Western Graduate&PostdoctoralStudies

Western University

Scholarship@Western

Electronic Thesis and Dissertation Repository

4-19-2013 12:00 AM

Joint outcome modeling using shared frailties with application to temporal streamflow data

Lihua Li, *The University of Western Ontario*

Supervisor: Charmaine Dean, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Statistics and Actuarial Sciences

© Lihua Li 2013

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Longitudinal Data Analysis and Time Series Commons

Recommended Citation

Li, Lihua, "Joint outcome modeling using shared frailties with application to temporal streamflow data" (2013). *Electronic Thesis and Dissertation Repository*. 1257.
https://ir.lib.uwo.ca/etd/1257

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

JOINT OUTCOME MODELING USING SHARED FRAILTIES WITH APPLICATION TO TEMPORAL STREAMFLOW DATA

(Thesis Format: Monograph)

by

Lihua Li

Graduate Program in Statistical and Actuarial Sciences

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
Western University
London, Ontario, Canada

©Lihua Li 2013

ABSTRACT

Recently there has been tremendous interest in the development of tools for joint analysis of longitudinal data and time-to-event data. This has gained emphasis particularly in clinical studies, where longitudinal measurements on a response may be recorded along with a time-to-event outcome. Joint analysis of multiple outcomes beyond longitudinal and survival have also been considered, for example, joint analysis of a variety of generalized linear models including continuous and count data, or continuous and binomial data. With joint analysis of multiple outcomes, the interest may be analysis of one outcome conditional on the others, or, more typically, analysis of all outcomes jointly using latent random effects to link the outcomes. In this project, we study joint-outcome models with the particular application being streamflow at two stations on the prairies. Here, streamflow at the two stations is linked via an annual random effect. Smoothers are used to flexibly account for temporal trends in the model. An important aspect is determining the amount of information required in order to estimate the link parameter which connects the two processes, and we investigate this via simulation in the context of the streamflow analysis.

Key words: Joint outcome modeling; Laplace approximation; Marginal likelihood; Random effect; Longitudinal data

ACKNOWLEDGEMENTS

First I would like to express sincerest and deepest gratitude to my supervisor, Dr. Charmaine Dean. It would be impossible to complete this work without her guidance, support and patience. What I have learnt from her is way beyond what I have presented here. I really appreciate the opportunity of being her student. Working with Dr. Dean has been a highlight of my time at Western. Dr. Amit Garg at the Kidney Clinical Research Unit at Victoria Hospital, also deserves my deepest gratitude and appreciation. Thanks for strongly supporting my academic pursuit and generously giving me the flexibility between work and study. I also really appreciate the opportunity of working with him. Also I would like to thank the Department of Statistical and Actuarial Sciences for the opportunity allowing me to balance my study and employment. Many thanks to the professors in the department for their advice and support, also for assistance whenever I turned to them for help.

Thanks also to my fellow graduate students and lab mates for their support, friendship and help. In particular, thanks to Yaofei Xiong and Chen Yang for help in programming and discussing research, thanks to Dr. Mark Wolters, Na Lei, Alisha and Elizabeth for helpful discussions.

It is such a great time of my life at Western being a student, classmate, lab mate. I will forever remember my experience here.

Finally, I would like to thank my family for their love and support. My parents made my dream of pursuing the degree in another country true; without the support and sacrifice of my husband, Jian, I don't think I would have been able to complete this work. And of course, thanks to my son, Zemu, for all the happiness he brings me.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Joint modeling is a term used to reflect a modeling approach whereby two response processes are linked via a common set of latent variables. It can be used to model two related outcomes such as a count and a binomial variable, two count outcomes, or two binomial outcomes, both of which have some shared effect; or to model a survival and recurrent event process; or, to model a survival and longitudinal variable. Under the joint modeling framework, we may, for example, use one process to inform the second, with the main emphasis being on analysis of one of the processes; alternatively, we may be interested in analyzing both outcomes jointly and using the shared latent structure to better inform both processes. Basically, the broad objective of joint modeling is to provide a framework for analyzing the systematic relationship among multiple outcomes while appropriately accounting for the correlation among these outcomes.

In practice, it is not uncommon that multiple outcomes, collected simutanenously, are measured repeatedly for each subject over time. In clinical settings, for example, longitudinal measurements on a response may be recorded along with a time-to-event outcome. When jointly modeling a survival and longitudinal variable, inference might focus on the time-to-event process while the longitudinal variable represents a time dependent covariate measured with error. A well-known illustration of this situation from HIV studies, is where measures of CD4 T-cell, a bio-marker of immunological status, are recorded longitudinally along with a time-to-event outcome, which is the progress to AIDS or death (eg. Faucet and Thomas (1996)). In another example, Fieuws et al. (2008) modeled the relationship between measures of serum creatinine and time to graft loss jointly. Wu and Carroll (1988) modeled longitudinal data and a censoring outcome simultaneously, because censoring was deemed to be informative of survival. They revealed that analyzing longitudinal data without incorporating the informative censoring (e.g. outcome-dependent drop-out) may lead to biased results. Faucet and Thomas (1996) and Wulfsohn and Tsiatis (1997) modeled time-to-event data and a longitudinal outcome including a time-varying covariate with measurement error, with a focus on differences resulting from joint analysis and a usual single-outcome survival analysis. They demonstrated that such differences may be large and advised that it is essential to model the longitudinal process and time-to-event process jointly when they are so related, since the longitudinal process may be highly informative for survival. Tsiatis and Davidian (2004) provided a comprehensive overview of the motivation and relevant literature on joint modeling of longitudinal and time-to-event data. More recently, Fitzmaurice et al. (2008) provided a thorough review of the literature, providing an update of the Tsiatis and Davidian (2004) review. In 2012, Wu et al. (2012) outlined commonly used methods, including the likelihood method and two-stage methods, and issues in joint modeling.

Dunson (2003) made popular the concept of joint modeling of several generalized linear outcomes. McCulloch (2008) quantified the construction of the correlation between mixed outcomes through theoretical and numerical calculations and also illustrated the efficiency and reduction in bias when utilizing a joint outcome approach.

The common approach of constructing a likelihood for a binary and continuous outcome by factorizing the joint distribution into a marginal component and a conditional component was considered much earlier by Krzanowski (1988), and Cox and Wermuth (1992).

Either of the outcomes may be conditioned upon, depending on the focus of the analysis. Another approach to model a binary outcome and a continuous outcome jointly is to assume that there exists an unobservable variable underlying the connection between the two outcomes, whereby the binary outcome may be assumed to occur if a latent variable exceeds a threshold. Catalano and Ryan (1992) considered this special case and indicated that the latent variable model provided a useful way to formalize the distribution of the discrete variable in the setting considered.

The factorization approach is particularly useful when additional hierarchies are included in a study, for example, cluster effects or repeated measurements. This was the case for Catalano (1997), who employed a latent variable to incorporate clustering. Fitzmaurice et al. (2008) constructed the joint density as the product of a marginal distribution for the binary outcome and conditional distribution for the continuous response given the binary outcome while accounting for clustering using a generalized estimating equation (GEE) approach.

For the joint analysis of continuous outcomes, some multivariate methods were introduced (Johnson and Wichern (2002)). Though multivariate analysis is a well developed field, when there is a clustering hierarchy involved in the outcomes, it may be useful to consider a joint modeling framework where shared random effects provide the link across outcomes.

The aim of this project is to explore the use of latent variables in a joint analysis of longitudinal data arising from two hierarchies, as an alternative to a typical multivariate analysis. Our interest is to accurately quantify the shared latent effect in such joint analyses. Our application models streamflow at two stations within the same general drainage area. Importantly, we explore here the sample size required to estimate the latent link parameter with reasonable power.

This project is organized as follows. In Chapter 2, we describe the motivating data and present an exploratory analysis of the streamflow data. In Chapter 3, we derive likelihood inference using a Laplace approximation, as this can be utilized for broad application of joint outcome analyses. Here we present the Laplace approximation for the case considered as well as employ the usual marginal density approach for the analysis of the joint streamflow outcomes. We present and discuss the results of our analysis in Chapter 4. In Chapter 5, we demonstrate, via simulation, the relationship between sample size and power to detect the link between the outcomes under different scenarios. We summarize and discuss future work in the final chapter.

Chapter 2

The Context of Joint Outcome Modeling and of the Application

We begin by discussing joint outcome modeling broadly to illustrate its utility. Because considerable work has been done in the context of survival analysis, our background discussion considers this area of application.

2.1 Background and Motivating Examples

Research on the relationship between longitudinal and time-to-event outcomes are most popular in the context of research on surrogates and biomarkers in medicine. As mentioned earlier, the most familiar example relates to HIV studies, where immunological and virological status, such as obtained by CD4 T-cell and viral RNA copy number, are collected on each patient, along with the time to progression to AIDS or death (see Wu and Ding (1999); Taylor and Wang (2002)). The objective of joint analysis is to model the mechanism underlying the evolution of the biomarkers and the event process in the presence of the treatment and to more efficiently estimate the treatment effect.

Let the event-time for individual i be denoted as T_i with censoring time C_i, so the observed event time is $U_i = min(T_i, C_i)$ and $\delta_i = I(U_i = C_i)$, the censoring indicator. The random variable $Y_i(t)$ represents the longitudinal response at time t; $X_{1i}(t)$ and $X_{2i}(t)$ are possibly time-dependent variables which affect both the longitudinal variable $Y_i(t)$ and time to event T_i. Assume that conditioning on a latent process, Y and U are independent. In this setting, Y is the longitudinally measured biomarker, CD4 cell count, and U is the time to progression to AIDs or death. The latent process could be considered as a patient's underlying health status.

Depending on the different causal paths of the relationship between the latent process and the random variables Y and U, the joint density function can be expressed as

$$\begin{aligned} f_{Y,U}(y,u) &= f_Y(y) f_{U|Y}(u|y) \\ &= f_Y(y) \int f_{U|Y,B}(u|y,b) f_B(b) db \end{aligned}$$

where b represents the latent variable, or perhaps

$$\begin{aligned} f_{Y,U}(y,u) &= \int f_{Y,U|B}(y,u|b) f_B(b) db \\ &= \int f_{Y|B}(y|b) f_{U|B}(u|b) f_B(b) db \end{aligned}$$

given b, Y and U are independent.

To be specific, we consider the joint model utilizing a linear mixed model for the longitudinal data and a proportional hazard model for the time-to-event data:

$$\begin{aligned} Y(t_{ij}) &= \mu(t_{ij}) + B_{1i}(t_{ij}) + \varepsilon_{ij} \\ \lambda(t_{ij}) &= \lambda_0(t_{ij}) \exp[\beta X_{2i} + B_{2i}(t_{ij})] \end{aligned}$$

where $\mu(t_{ij})$ is the mean function depending on covariates, X_{1i}; $\varepsilon_{ij} \sim N(0, \sigma^2)$ denotes the measurement error; X_{2i} are covariates which may or may not be the same as X_{1i} affecting survival directly, and $B_{1i}(t_{ij})$ and $B_{2i}(t_{ij})$ random effects at the individual level which influence the longitudinal process and survival process respectively.

These two processes B_{1i} and B_{2i} are the factors which induce the correlation between Y and U; the correlation structure may be, for example, a multivariate Gaussian process. However the dimension of the random effects and also the hierarchy of the models increase the complexity of computation of the correlation structure. As an alternative and also because of specific useful motivating contexts which drive these linkages, shared-parameter models have become more prominent. In shared-parameter models, $B_{2i}(t)$ is assumed a function of some or all components of $B_{1i}(t)$. For example, $B_{2i}(t) = \rho B_{1i}(t)$, or $B_{1i} = b_{1i} + b_{2i}t, B_{2i}(t) = \rho_1 b_{1i} + \rho_2 b_{2i}t$. Such shared-parameter models have been very frequently utilized in the medical field.

