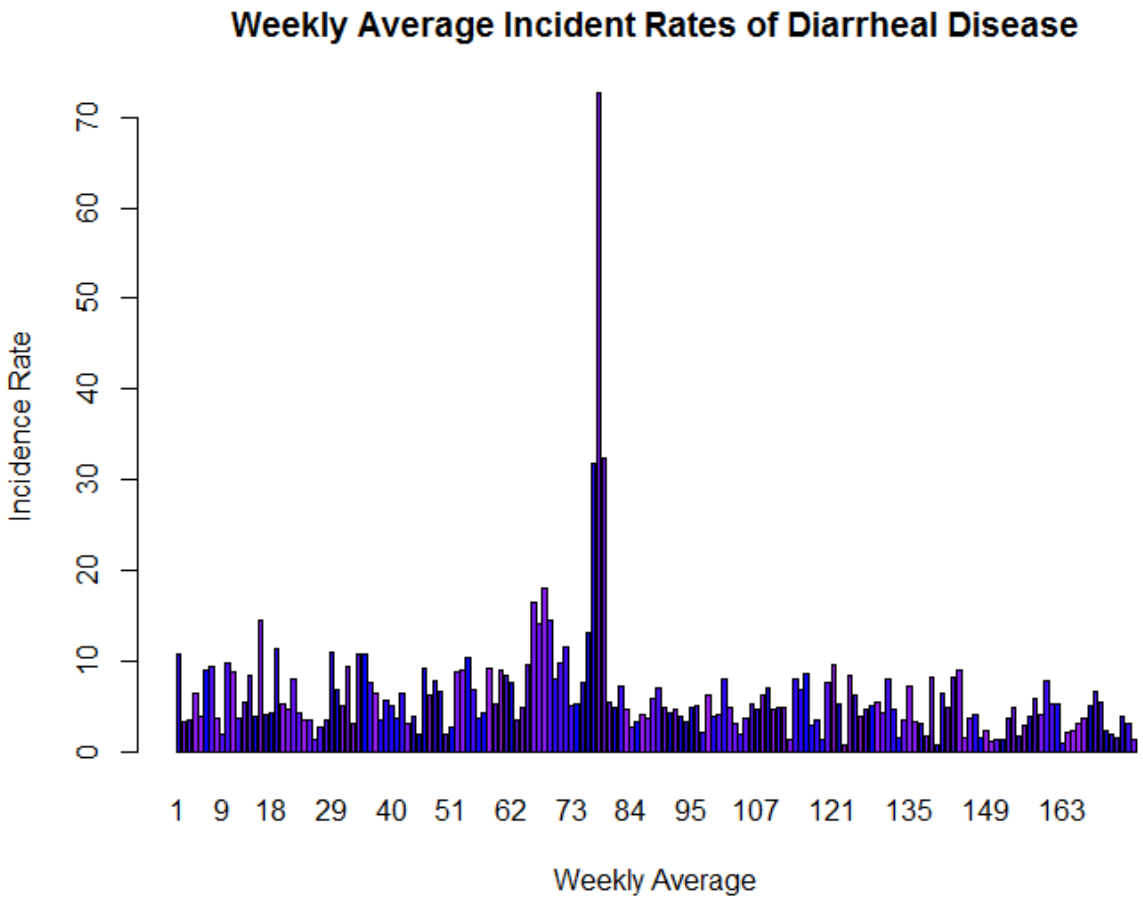
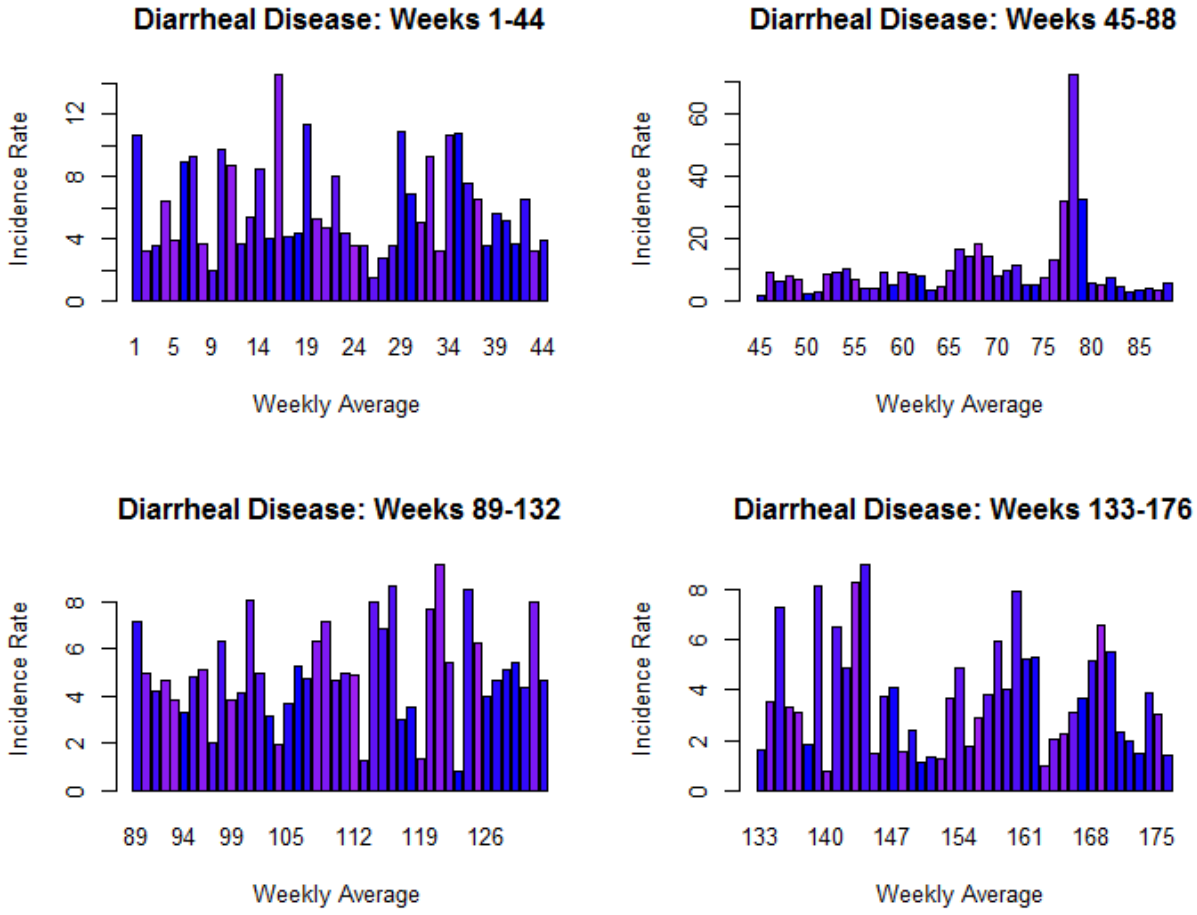