Another example is the common spatial factor model for joint modeling of spatial count outcomes. In public health and ecological studies, variables measured at the same spatial locations may be correlated. It is important then to consider the outcomes jointly, whereby they are characterized by a common spatial factor. In Feng and Dean (2012), an analysis of Ontario lung cancer for men and women is conducted by using common spatial factor models. The incidence and expected counts of lung cancer in 37 public health units over 1995-2002 in Ontario were considered jointly for men and women. Let y_{im} denote the lung cancer count in region i for men, $y_{im} \sim Poisson(\mu_{im})$, and let E_{im} be the expected count of lung cancer in region i for men. Correspondingly let y_{if} denote the lung cancer count in region i for women, $y_{if} \sim Poisson(\mu_{if})$ with E_{if} being the expected count of lung cancer for women. Expected counts are calculated based on age-gender-distribution and some standardized values of rates for each age-gender distribution.

The model is specified as:

$$\log(\mu_{im}) = \alpha_m + \log(E_{im}) + b_i + h_{im}$$
$$\log(\mu_{if}) = \alpha_f + \log(E_{if}) + \gamma b_i + h_{if}$$

where b_i represents a spatially correlated regional risk; h_{im} and h_{if} are independent random effects representing variation over and above the spatial effects. The common spatial structure $b = (b_1, ...b_n)^T \sim MVN(0, \sum_b)$, $h_m = (h_{1m}, ..., h_{nm})^T \sim MVN(0, \sigma^2_{hm} I)$, $h_m = (h_{1f}, ..., h_{nf})^T \sim MVN(0, \sigma^2_{hf} I)$, b, h_m and h_f are independent, and γ is termed the factor loading for the the shared spatial random effect. In the lung cancer setting, for example, γ is expected to be unity.

More complex models involving more than two outcomes and several layers in a hierarchy may also be considered. Consider, for example, the developmental toxicity study of ethylene glucol in mice conducted by the National Toxicology Program (Price et al. (1985)). In these experiments, the outcomes are litter size; the malformation status of a live fetus, a binary outcome; and birth weight. Such data were explored by Catalano and Ryan (1992), Molenberghs and Ryan (2002) and Gueorguieva and Agresti (2001) and Dunson et al. (2003) among others. In Dunson et al. (2003), each litter is treated as a cluster, and the two outcomes of fetal weight and malformation status are analyzed jointly. Let y_{ij1} denote the fetal weight for the jth pup in the litter i and y_{ij2} be the malformation status;

y^*_{ij2} denotes a normal variable underlying y_{ij2} such that $y_{ij2} = I(y^*_{ij2} > 0)$. Let s_i be the size of litter i, which can take values from 1 to T, where T is the maximum number of pups in a litter. The covariate x_i is the dose of ethylene glycol administered, $\xi_i \sim N(0,1)$ is a latent variable for the i^{th} litter, which is hence operating at the cluster level; η_{ij} is an individual level latent variable for the j^{th} pup in the i^{th} litter.

The model is specified as:

$$\begin{aligned} y_{ij1} &= \mu_1 + \alpha_1 x_i + \lambda_1 \xi_i + \gamma_1 \eta_{ij} + \varepsilon_{1ij1}, \\ y^*_{ij2} &= \mu_2 + \alpha_2 x_i + \lambda_2 \xi_i + \gamma_2 \eta_{ij} + \varepsilon_{1ij2}, \\ \Pr(s_i = j | x_i, \xi_i) &= \Phi\left(\delta_j - \beta x_i - \lambda_3 \xi_i\right) \prod_{h=1}^{j-1} \{1 - \Phi\left(\delta_h - \beta x_i - \lambda_3 \xi_i\right)\} \end{aligned}$$

Here μ_1 and μ_2 are intercepts in the weight and malformation model, while $\varepsilon_{1ij1} \sim N(0, \sigma^2)$ and $\varepsilon_{1ij2} \sim N(0,1)$ are the error terms in two models respectively; Φ is the standard normal distribution function; $\delta = (\delta_1, ..., \delta_{T-1})'$ are parameters characterizing the baseline litter size distribution among dams given $x_i = 0$.

It is not uncommon to collect several types of outcomes simultaneously in some studies such as social science surveys where questionnaires collect responses on several behaviors.

2.2 Streamflow Data and Exploratory Analysis

The example considered in this project relates to streamflow on the prairies. Streamflow is of vital importance in semi-arid regions from the perspective of both human and wildlife activities. Accurately predicting streamflow not only helps detect change due to landuse or climate variation but also facilitates government regulation. We consider two stations in the same general spatial location. Generally, streamflow on the prairies is dominated by snowmelt and spring rains; there is likely some similarity in flow at the stations, and this depends on the soil and drainage features surrounding the stations. In particular, annual effects are likely similar; these are of interest to predict return rates of flood and drought years. Our joint modeling technique is used in an exploratory way for this streamflow analysis. The joint model for streamflow we propose in this project permits handling the seasonality by using smoothers and also accounts for the correlation rooted in common random effects.

The streamflow data is obtained from Environment Canada. After exploring a few stations in the Canadian Reference Hydrologic Basin Network (RHBN), station 05ND007 and 05NF012 were determined for joint analysis in this project, considering data quality and sample size for a meaningful illustration. The data are extracted from the flood risk period of March 1 to May 31 1964-2003. Station 05ND007 is the Souris river at Sherwood; station 05NF012 is the Souris River at Westhope.

The Souris River or Mouse River (as it is alternatively known in the U.S.) is a river in central North America. The two stations are both located in North Dakota. Table 1 lists the basic geographical information about two stations. For simplicity, station 05ND007 and 05NF012 are denoted as site A and site B respectively.

Though these two stations are from the same river, there are different characteristics, for example, the presence of impoundments and dams, which will strongly affect flows, as the management of these activities modifies streamflow patterns.

Table 2.1: Attributes of two stations considered in this project

Station ID	Station Name	state	latitude	longitude	drainage area
05ND007(A)	Souris river near SHERWOOD	North Dakota	48°59′24″ N	101°57′28″ W	23100.00
05NF012(B)	Souris river near WESTHOPE	North Dakota	48°59′47″ N	100°57′29″ W	43700.00

Figure 2.1 is a time series plot of daily streamflow at the two stations. Though the temporal patterns are somewhat similar, there are differences in flow magnitude. High flows were evident between 1969 and 1979, then later on, through 1994 and 2002, another series of flood periods is observed in our study window. The daily flow of both stations reaches a maximum in 1976. Figure 2.2 is a three-dimensional plot of streamflow values by day and year, while Figure 2.3 is a heatmap of daily streamflow by day and year. The colors vary from grey, representing the lowest flow rate, to red, representing the highest flow rate.

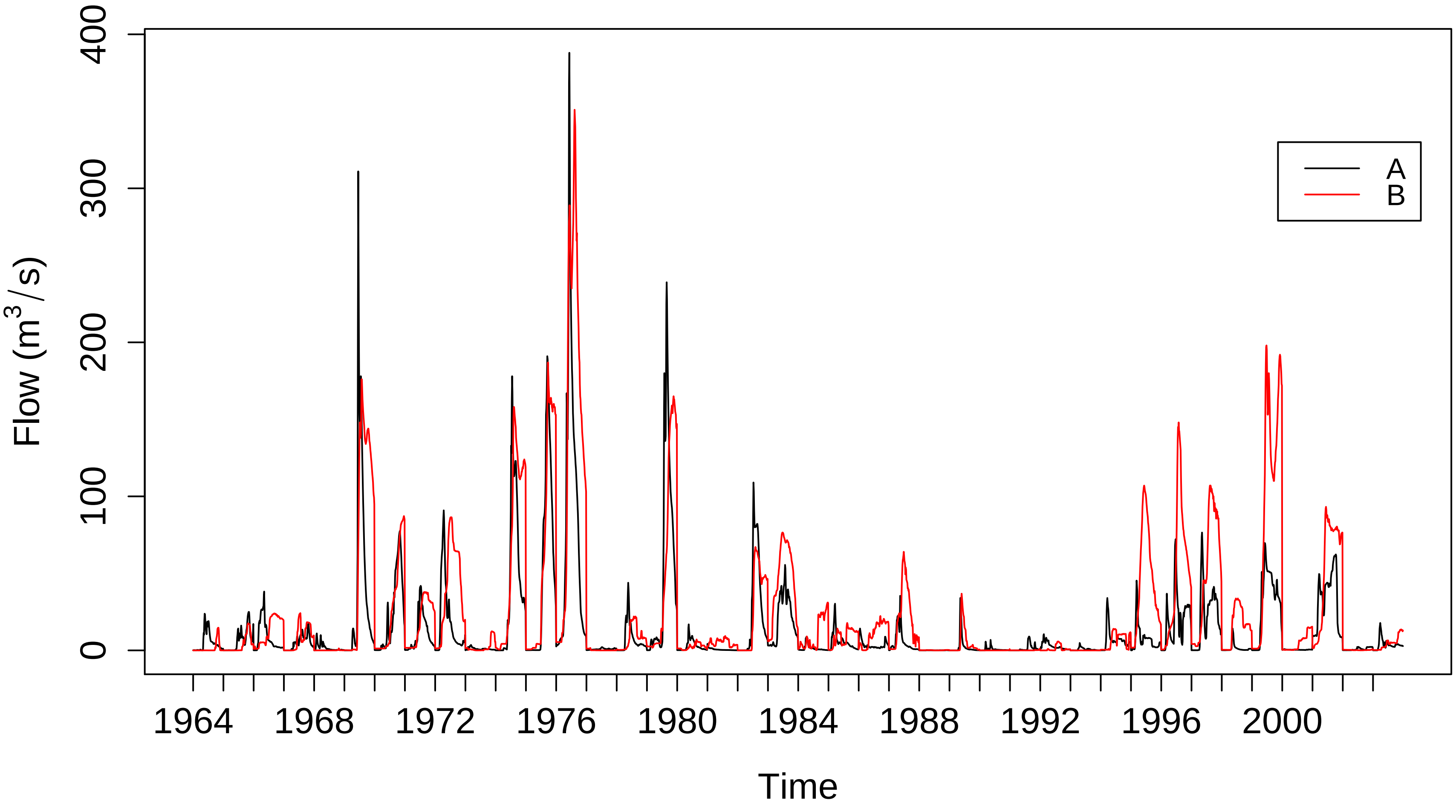


Figure 2.1: Streamflow at two stations, A and B, in March, April and May from 1964 to 2003

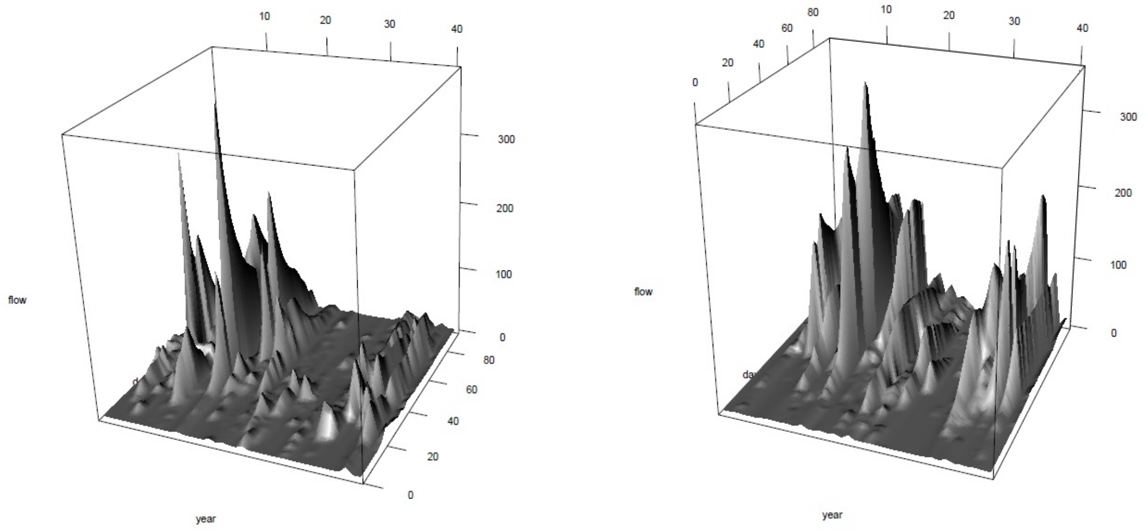


Figure 2.2: Three-dimensional plot of daily streamflow at two stations by day of year (1=March 1) and year (1=1964)