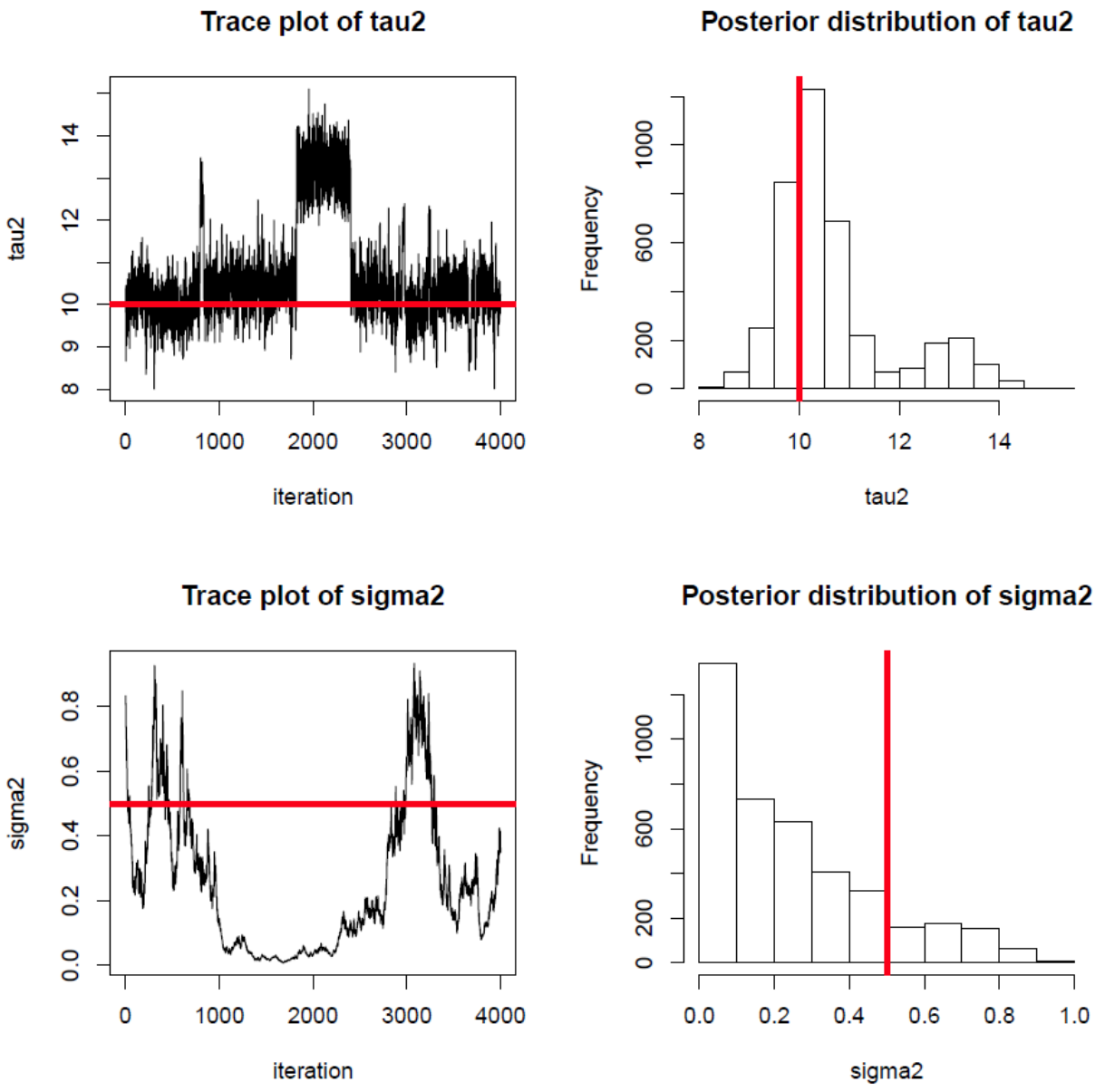


Figure 6. Weekly Average Incidence Rates of Diarrheal Disease

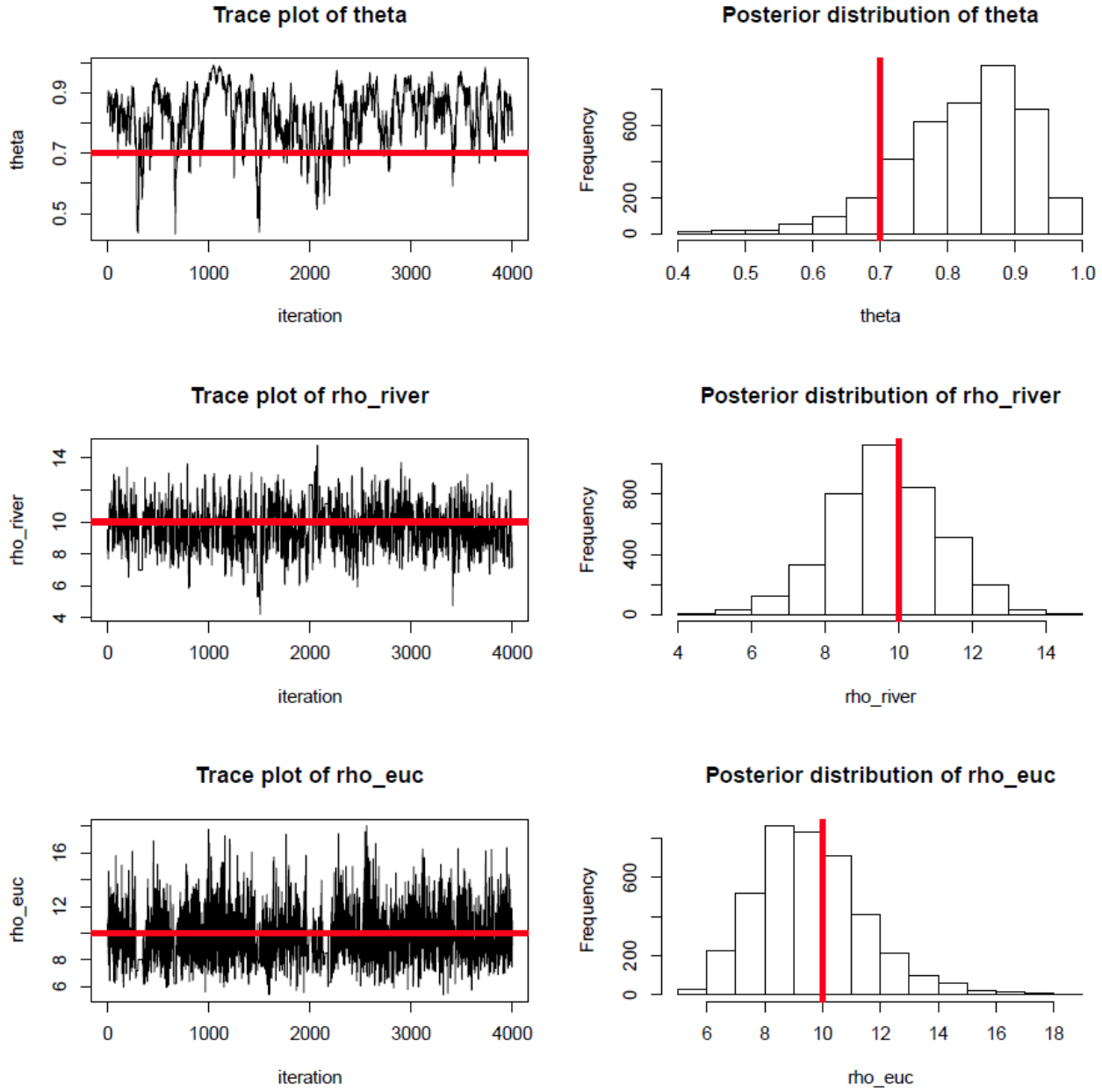


APPENDIX

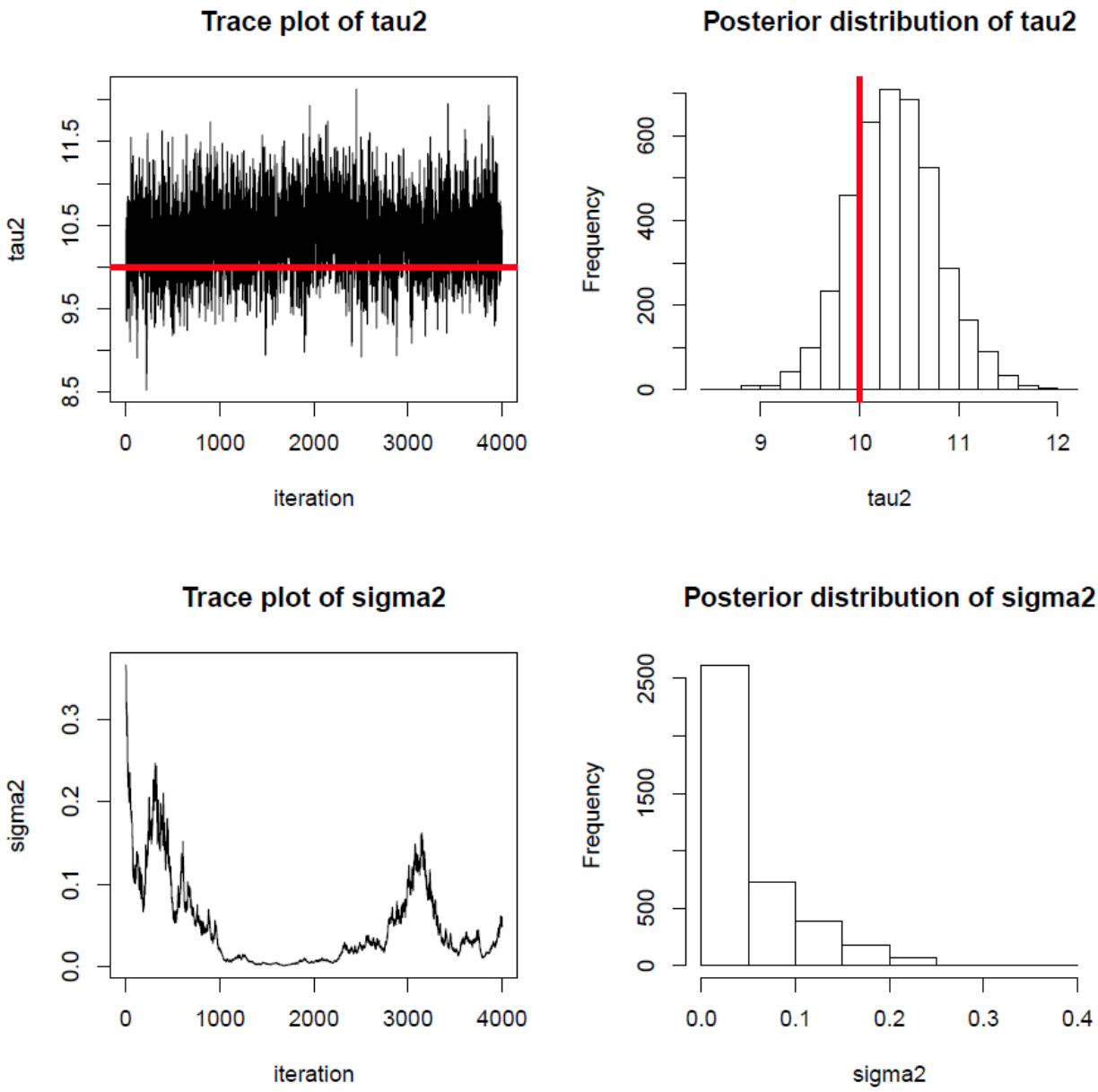
Appendix Figure 1a. Trace plots and histograms of the posterior distributions of the posterior distributions of τ^2, and σ^2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices with 90% of the points connected by river