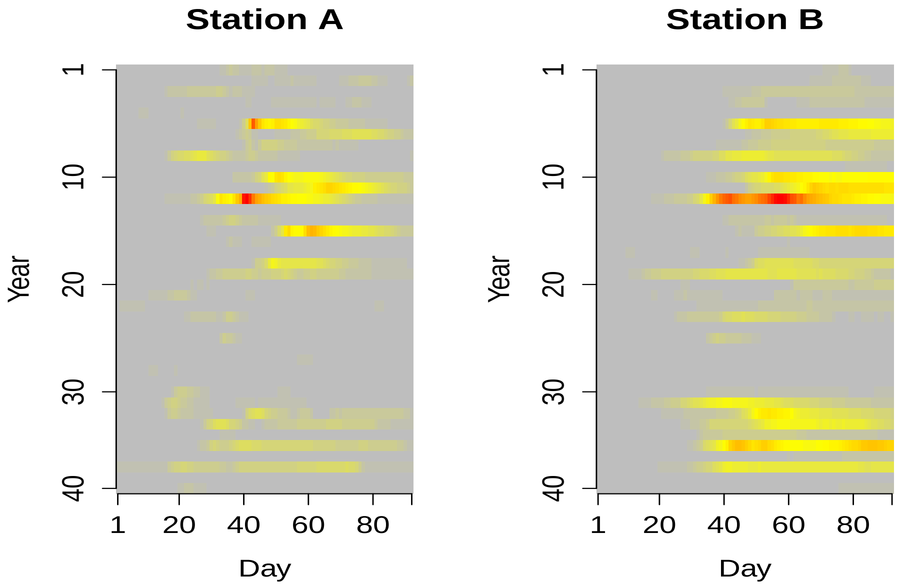


Figure 2.3: Plot of daily streamflow at two stations by day of year (1=March 1) and year (1=1964); grey represents the lowest flow values; yellow, moderate flows; orange and red represent the highest flow values

To examine the relationship between the magnitude and frequency of daily flow for the two stations, Flow Duration Curves (FDCs) are provided in Figure 2.4. Flow Duration Curves (FDCs) are hydrological curves showing the percentage of time that the flow in a stream equals or exceeds some specified values of interest over the study period. An FDC cuve can visually illustrate the variability in stream flow. Statistically, an FDC is the complement of the cumulative distribution function (cdf) of daily flow, Q. The FDC plots Q_p, the p^{th} quantile of daily flow. It is calculated as $p = 1 - P(Q \leq Q_p)$, where $P(A)$ refers to the probability of the event A. The sharp decline on the left of the flow duration curves as observed in Figure 2.4 reflects extreme events. Also we notice that station B's flow equalled or exceeded between 30 and 160 m^3/s more often than station A's flow, reflecting larger seasonal daily flow at station B versus station A.

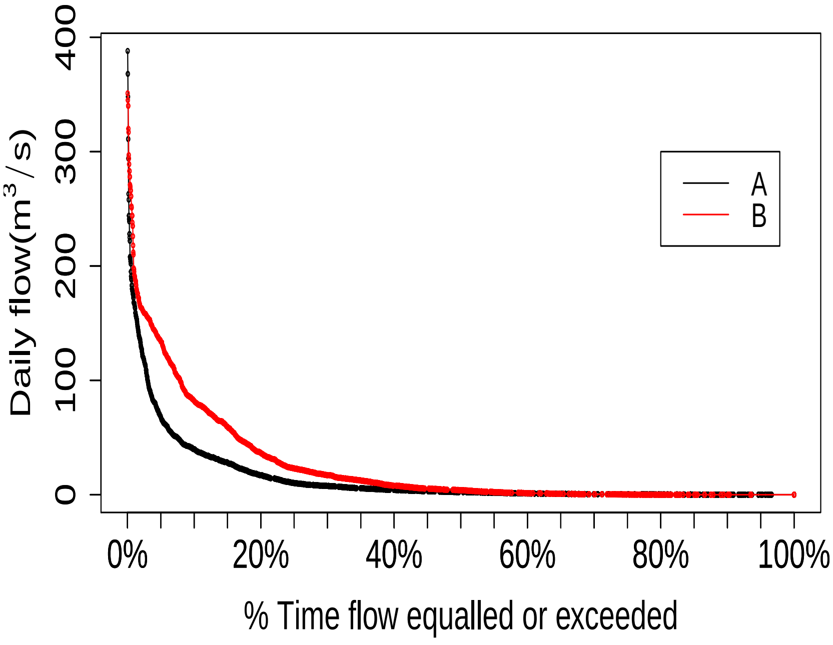


Figure 2.4: Flow duration curves for station A and B

Figure 2.5 provides box plots of streamflow over 40 years. Years with large flows tend to be the same and are 1969, 1975, 1976, 1977 and 1980.

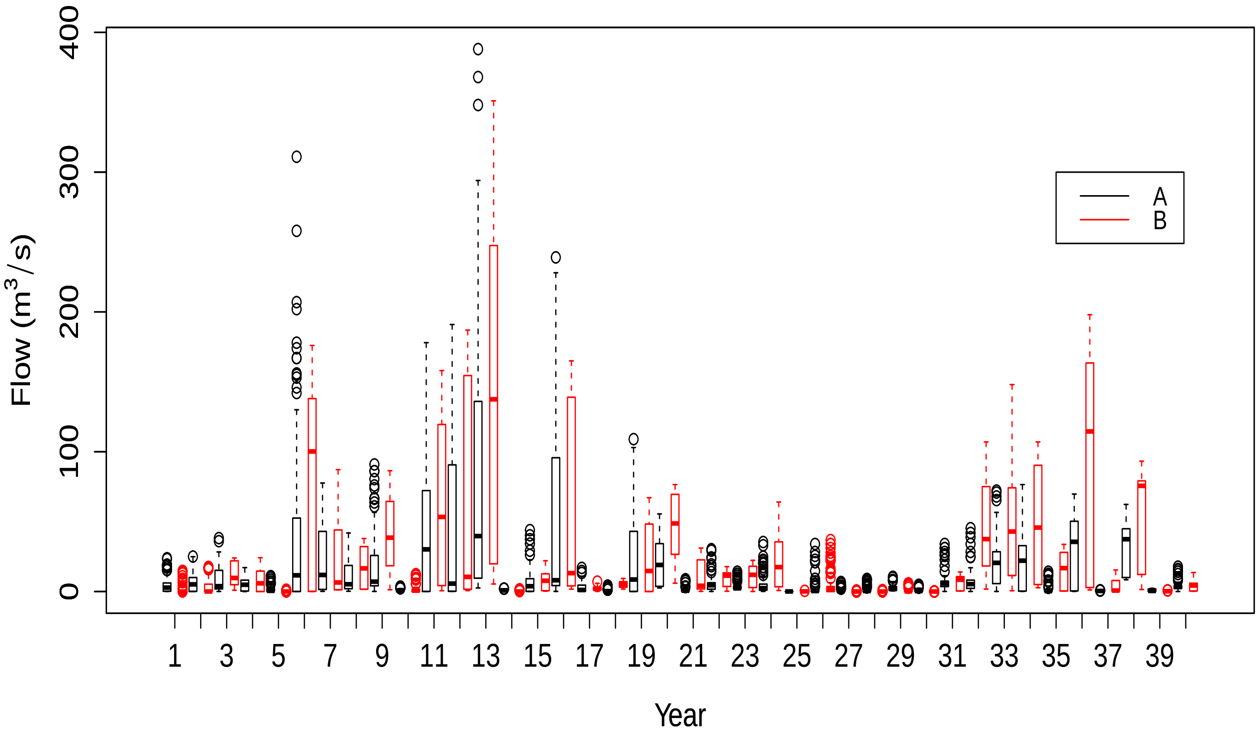
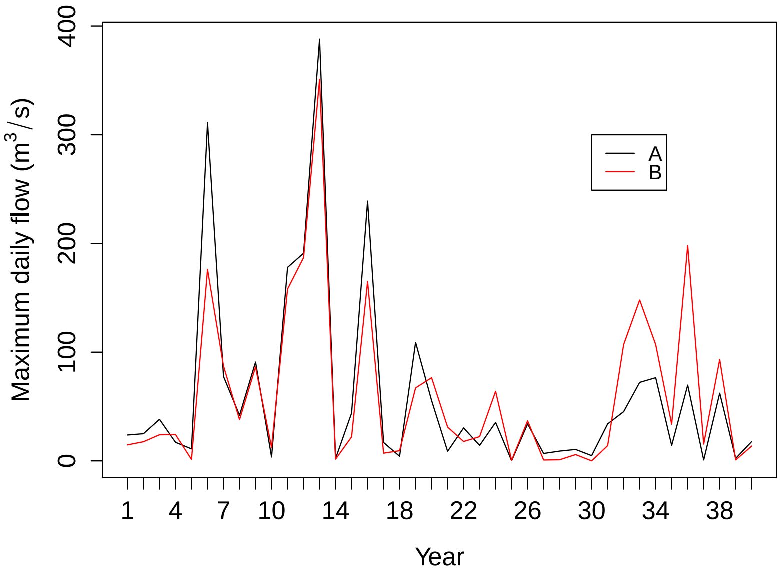


Figure 2.5: Box plot of daily streamflow at stations A and B for each of the 40 years in our study period, year 1=1964, while year 40=2003

Maximum and median daily flow over 40 years are plotted in Figure 2.6. It is seen that the two stations tend to experience similar annual maxima.



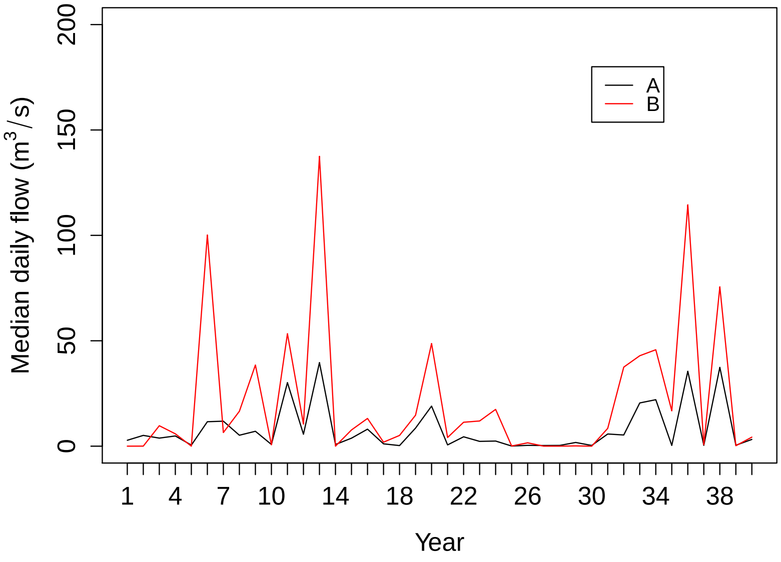


Figure 2.6: Maximum and median daily flow from 1964 (=year 1) to 2003 (=year 40) at two stations considered in this study

To investigate patterns in extremes, we present a correspondence table for the presence of extremes at the two stations. Let y_i^A and y_i^B be the mean daily flow at station A and B in the i^{th} year. Let Z_i be an indicator function defining extreme flow defined as

$$Z_i = \begin{cases} 1 & \text{if } y_i > L \\ 0 & \text{if } y_i \leq L \end{cases}$$

The cutpoint L is station specific and is selected at the 75^{th} percentile, 80^{th} percentile, 90^{th} percentile and 95^{th} percentile of the mean daily flow each year at each station. Table 2 lists the correspondence in extremes at these stations for these four different cutpoints. As seen from Table 2, the two stations exhibit some similarity in extremes.

Table 2.2: The correspondece table for different values of the quantile cutpoints

cutoff point	$Z_i^A = 1, Z_i^B = 1$	$Z_i^A = 0, Z_i^B = 0$	$Z_i^A = 1, Z_i^B = 0$	$Z_i^A = 0, Z_i^B = 1$
95%	0.025	0.925	0.025	0.025
90%	0.050	0.850	0.050	0.050
80%	0.175	0.775	0.025	0.025
75%	0.200	0.700	0.050	0.050

The average daily flow for each year at station A is plotted against that of station B in Figure 2.7. Strong correlation is evident in the annual flows for the two stations.

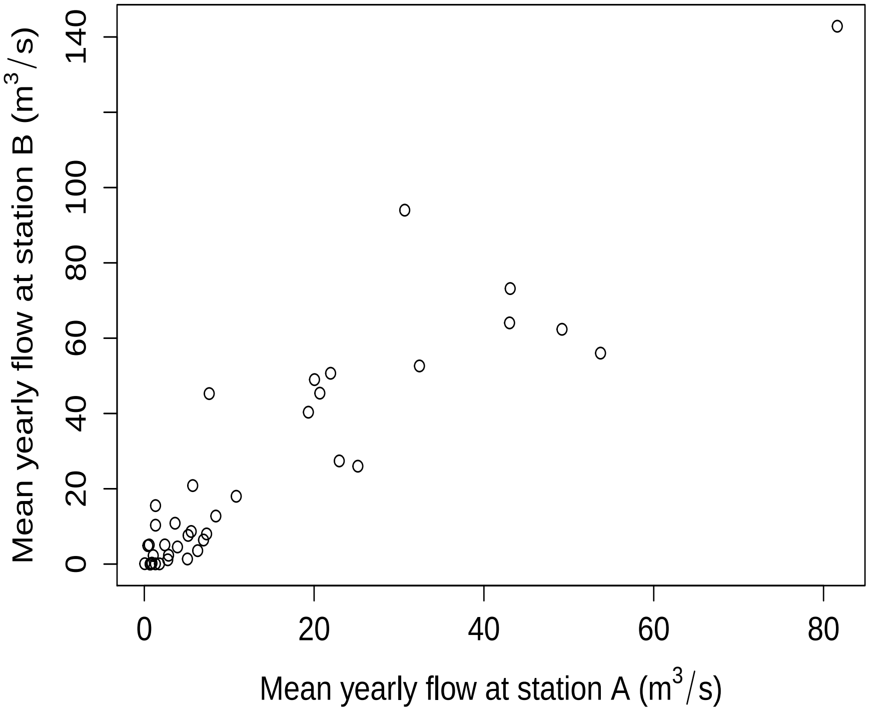


Figure 2.7: Plot of average daily flow at station A versus average daily flow at station B over 40 years

Based on all these exploratory analyses, we hypothesize that these two stations share similar annual effects and that the annual effect plays an important role in explaining the longitudinal outcomes. We will explore this in the subsequent section.

CHAPTER 3

LIKELIHOOD INFERENCE

Let y_{ij}^{A} and y_{ij}^{B} denote the stream flow at station A and B, respectively, on the j^{th} day of the i^{th} year, $i = 1, 2, ..., k$ and $j = 1, 2, ..., n$, where $y_{ij}^{A}|b_i \sim lognormal\left(\mu_{ij}^{A}, \sigma_{A}^{2}\right)$, $y_{ij}^{B}|b_i \sim lognormal\left(\mu_{ij}^{B}, \sigma_{B}^{2}\right)$, and b_i is an annual effect, $b_i \sim N\left(0, \sigma_{b}^{2}\right)$, influencing the means as:

$$\begin{aligned} \mu_{ij}^{A} &= X_{ij}^{A}\beta^{A} + b_i \\ \mu_{ij}^{B} &= X_{ij}^{B}\beta^{B} + \rho b_i. \end{aligned}$$

Here the means μ_{ij}^{A} and μ_{ij}^{B} are modeled as smoothing splines modulated by the annual effect b_i. Hence X_{ij} represents the ij^{th} row in a matrix representing spline basis functions modeling the overall seasonality term.

Smoothing splines flexibly capture the seasonal pattern existing in data using a spline function. Wood (2006) describes several smoothers including regression splines, P-splines and thin plate splines. For univariate smoothers, these smooth functions may be piecewise polynomial functions such as cubic regression splines. A cubic regression spline is a curve constructed from segments of cubic polynomials joined together so that the curve is continuous in values at both first and second derivatives. The points at which the segments are joined are termed the knots of the spline.

As conventional, knots are evenly spaced here through the range of observed x values. Given knot locations at $z^* : i = 1, 2, ..., q-2$, we use a cubic spline basis (see Wood (2006)) with basis functions expressed as: $s_1(x) = 1$, $s_2(x) = x$ and $s_{i+2} = f(x, z^*)$ for $i = 1, 2, ..., q$, where

$$f(x,z^*) = \left[(z^* - \frac{1}{2})^2 - \frac{1}{12}\right]\left[(x - \frac{1}{2})^2 - \frac{1}{12}\right]/4 - \left[(|x - z^*| - \frac{1}{2})^4 - \frac{1}{2}(|x - z| - \frac{1}{2})^2 + \frac{7}{240}\right]/24.$$

Using this cubic spline basis for the variable day of year means that X is a $kn \times q$ matrix with X_{ij}, the ij^{th}row of the basis matrix written as

$$X_{ij} = [1, x_{ij}, f(x_{ij}, z_1^*), f(x_{ij}, z_2^*), ..., f(x_{ij}, z_{q-2}^*)].$$

3.1 Laplace Approximation

For the broader context of joint outcome modeling, inference commonly proceeds via a Laplace approximation (see, for example, Vonesh et al. (2002) and Lee et al. (2006)). It has been established that the Laplace approximation works quite well in a wide variety of joint outcome models (see Skaug and Fournierb (2006); and Rue et al. (2009)). The main advantage of this approach is that it avoids complex numerical integration and is computationally efficient (see Millar (2011)). For a simple illustration, let b be a random effect, y be the response variable, and θ be all the parameters to be estimated in the model. Assume b is one dimensional, $b \in \mathbb{R}$. Let $g(b) = f(y, b; \theta)$, the probability density function of y and b. The likelihood function can be written

$$L(\boldsymbol{\theta};y)=\int_{\mathbb{R}} g(b)\,\mathrm{d}b=\int_{\mathbb{R}} e^{\log g(b)}\,\mathrm{d}b.$$

Let $\hat{b}$ denote the value which maximizes $g(b)$, and hence also $\log g(b)$. Then $\log g(b)$ can be expanded around $\hat{b}$ as below using a second-order Taylor expansion,

$$\log g(b)\approx\log g(\hat{b})-\frac{c(b-\hat{b})^2}{2},$$

where c is given by

$$c=-\frac{\partial^2 \log g(b)}{\partial b^2}|_{b=\hat{b}},$$

Thus the likelihood function can be written as:

$$L(\boldsymbol{\theta};y)\approx g(\hat{b})\int_{\mathbb{R}}\exp\left\{-\frac{c(b-\hat{b})^2}{2}\right\}\mathrm{d}b.$$

The second term on the right-hand side of the equation can be seen as the density function of a normal random variable with mean $\hat{b}$ and variance c^{-1}. Therefore, the Laplace approximation of the likelihood function can be obtained by integrating the normal density function, i.e.:

$$\begin{aligned}L(\boldsymbol{\theta};y) &\approx g(\hat{b})\int_{\mathbb{R}}\exp\left\{-\frac{(b-\hat{b})^2}{2/c}\right\}\mathrm{d}b\\ &= g(\hat{b})\sqrt{\frac{2\pi}{c}}\\ &= f(y,\hat{b};\boldsymbol{\theta})\sqrt{\frac{2\pi}{c}}.\end{aligned}$$

In the multi-dimension case where $u \in \mathbb{R}^q$, the Laplace approximation can be extended as

$$\begin{aligned} L(\theta;y) &= \int_{\mathbb{R}} g(b)db \\ &\approx g(\hat{b})(2\pi)^{\frac{q}{2}}\det(-H(\hat{b}))^{-\frac{1}{2}} \\ &= f(y,\hat{b};\theta)(2\pi)^{\frac{q}{2}}\det(-H(\hat{b}))^{-\frac{1}{2}}, \end{aligned} \tag{3.1}$$

where $\det(-H(\hat{b}))$ is the determinant of the negative of the $q \times q$ Hessian matrix of $g(b)$, given $b = \hat{b}$, which is defined as:

$$H(\hat{b}) = -\frac{\partial^2 \log g(b)}{\partial b^2}|_{b=\hat{b}}.$$

The term $f(y,b;\theta)$ can be viewed as a "complete data" likelihood element, while $f(y,\hat{b};\theta)(2\pi)^{\frac{q}{2}}\det(-H(\hat{b}))^{-\frac{1}{2}}$ (see equation 3.1) is termed the"observed data" likelihood. For the joint model proposed earlier, the likelihood function can be written:

$$L(\theta;\mathbf{y}) = \prod_{i=1}^{k} f\left(\mathbf{y}_i^A|b_i\right) f\left(\mathbf{y}_i^B|b_i\right) f\left(b_i\right),$$

where $\theta = (\beta^A, \beta^B, \rho, \sigma^A, \sigma^B, \sigma_b)$.

Then the "complete data" log-likelihood function can be expressed as:

$$l_c(\theta;\mathbf{y},b) = \sum_{i=1}^{k}\left[\log f\left(\mathbf{y}_i^A|b_i\right) + logf\left(\mathbf{y}_i^B|b_i\right) + \log f\left(b_i\right)\right].$$

Since the probability density function of $\mathbf{y}_i$ and b_i are

$$
\begin{aligned}
f\left(y_{ij}|b_i\right) &= \frac{1}{\sqrt{2\pi}\sigma y_{ij}}\exp\left\{-\frac{(\log y_{ij}-X_{ij}\beta-b_i)^2}{2\sigma^2}\right\} \\
f\left(b_i\right) &= \frac{1}{\sqrt{2\pi}\sigma_b}\exp\left\{-\frac{(b_i-0)^2}{2\sigma_b^2}\right\},
\end{aligned}
$$

the “complete data” likelihood function may be rearranged as

$$
\begin{aligned}
l_c(\theta;\mathbf{y},b) &= \sum_{i=1}^{k}\left[n\log\left(\frac{1}{\sqrt{2\pi}\sigma_A}\right)-\sum_{j=1}^{n}\log y_{ij}^{A}-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{\log y_{ij}^{A}-(X_{ij}\beta^{A}+b_i)}{\sigma_A}\right)^2\right] \\
&\quad+\sum_{i=1}^{k}\left[n\log\left(\frac{1}{\sqrt{2\pi}\sigma_B}\right)-\sum_{j=1}^{n}\log y_{ij}^{B}-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{\log y_{ij}^{B}-(X_{ij}\beta^{B}+\rho b_i)}{\sigma_B}\right)^2\right] \\
&\quad+\sum_{i=1}^{k}\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_b}\right)-\frac{1}{2}\left(\frac{b_i}{\sigma_b}\right)^2\right],
\end{aligned}
$$

The Laplace approximation to the log likelihood is

$$
\tilde{l_o}(\theta;\mathbf{y},\tilde{b}) = l_c(\theta;\mathbf{y},\tilde{b})-\frac{1}{2}\log\left|-\frac{1}{2\pi}\frac{\partial^2 l_c}{\partial b^2}\right|_{b=\tilde{b}}.
$$