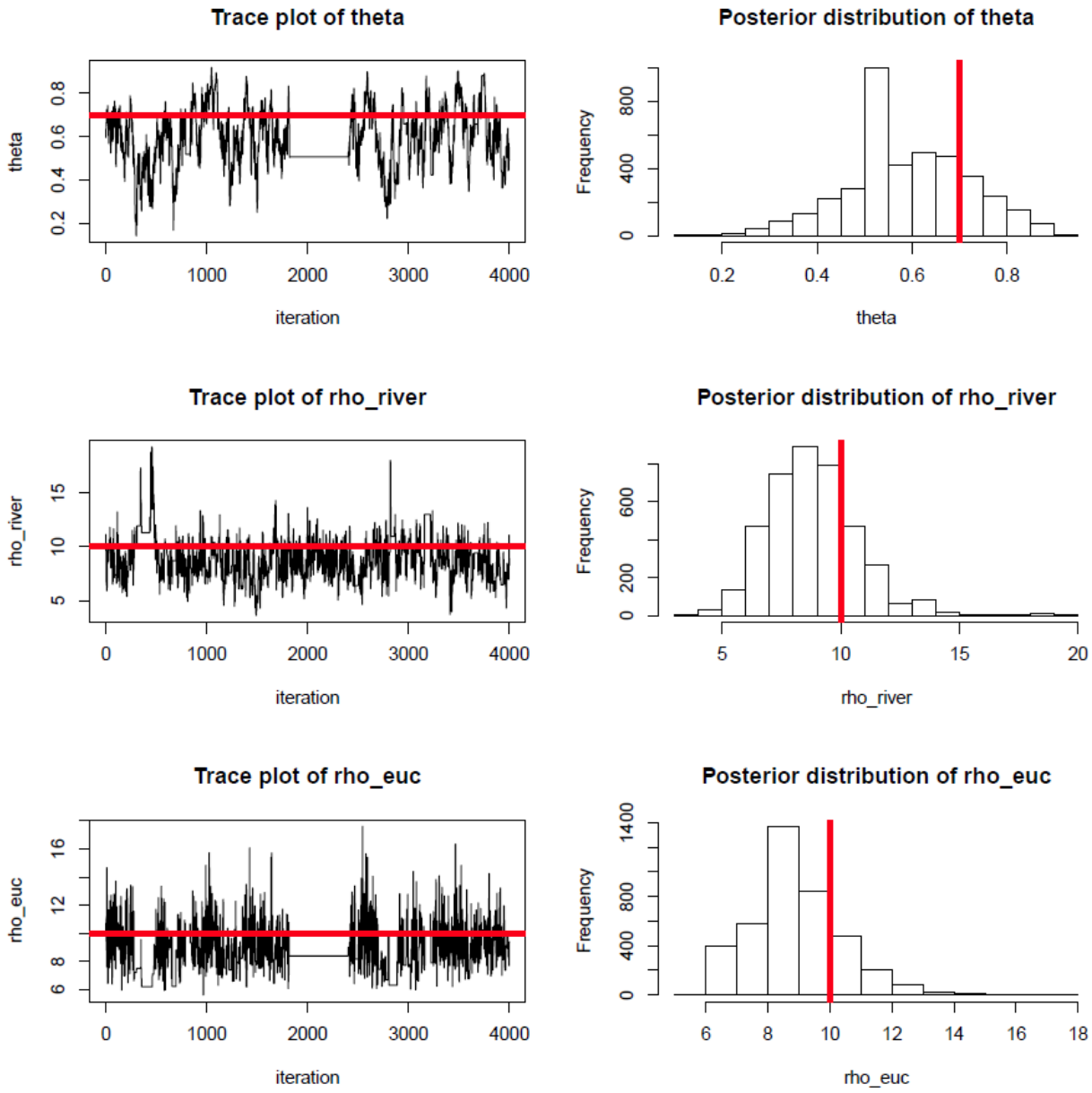
Appendix Figure 1b. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices with 90% of the points connected by river



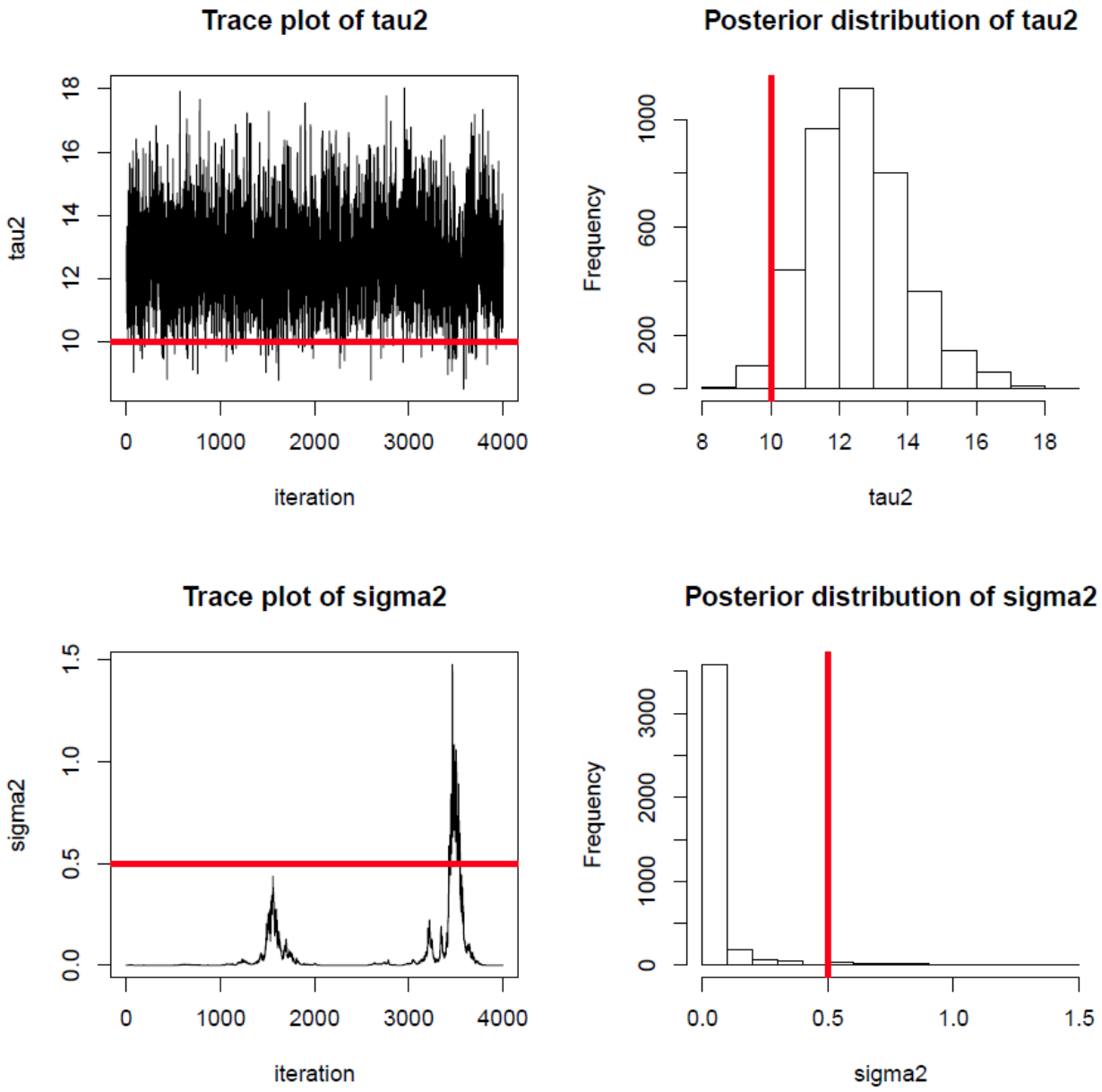
Appendix Figure 2a. Trace plots and histograms of the posterior distributions of the posterior distributions of τ^2, and σ^2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices with 75% of the points connected by river



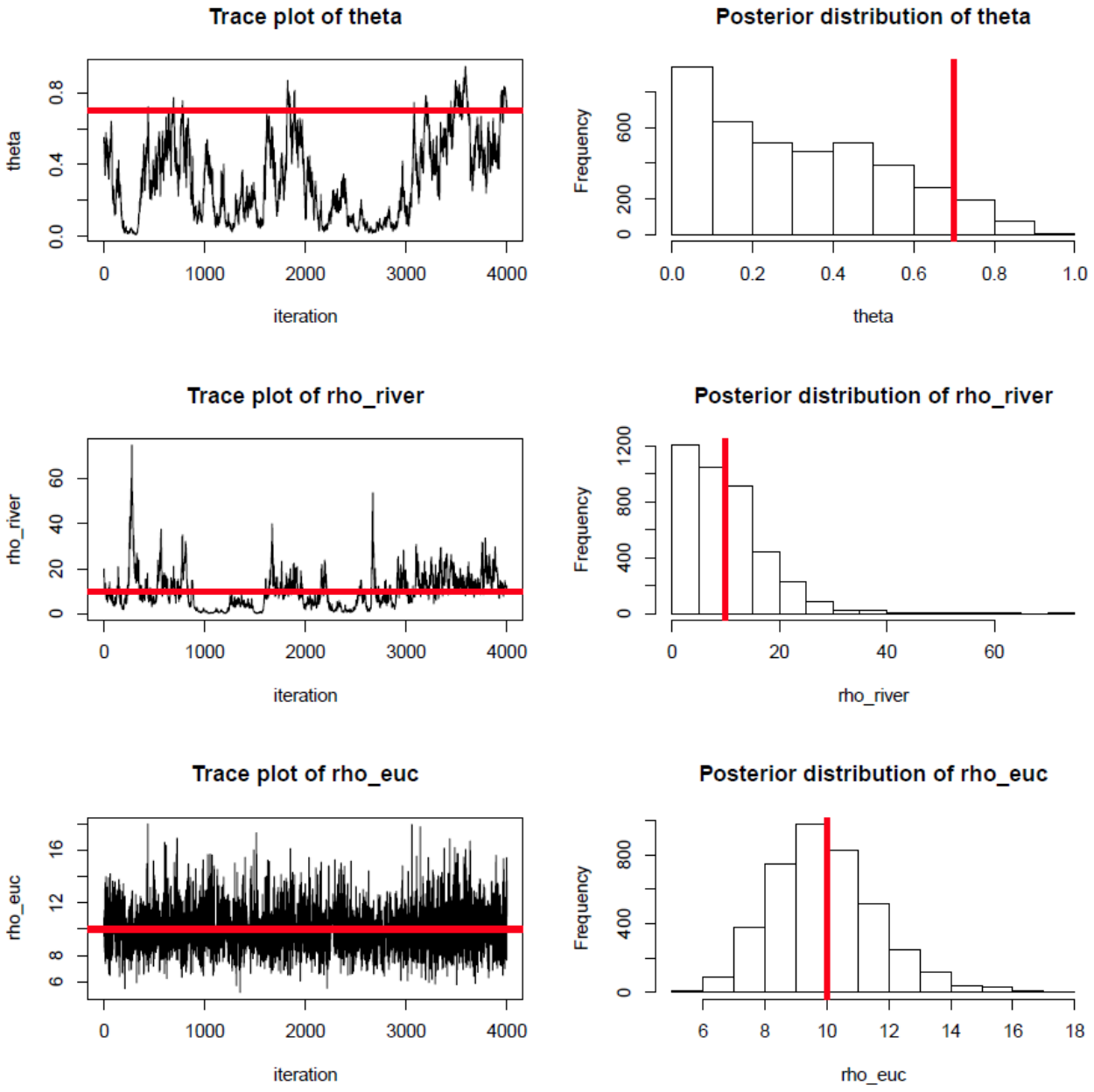
Appendix Figure 2b. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices with 75% of the points connected by river