Since $\partial^2 l_c/\partial b^2$ is an a $k\times k$ matrix:

$$
-\begin{pmatrix}
\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2} & 0 & \cdots & 0 \\
0 & \frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2} & \cdots & 0 \\
\cdots & \cdots & \ddots & \cdots \\
0 & 0 & \cdots & \frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2}
\end{pmatrix}_{k\times k},
$$

we have

$$\tilde{l_o}(\theta;\mathbf{y},\tilde{b}) \quad = \quad l_c(\theta;\mathbf{y},\tilde{b}) - \frac{1}{2}\log\left[\left(\frac{1}{2\pi}\right)^k\left(\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2}\right)^k\right],$$

where $\tilde{b} = b_i$, $i = 1,2,...,k$ is the solution to

$$\frac{\partial l_c^{(i)}(\theta;\mathbf{y},b_i)}{\partial b_i} = 0, i = 1,2,...,k$$

$$\Rightarrow \frac{\sum_{j=1}^n (X_{ij}\beta^A - \log y_{ij}^A) + nb_i}{\sigma_A^2} + \frac{\sum_{j=1}^n (\rho X_{ij}\beta^B - \rho\log y_{ij}^B) + n\rho^2 b_i}{\sigma_B^2} + \frac{b_i}{\sigma_b^2} = 0$$

$$\Rightarrow \tilde{b_i} = \left[\frac{\sum_{j=1}^n (X_{ij}\beta^A - \log y_{ij}^A)}{\sigma_A^2} + \frac{\sum_{j=1}^n (\rho X_{ij}\beta^B - \rho\log y_{ij}^B)}{\sigma_B^2}\right] \Big/ \left(\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2}\right).$$

Once $\tilde{b_i}$ is derived, then the "observed data" likelihood function l_o is updated using the current estimates $\tilde{b_i}$. Subsequently, estimates of the parameters $\beta^A, \beta^B, \sigma_A, \sigma_B, \rho, \sigma_b$ can be obtained by solving the equation

$$\frac{\partial \tilde{l_o}(\theta;\mathbf{y},\tilde{b})}{\partial \beta^A} = 0$$
$$\frac{\partial \tilde{l_o}(\theta;\mathbf{y},\tilde{b})}{\partial \beta^B} = 0$$
$$\frac{\partial \tilde{l_o}(\theta;\mathbf{y},\tilde{b})}{\partial \sigma_A} = 0$$
$$\frac{\partial \tilde{l_o}(\theta;\mathbf{y},\tilde{b})}{\partial \sigma_B} = 0$$
$$\frac{\partial \tilde{l_o}(\theta;\mathbf{y},\tilde{b})}{\partial \rho} = 0$$
$$\frac{\partial \tilde{l_o}(\theta;\mathbf{y},\tilde{b})}{\partial \sigma_b} = 0.$$

The above procedures is iterated to convergence, given the starting the value $\beta^{A(0)}$, $\beta^{B(0)}$, $\sigma_A^{(0)}$, $\sigma_B^{(0)}$, $\rho^{(0)}$, $\sigma_b^{(0)}$ and $b^{(0)}$. This yields the estimates $\hat{\beta^A}$, $\hat{\beta^B}$, $\hat{\sigma_A}$, $\hat{\sigma_B}$, $\hat{\rho}$ and $\hat{\sigma_b}$ of the

MLE.

The variance of the estimated parameters can be estimated using the diagonal elements of the variance-covariance matrix

$$Cov(\hat{\theta}) = -\left[\frac{\partial^2 l_0(\theta|b^*)}{\partial\theta\partial\theta^T}\right]^{-1}_{\theta=\hat{\theta}}.$$

The elements of this matrix are provided in the appendix.

3.2 Inference Using the Marginal Likelihood

Inference using a marginal likelihood approach is not typically trivial for joint modeling. However, in our situation, it is straightforward because it can be equivalently written in a linear model form.

In this situation, we rearrange the proposed model to the equivalent joint linear models:

$$\begin{aligned} z^A_{ij} &= X^A_{ij}\beta^A + b_i + \varepsilon^A_{ij} \\ z^B_{ij} &= X^B_{ij}\beta^B + \rho b_i + \varepsilon^B_{ij}, \end{aligned}$$

where $z_{ij} = \log(y_{ij})$, the logarithm of daily flow y_{ij}.

Let $\mathbf{z}^A_i$ and $\mathbf{z}^A_i$ be the logarithm of daily flow in year i for station A and station B, respectively. Based on the linear models, the marginal joint density in our case can be easily derived as:

$$
\begin{aligned}
& L(\theta;\mathbf{z}_i^A,\mathbf{z}_i^B) \\
= \; & f(z_i^A,z_i^B;\theta) \\
= \; & \int f(\mathbf{z}_i^A,\mathbf{z}_i^B,b_i;\theta)db_i \\
= \; & \int f(\mathbf{z}_i^A|b_i;\theta)f(\mathbf{z}_i^B|b_i;\theta)f(b_i)db_i \\
= \; & \int \prod_{j=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma_A}\exp\left\{-\frac{(z_{ij}^A-X\beta^A-b_i)^2}{2\sigma_A^2}\right\}\right] \\
& \left[\frac{1}{\sqrt{2\pi}\sigma_B}\exp\left\{-\frac{(z_{ij}^B-X_{ij}\beta^B-\rho b_i)^2}{2\sigma_B^2}\right\}\right]\left[\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left\{-\frac{b_i^2}{2\sigma_b^2}\right\}\right]db_i \\
= \; & \int\left(\frac{1}{\sqrt{2\pi}}\right)^{2n+1}\frac{1}{\sigma_A^n\sigma_B^n\sigma_b} \\
& \exp\left\{-\frac{\sum_{j=1}^n(z_{ij}^A-X_{ij}\beta^A-b_i)^2}{2\sigma_A^2}-\frac{\sum_{j=1}^n(z_{ij}^B-X_{ij}\beta^B-\rho b_i)^2}{2\sigma_B^2}-\frac{b_{ij}^2}{2\sigma_b^2}\right\}db_i \\
= \; & \left(\frac{1}{\sqrt{2\pi}}\right)^{2n+1}\frac{1}{\sigma_A^n\sigma_B^n\sigma_b}\exp\left\{-\frac{\sum_{j=1}^n(y_{ij}^A-X_{ij}\beta^A)^2}{2\sigma_A^2}-\frac{\sum_{j=1}^n(y_{ij}^B-X\beta^B)^2}{2\sigma_B^2}\right\} \\
& \int\exp\left\{-\left[\frac{n}{2\sigma_A^2}+\frac{n\rho^2}{2\sigma_B^2}+\frac{1}{2\sigma_b^2}\right]b_i^2\right. \\
& \left.+\left[\frac{\sum_{j=1}^n(z_{ij}^A-X_{ij}\beta^A)}{\sigma_A^2}+\frac{\sum_{j=1}^n\rho(z_{ij}^B-X_{ij}\beta^B)}{\sigma_B^2}\right]b_i\right\}db_i.
\end{aligned}
$$

Let

$$
\begin{aligned}
\sigma^{*2} &= \left(\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2}\right)^{-1} \\
\mu_i^* &= \left[\frac{\sum_{j=1}^n(z_{ij}^A-X_{ij}\beta^A)}{\sigma_A^2}+\frac{\sum_{j=1}^n\rho(z_{ij}^B-X_{ij}\beta^B)}{\sigma_B^2}\right]\sigma^{*2}.
\end{aligned}
$$

Therefore,

$$\begin{aligned}
& L(\theta;\mathbf{z}_i^A,\mathbf{z}_i^B) \\
= \; & (\frac{1}{\sqrt{2\pi}})^{2n+1}\frac{1}{\sigma_A^n\sigma_B^n\sigma_b}\exp\left\{-\frac{\sum_{j=1}^n(z_{ij}^A-X_{ij}\beta^A)^2}{2\sigma_A^{*2}}-\frac{\sum_{j=1}^n(z_{ij}^B-X_{ij}\beta^B)^2}{2\sigma_B^2}+\frac{\mu_i^{*2}}{2\sigma^{*2}}\right\} \\
& \sqrt{2\pi}\sigma^*\int\frac{1}{\sqrt{2\pi}\sigma^*}\exp\left\{-\frac{(b_i-\mu_i^*)^2}{2\sigma^{*2}}\right\}db_i \\
= \; & \left(\frac{1}{\sqrt{2\pi}}\right)^{2n}\frac{\sigma^*}{\sigma_A^n\sigma_B^n\sigma_b}\exp\left\{-\frac{\sum_{j=1}^n(z_{ij}^A-X_{ij}\beta^A)^2}{2\sigma_A^2}-\frac{\sum_{j=1}^n(z_{ij}^B-X_{ij}\beta^B)^2}{2\sigma_B^2}+\frac{\mu_i^{*2}}{2\sigma^{*2}}\right\}.
\end{aligned}$$

The log likelihood function over all observations $\mathbf{z}_i^A$ and $\mathbf{z}_i^B$, $i=1,2,...,k$ becomes:

$$\begin{aligned}
& \sum_{i=1}^k\left[2n\log\sqrt{2\pi}+\log\sigma^*-n\log\sigma_A-n\log\sigma_B-\log\sigma_b\right] \\
& -\sum_{i=1}^k\left[\frac{\sum_{j=1}^n(z_{ij}^A-X_{ij}\beta^A)^2}{2\sigma_A^2}+\frac{\sum_{j=1}^n(z_{ij}^B-X_{ij}\beta^B)^2}{2\sigma_B^2}+\frac{\mu_i^{*2}}{2\sigma^{*2}}\right].
\end{aligned}$$

CHAPTER 4

STREAMFLOW ANALYSIS

We apply the shared parameter model to the analysis of the streamflow data. Table 2 provides estimates of the parameters from fitting the joint model. The estimates of σ_A and σ_B are very close, indicating hydrological connection in variation of flows for the two stations. We note that the link parameter ρ is significant. Though not shown here, we note that bootstrap estimates of the parameters show the normal approximation to be valid.

Table 4.1: Estimates and standard errors of parameters in the joint model

Parameter	Laplace approximation		Marginal Density	
	Estimate	SE	Estimate	SE
σ_A	1.831	0.021	1.835	0.022
σ_B	1.696	0.020	1.690	0.021
ρ	2.000	0.021	2.190	0.060
σ_b	1.379	0.155	1.261	0.144

Figure 4.1 displays the posterior estimated value of log flow against day of year, overlaid on the mean observed log daily value averaged over 40 years. The seasonal smoothers use 8 interior knots and seem to capture the seasonality for both stations reasonably well. As shown in Figure 4.1, the mean flow for station A seems to peak around the 50^{th} day in the study window while mean flow for station B peaks at about day 80. As well, mean flows have somewhat different shape with flows for station A tending to remain high over the period. Figure 4.2 plots the posterior estimated values and observed values of log daily flow averaged over 5-day windows and the 40-year period; this demonstrates very good correspondence between observed and expected at the 5-day level of grouping.