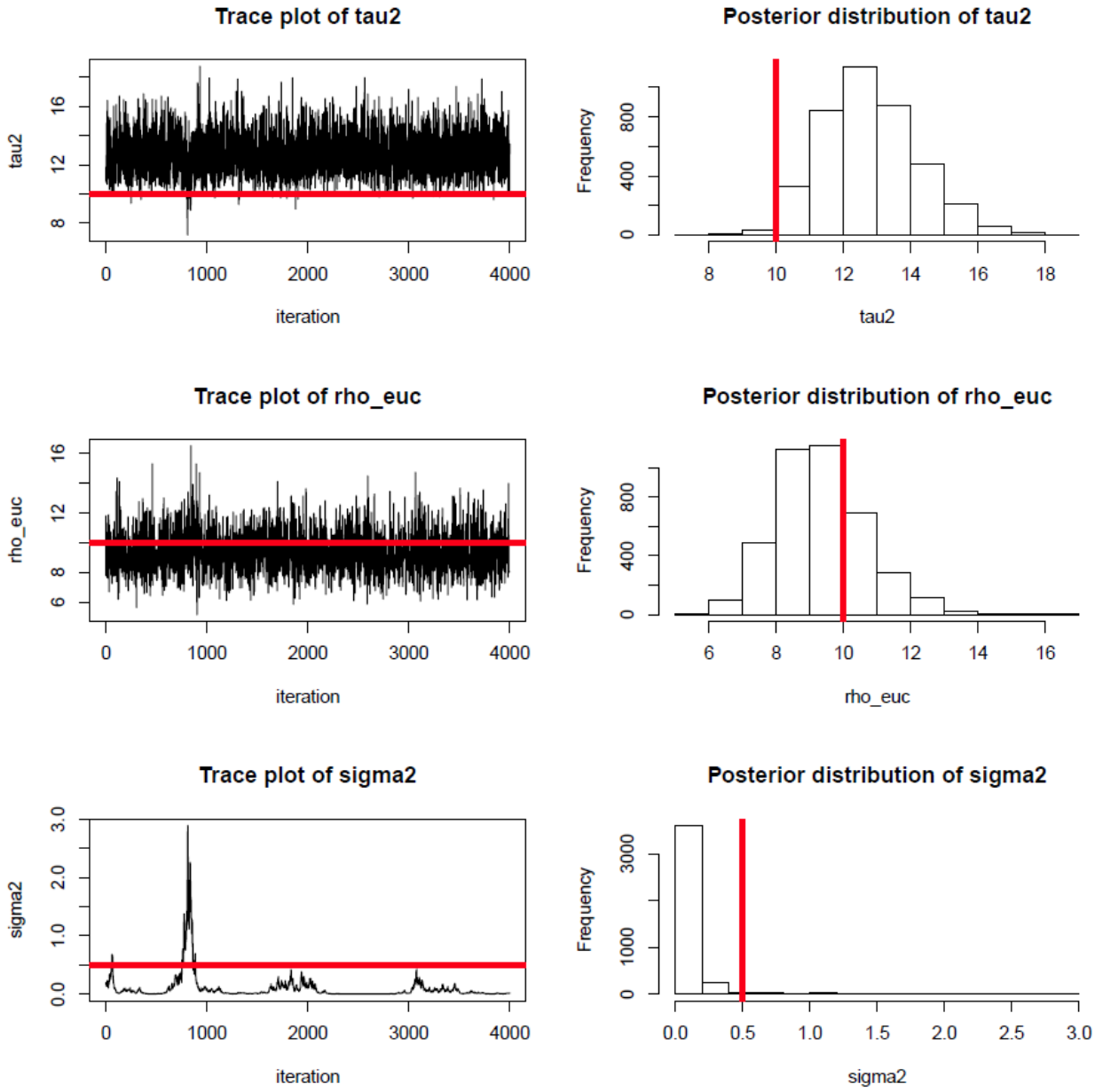
Appendix Figure 3a. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices for s=200, t=10