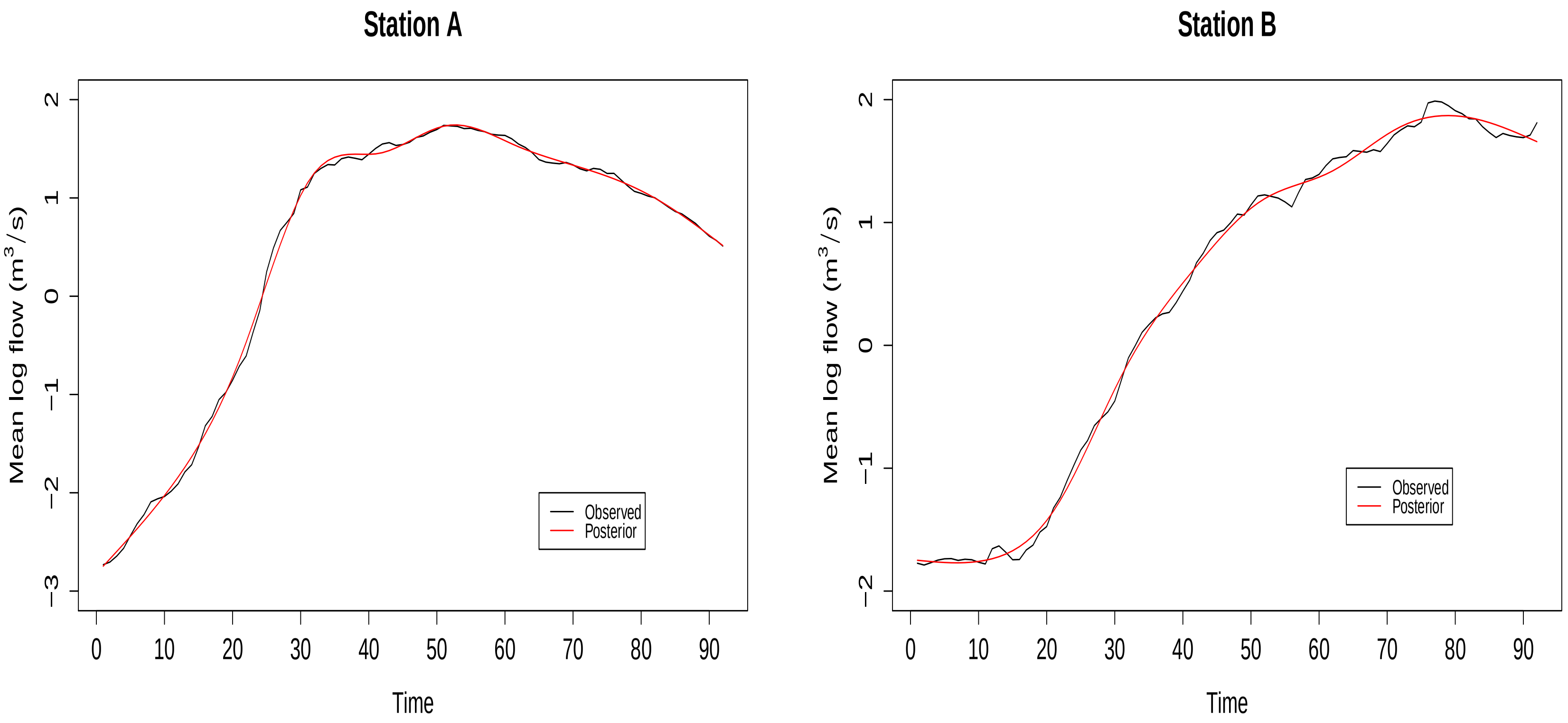


Figure 4.1: Posterior estimated values and mean observed values of log flow averaged over 40 years with day 1=March 1

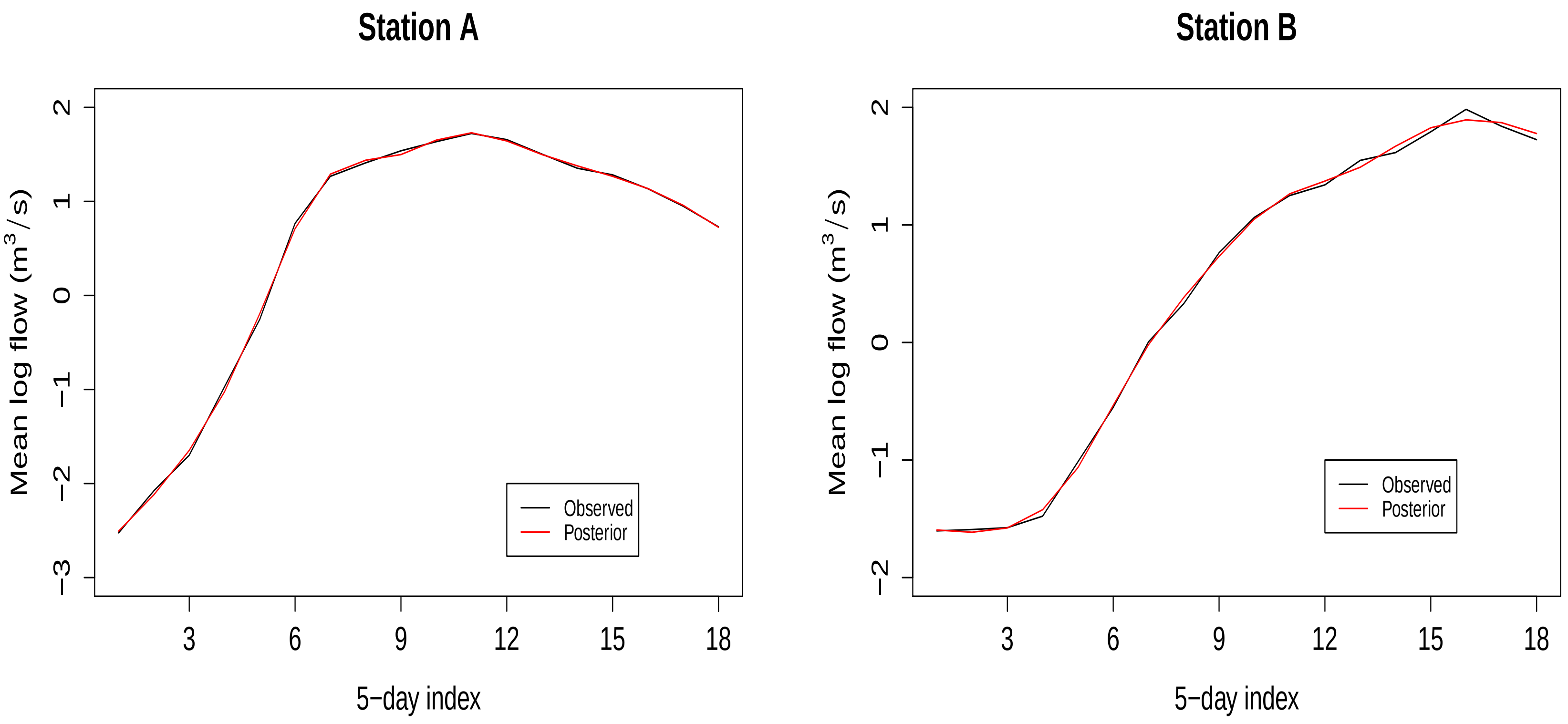


Figure 4.2: Mean posterior estimated values and mean observed values of log flow averaged over 40 years, grouped by 5-day window

The fitted model is also illustrated through the plot of posterior estimates of log daily flow over 40 years. Note that some very low flows are not well captured by the model. As well note the lower values of flow for both stations from 1988 to 1993 identified in the Figure 4.3.

To assess the goodness of fit for the specified model, posterior estimated values vs observed values of log flow over 5-day windows averaged over 40 years are examined in Figure 4.4. The figure shows fair correspondence between observed and fitted values.

Figure 4.5 presents mean and observed annual log flow values, again, illustrating the drought period in 1988 to 1993.

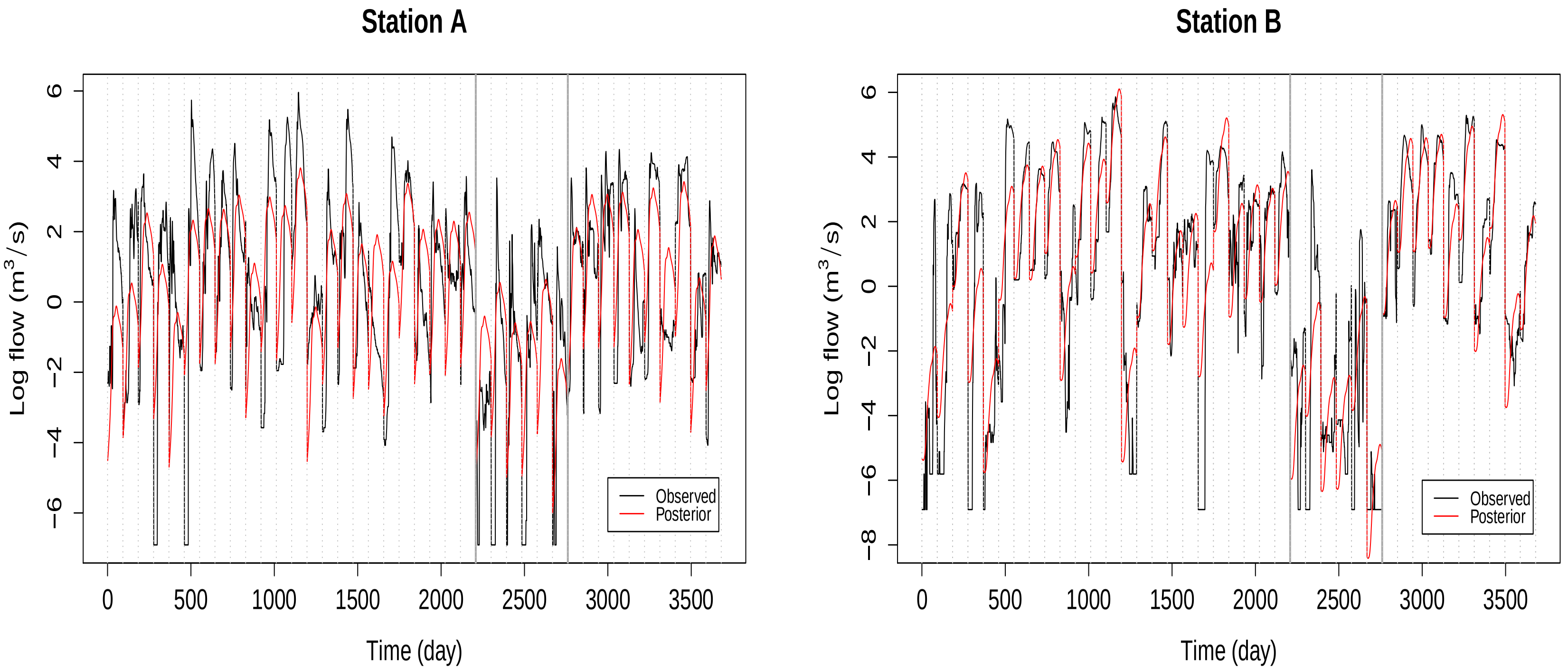


Figure 4.3: Posterior estimated values and observed values of log daily flow by day over 40 years for station A and B

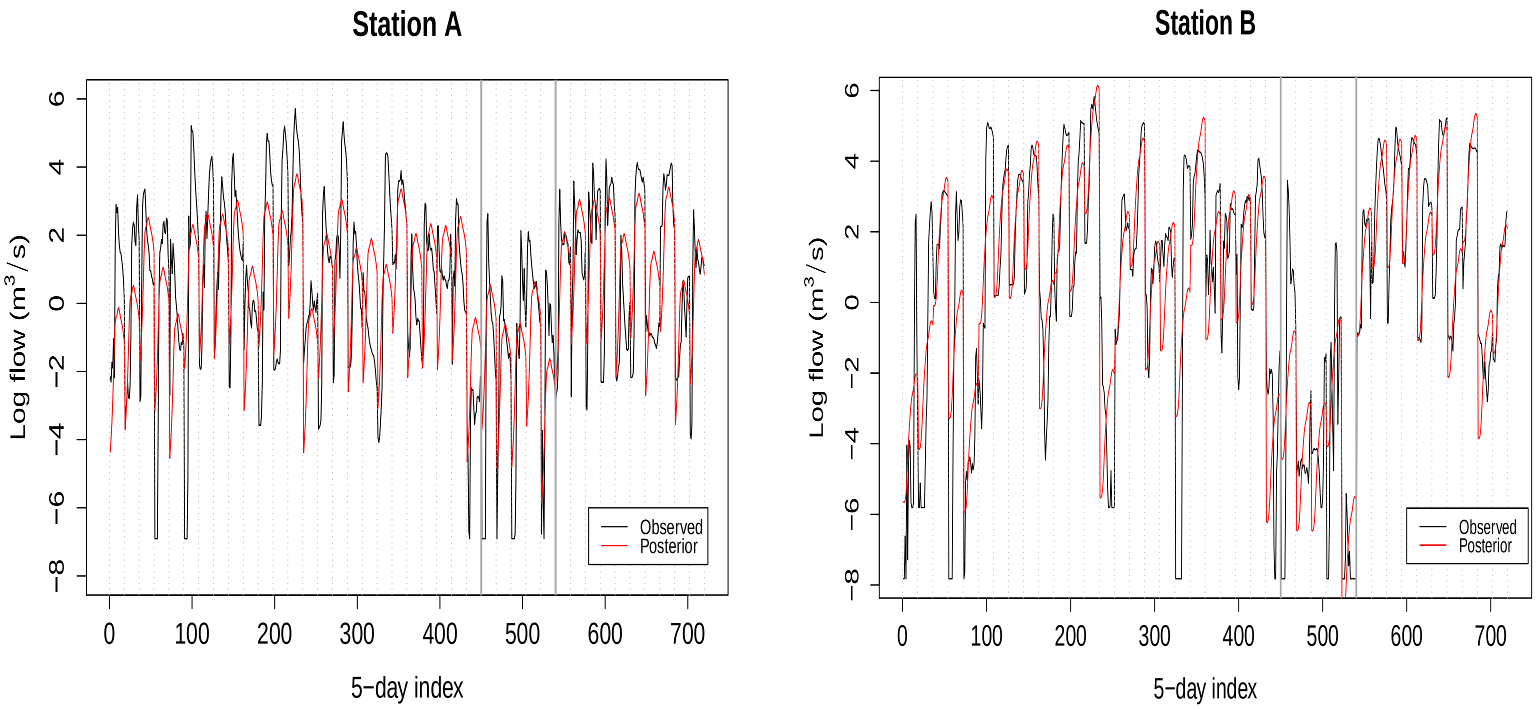


Figure 4.4: Posterior estimated values and observed values of log flow over 40 years, grouped by 5-day window

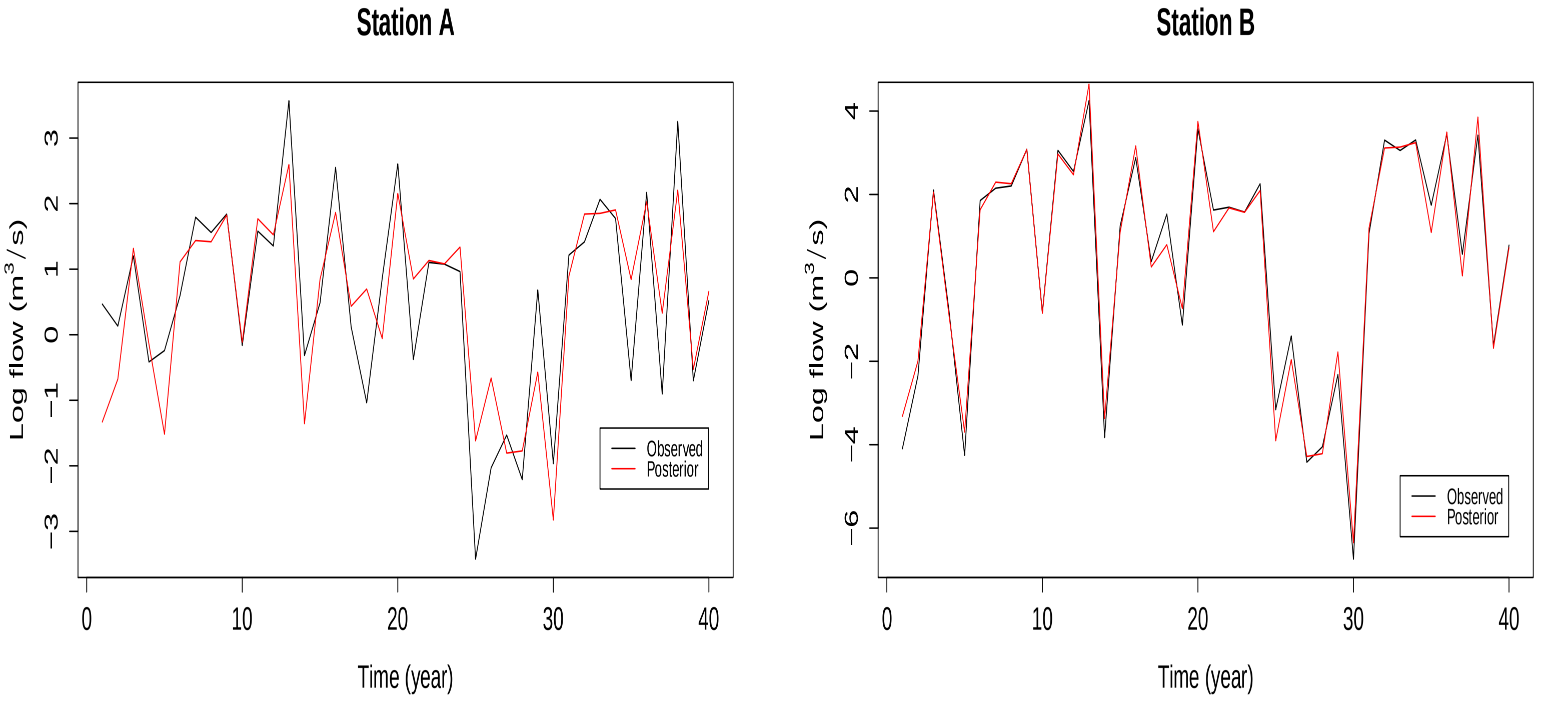


Figure 4.5: Posterior estimated values and observed values of annual log flow over 40 years

Figure 4.6 plots posterior estimates of b_i and their 95% confidence intervals. We observe the largest annual effect in 1976 (year 13), and low values in 1988-1993 (year 25-30); this is consistent with observations in Figure 4.5, and as well in the exploratory analysis.

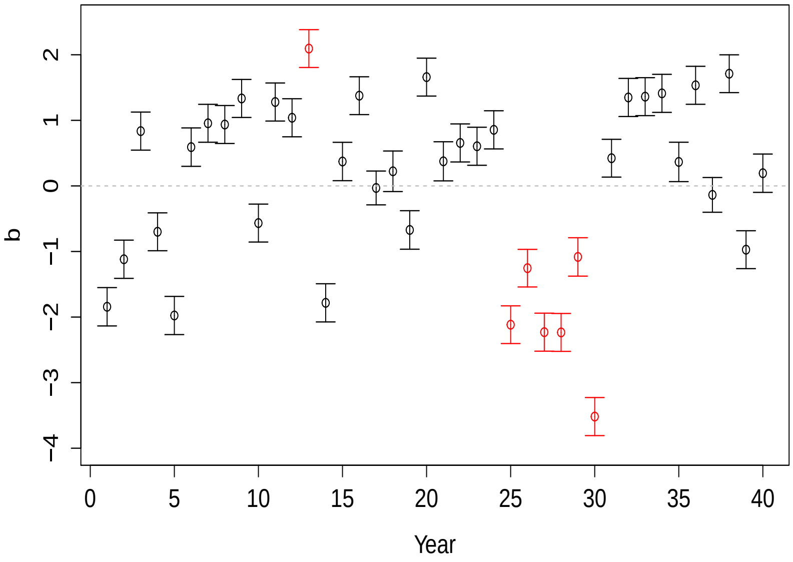


Figure 4.6: Posterior estimates of the annual effect and their 95% confidence intervals. Some extreme values are highlighted in red

To further assess goodness of fit for the proposed model, the residuals for both models are examined. This is mainly illustrated by the density plot of residuals. As we see from Figure 4.6, the two plots suggests that the normality assumption is reasonably satisfied. The density plot for the first station shows slight left skewness, which is not surprising. As we pointed out previously, this is mainly because of some low values for that station.

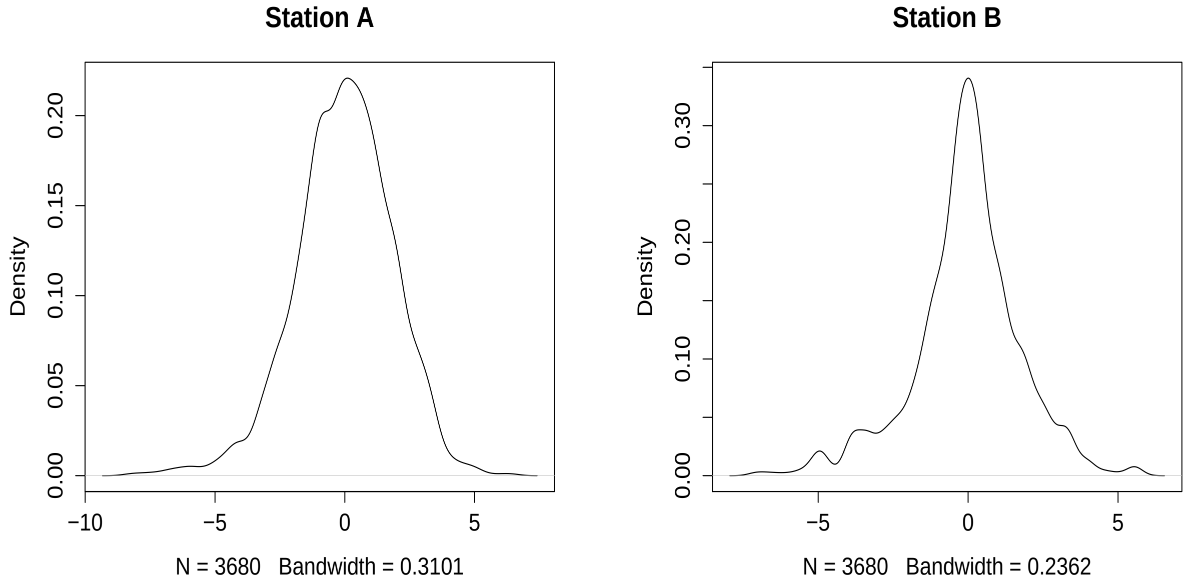


Figure 4.7: Density plot of residuals for station A and station B

The analysis here is based on a cubic spline with eight interior knots. In order to capture the seasonality adequately while avoiding over-fitting, a small sensitivity analysis is conducted by varying the number of knots utilized. We explore cubic spline smoothing with 5, 6, 7, 8, 9 and 10 knots. The residual sum of squares (RSS), calculated as $\sum_{i=1}^{k}\sum_{j=1}^{n}(\log y_{ij} - X\hat{\beta} - \hat{b}_i)$, here $i = 40$ and $j = 92$, is listed below. Eight knots seem to provide a reasonable fit for both stations; changes in SSE are very small with a larger number of knots; even seven knots may be sufficient.

Table 4.2: Residual sum of squares by varying the number of knots for both stations

number of knots	SSR for station A	SSR for station B
5	14597.975	13688.077
6	14595.357	13687.495
7	14575.561	13681.641
8	14569.024	13680.308
9	14572.067	13679.824
10	14573.103	13679.923

CHAPTER 5

POWER AND SAMPLE SIZE REQUIRED FOR TESTING THE LINK PARAMETER

It is useful in practice to investigate what sample size should be required in order to achieve a reasonable power for testing a hypothesis. Therefore, it is important to routinely evaluate the power of testing procedures. In joint modeling, the emphasis is typically on the shared parameter and assessment of linkage across outcomes. There has been little discussion on whether and in what situations that parameter may be estimated well. This section draws attention to this concern by considering the problem for the streamflow analysis context.

To examine the strength of the evidence of the shared parameter between two sites, or in general, between two outcomes modeled as in this study, a simulation study is conducted.

We are concerned with a test of

$H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$

Under H_0, there is no shared common effect for the two outcomes. In other situations of joint modeling, testing $H_0 : \rho = 1$ versus $H_1 : \rho \neq 1$ may be of interest.

Keeping the length of the series within annual clusters the same as in the streamflow analysis and using the estimated mean values as derived in the analysis, we generate the longitudinal data for our study using a series of values of σ_A, σ_B, σ_b and ρ. To be specific, set $r = 1$, then the iterative procedure is described as follows:

1. At the rth replicate, generate $b^{(r)} = \left(b_1^{(r)}, b_2^{(r)}, ..., b_k^{(r)}\right) \sim N(0, \sigma_b)$, $y_{ij}^{A(r)} \sim N\left(X\beta^A + b_i, \sigma_A^2\right)$ and $y_{ij}^{B(r)} \sim N\left(X\beta^B + \rho b_i, \sigma_B^2\right)$,
2. Fit the joint model using $y_{ij}^{A(r)}$ and $y_{ij}^{B(r)}$ to obtain the estimate of $\hat{\rho}^{(r)}$ and standard deviance of $\hat{\rho}^{(r)}$, then construct the 95% confidence interval for ρ, which is denoted as $\hat{\rho}_L^{(r)}$ and $\hat{\rho}_U^{(r)}$
3. Set r to $r+1$. If $r \leq R$, return to step 1; else stop.