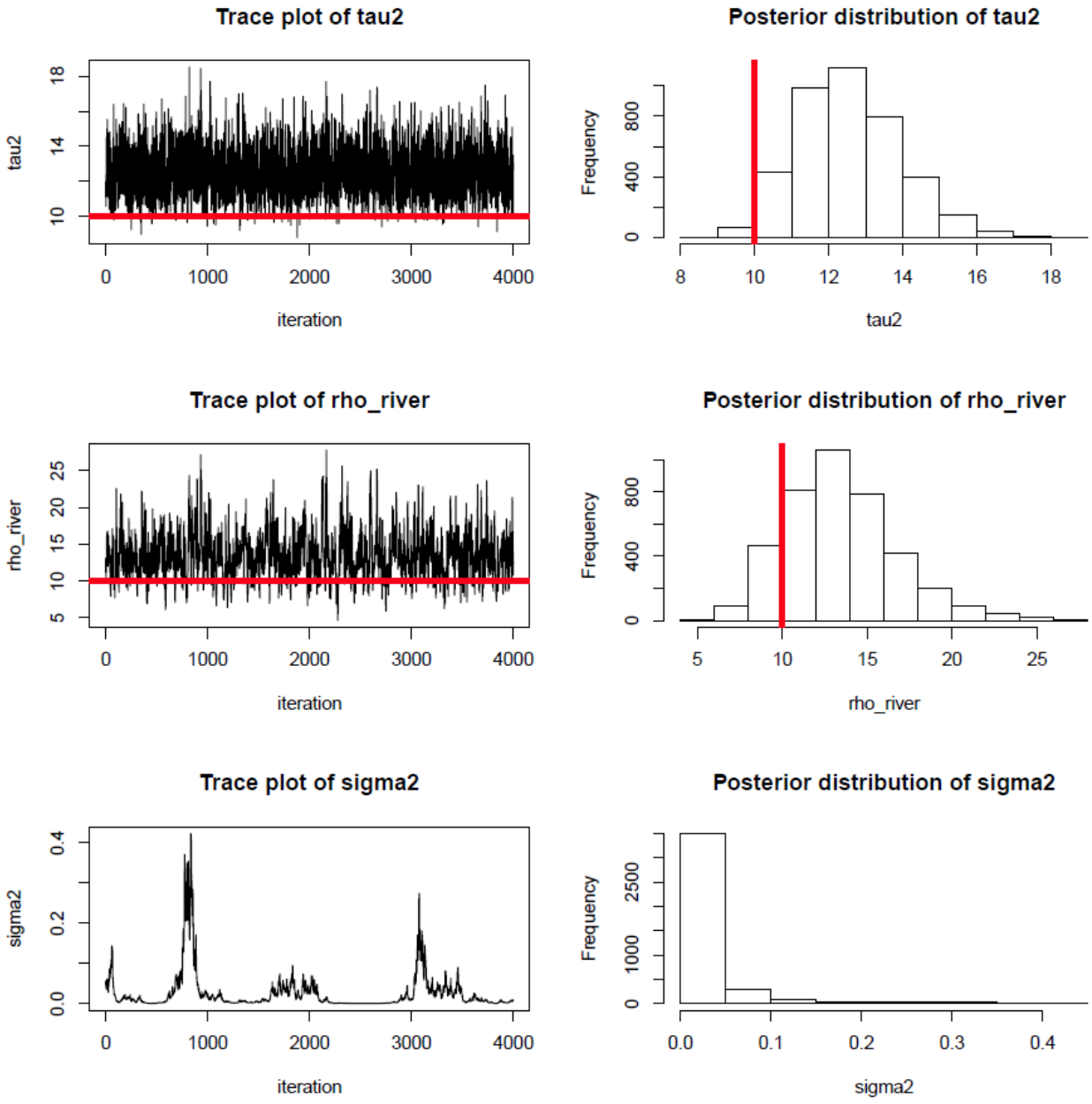
Appendix Figure 3b. Trace plots and histograms of the posterior distributions of θ, ρ_1, and ρ_2 for combined covariance function ($\mathbf{S_1}$) for simulated outcome and simulated distance matrices for s=200, t=10



Appendix Figure 3c. Trace plots and histograms of the posterior distributions of τ^2, ρ_1, and σ^2 for Euclidean Only covariance function ($\mathbf{S_2}$) for simulated outcome and simulated distance matrices



Appendix Figure 3d. Trace plots and histograms of the posterior distributions of τ^2, ρ_2, and σ^2 for River Only covariance function ($\mathbf{S}_3$) for simulated outcome and simulated distance matrices for s=200, t=10



WORKS CITED

1.) Ver Hoef, Jay M., Erin Peterson, and David Theobald. "Spatial statistical models that use flow and stream distance." *Environmental and Ecological statistics* 13.4 (2006): 449-464.

2.) Money, Eric S., Gail P. Carter, and Marc L. Serre. "Modern space/time geostatistics using river distances: data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey." *Environmental science & technology* 43.10 (2009): 3736.

3.) Money, Eric, Gail P. Carter, and Marc L. Serre. "Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in New Jersey." *water research* 43.7 (2009): 1948-1958.

4.) Peterson, Erin E., and Jay M. Ver Hoef. "A mixed-model moving-average approach to geostatistical modeling in stream networks." *Ecology* 91.3 (2010): 644-651.

5.) Peterson, Erin E., David M. Theobald, and Jay M. ver Hoef. "Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow." *Freshwater Biology* 52.2 (2007): 267-279.

6.) Peterson, Erin E., et al. "Patterns of spatial autocorrelation in stream water chemistry." *Environmental Monitoring and Assessment* 121.1 (2006): 569-594

7.) Cressie, Noel, et al. "Spatial prediction on a river network." *Journal of Agricultural, Biological, and Environmental Statistics*11.2 (2006): 127-150.

8.) Saby, N., et al. "Geostatistical assessment of Pb in soil around Paris, France." *Science of the total environment* 367.1 (2006): 212-221.

9.) Reich, Brian J., et al. "A class of covariate-dependent spatiotemporal covariance functions." *The annals of applied statistics* 5.4 (2011): 2265.

10.) Schmidt, Alexandra M., Peter Guttorp, and Anthony O'Hagan. "Considering covariates in the covariance structure of spatial processes." *Environmetrics* 22.4 (2011): 487-500.

11.) Schmidt, Alexandra M., and Anthony O'Hagan. "Bayesian inference for non-stationary spatial covariance structure via spatial deformations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.3 (2003): 743-758.

12.) Schmidt, Alexandra M., and Marco A. Rodriguez. "Modelling multivariate counts varying continuously in space." *Bayesian Statistics* 9 (2011): 611-638.

13.) Bernardo, J. M., et al. "Modelling Multivariate Counts Varying Continuously in Space."

14.) http://www.who.int/mediacentre/factsheets/fs330/en/

15.) Parashar, Umesh D., Joseph S. Bresee, and Roger I. Glass. "The global burden of diarrhoeal disease in children." *Bulletin of the World Health Organization* 81.4 (2003): 236-236.

16.) Lanata, Claudio F., et al. "Global causes of diarrheal disease mortality in children< 5 years of age: a systematic review." *PloS one* 8.9 (2013): e72788.

17.) Boschi-Pinto, Cynthia, Lana Velebit, and Kenji Shibuya. "Estimating child mortality due to diarrhoea in developing countries." *Bulletin of the World Health Organization* 86.9 (2008): 710-717.

18.) Kotloff, Karen L., et al. "Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study." *The Lancet* 382.9888 (2013): 209-222.

19.) Patz, Jonathan A., et al. "Impact of regional climate change on human health." *Nature* 438.7066 (2005): 310-317.

20.) Jones, Kate E., et al. "Global trends in emerging infectious diseases." *Nature* 451.7181 (2008): 990-993.

21.) Lipp, Erin K., et al. "The effects of seasonal variability and weather on microbial fecal pollution and enteric pathogens in a subtropical estuary." *Estuaries* 24.2 (2001): 266-276.

22.) Fink G, Gunther I, Hill K. The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986–2007. *Int J Epidemiol*. 2011;40:1196–1204.

23.) 3. Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull World Health Organ*. 2003;81:197–204.

24.) 4. Pruss A, Kay D, Fewtrell L, Bartram J. Estimating the burden of disease from water, sanitation, and hygiene at a global level. *Environ Health Perspect*. 2002;110:537–542

25.) Bhavnani, Darlene, et al. "Impact of rainfall on diarrheal disease risk associated with unimproved water and sanitation." *The American journal of tropical medicine and hygiene* 90.4 (2014): 705-711.

26.) Checkley, William, et al. "Effects of EI Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children." *The Lancet* 355.9202 (2000): 442-450.

27.) Thomas, Kate M., et al. "A role of high impact weather events in waterborne disease outbreaks in Canada, 1975–2001." *International journal of environmental health research* 16.03 (2006): 167-180.

28.) Curriero, Frank C., et al. "The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994." *American journal of public health* 91.8 (2001): 1194-1199.

29.) Hashizume, Masahiro, et al. "The effect of rainfall on the incidence of cholera in Bangladesh." *Epidemiology* 19.1 (2008): 103-110.

30.) Alexander, Kathleen A., et al. "Climate change is likely to worsen the public health threat of diarrheal disease in Botswana." *International journal of environmental research and public health* 10.4 (2013): 1202-1230.

31.) Singh, Reena B., et al. "The influence of climate variation and change on diarrheal disease in the Pacific Islands." *Environmental health perspectives* 109.2 (2001): 155.

32.) Bi, Peng, et al. "Seasonal rainfall variability, the incidence of hemorrhagic fever with renal syndrome, and prediction of the disease in low-lying areas of China." *American journal of epidemiology* 148.3 (1998): 276-281.

33.) Auld, Heather, D. MacIver, and J. Klaassen. "Heavy rainfall and waterborne disease outbreaks: the Walkerton example." *Journal of Toxicology and Environmental Health, Part A* 67.20-22 (2004): 1879-1887.

34.) Nichols, Gordon, et al. "Rainfall and outbreaks of drinking water related disease and in England and Wales." *Journal of water and health* 7.1 (2009): 1-8.

35.) Eisenberg, Joseph NS, et al. "Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural Ecuador." *Proceedings of the National Academy of Sciences* 103.51 (2006): 19460-19465.

36.) Ahn, Jaeil, et al. "A space-time point process model for analyzing and predicting case patterns of diarrheal disease in northwestern Ecuador." *Spatial and spatio-temporal epidemiology* 9 (2014): 23-35.

37.) Eckmann, J-P., and David Ruelle. "Ergodic theory of chaos and strange attractors." *Reviews of modern physics* 57.3 (1985): 617.

38.) Andrieu, Christophe, et al. "An introduction to MCMC for machine learning." *Machine learning* 50.1-2 (2003): 5-43.

39.) Chib, Siddhartha, and Edward Greenberg. "Understanding the metropolis-hastings algorithm." *The american statistician* 49.4 (1995): 327-335.

40.) Spiegelhalter, David J., et al. "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4 (2002): 583-639.

41.) Watanabe, Sumio. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research* 11.Dec (2010): 3571-3594.

42.) Gelfand, Alan E., and Sujit K. Ghosh. "Model choice: a minimum posterior predictive loss approach." *Biometrika* 85.1 (1998): 1-11.

43.) Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." *IEEE transactions on information theory* 13.1 (1967): 21-27.