Here R is set to 500. Then the power can be calculated as

$$1 - \beta = 1 - \sum_{r=1}^{R} I\left(\hat{\rho}_L^{(r)} < 0 < \hat{\rho}_U^{(r)}\right)/R$$

Where $I(A)$ is the indicator for event A. Our study design considers 3 scenarios:

- S1:$\sigma_A = \sigma_B = 1.8$, $\sigma_b = 1.3$, values close to the estimates obtained in the streamflow analysis; k, the number of years of data, takes values 5, 10, 20 or 40; $\rho = 0$, 0.05, 0.20, 0.6, 0.8, 1;
- S2: $\sigma_A = \sigma_B = 1.8$, $\sigma_b = 0.5$, $k = 20$, 40, 60 or 80; $\rho = 0$, 0.05, 0.20, 0.6, 0.8, 1;
- S3: $\sigma_A = \sigma_B = 1$, 2, 4, 6, 8, 10, $\sigma_b = 1$; $\rho = 0.2$, 0.5, 1

Figure 5.1 and 5.2 provide power curves for S1 and S2 respectively. Under S1, with 40 years of data as in our study, power is reasonably high. With 5 years of data, there is lower power to detect smaller values of ρ. Under S2, with the annual effect having a less dominant effect, about 80 years of data is required to achieve reasonable power for values of ρ greater than 0.2.

This is also seen in Figure 5.3, which compares power curves for the same number of years of data but with different values of σ_b (S1 vs S2).

Figure 5.4 provides power curves under S3. We consider the power as a function of σ_A by varying $\sigma_A (= \sigma_B)$ from 1, 2, 4,... to 10 and keeping σ_b equal to 1. Three different values of ρ are considered. As shown in the Figure 5.4, as the dominance of $\sigma_A (= \sigma_B)$ decreases, the power increases for fixed ρ. Of course, this is modulated by the size of ρ.

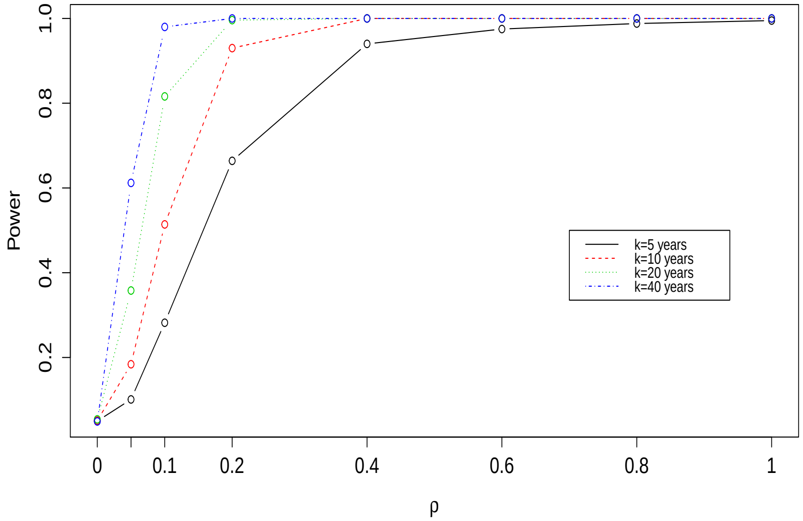


Figure 5.1: Power curve for testing the shared common effect over varying values of the link parameter from 0 to 1, for different numbers of years of data, under S1

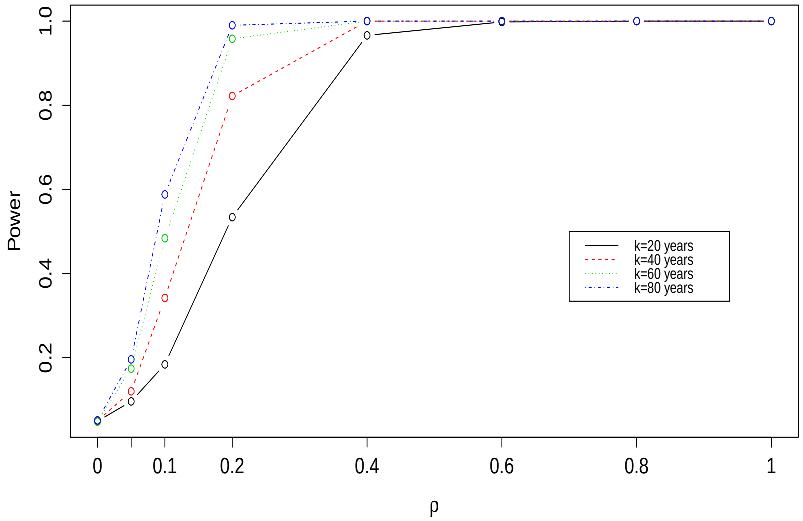


Figure 5.2: Power curve for testing shared common effect over varying values of the link parameter from 0 to 1, for different numbers of years of data, under S2

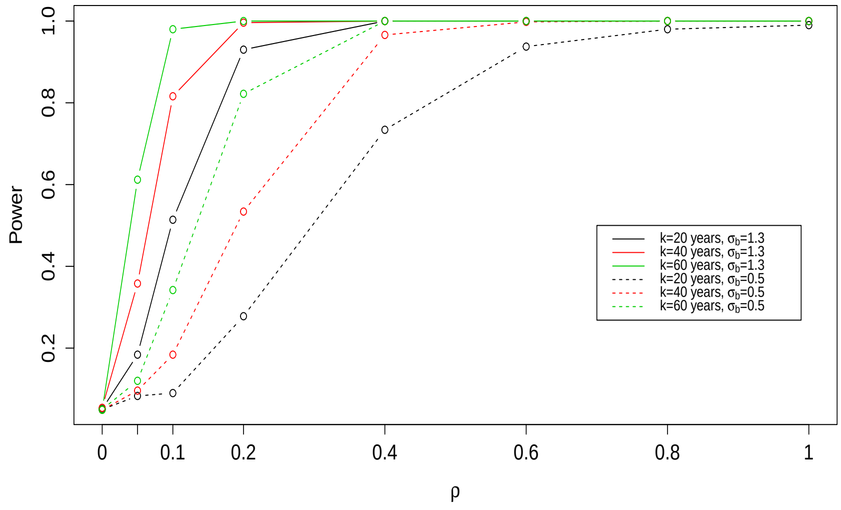


Figure 5.3: Power curve for testing the shared common effect over varying values of the link parameter from 0 to 1, for different number of years of data, under S1 and S2

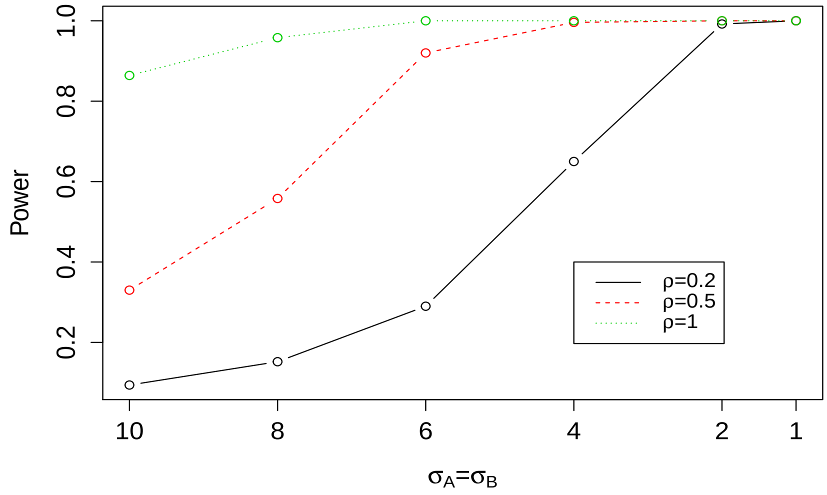


Figure 5.4: Power curve for testing the shared common effect over varying values of σ_A from 1 to 10 , for 3 different values of ρ, under S3

CHAPTER 6

FUTURE WORK

In this project, we developed a joint model for the streamflow data with a cubic spline smoother for the temporal trend and with an annual shared random effect across the outcomes. This joint outcome modeling approach provided a fair description of the pattern of streamflow at two stations. However, there are several extensions required to consider streamflow well, including smoother selection, incorporation of additional random effects, handling many zeros and accounting for auto-correlation.

6.1 Penalized Spline Smoothers

To assess the fit of the smoothers we performed sensitivity analysis by altering the number of knots in our cubic spline. Alternatively, we may fit a model with a large number of evenly spaced knots and control for overfitting by including a penalty term in the optimization. By employing penalized spline smoothing, the likelihood function criteria becomes $l_p(\beta) = l(\beta) - 1/2\sum_j \lambda_j \beta^T S_j \beta$, where S is a matrix of known coefficients and λ is a smoothing parameter which controls the trade-off between model fitting and model smoothness (see Wood (2006)). By doing this, the problem of knot selection is reduced to estimating the smoothing parameter.

Multi-dimensional smoothers, for example, tensor product splines, which produce knot free bases for multiple predictors and are scale invariant, can also be considered. In the future, we may also extend our model by adding a smoother of year.

Dealing with uncertainty in the basis function in joint models was considered by Bigelow and Dunson (2009). Here, in a Bayesian framework, the number and location of knots were determined by averaging models of the same class of multivariate linear splines but with different numbers and locations of knots.

6.2 Station-Specific Random Effects

In our analysis, the models for streamflow data from two stations were joined via the use of one shared random effect. However, the use of an additional station-specific random effect can be explored to account for the variability arising from an annual effect at station B which is not currently explained by the shared common random effect. In this case, the extended model can be written as:

$$\begin{aligned} \mu_{ij}^A &= X_{ij}^A \beta^A + b_i \\ \mu_{ij}^B &= X_{ij}^B \beta^B + \rho b_i + a_i, \end{aligned}$$

where $a_i \sim N(0, \sigma_a)$ is the additional annual random effect for station B. The "complete data" likelihood function can then be written as:

$$L_c(\theta; \mathbf{y}, a, b) = \prod_{i=1}^{k} \left[f(\mathbf{y}_i^A | a_i, b_i) f(\mathbf{y}_i^B | a_i, b_i) f(a_i) f(b_i) \right]$$

and the "complete data" log-likelihood function can be written as:

$$\begin{aligned}
l_c(\theta;\mathbf{y},a,b) &= \sum_{i=1}^{k}\left[n\log\left(\frac{1}{\sqrt{2\pi}\sigma_A y_{ij}}\right)-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{\log y_{ij}^A-(X_{ij}\beta^A+b_i)}{\sigma_A}\right)^2\right]\\
&+\sum_{i=1}^{k}\left[n\log\left(\frac{1}{\sqrt{2\pi}\sigma_B y_{ij}}\right)-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{\log y_{ij}^B-(X_{ij}\beta^B+\rho b_i+a_i)}{\sigma_B}\right)^2\right]\\
&+\sum_{i=1}^{k}\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_b}\right)-\frac{1}{2}\left(\frac{b_i}{\sigma_b}\right)^2\right]\\
&+\sum_{i=1}^{k}\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_a}\right)-\frac{1}{2}\left(\frac{a_i}{\sigma_a}\right)^2\right].
\end{aligned}$$

Using the Laplace approximation as derived in Section 3, the "observed data" likelihood function l_o can be derived as:

$$\tilde{l_o}(\theta;\mathbf{y},\tilde{a},\tilde{b}) = l_c(\theta;y,\tilde{a},\tilde{b})-\frac{1}{2}\log\left|-\frac{1}{2\pi}\frac{\partial^2 l_c}{\partial(a,b)^2}\right|_{a=\tilde{a},b=\tilde{b}},$$

where $\theta=(\beta^A,\beta^B,\rho,\sigma_A,\sigma_B,\sigma_a,\sigma_b)$; here $\partial^2 l_c/\partial(a,b)^2$ has elements:

$$\begin{aligned}
\frac{\partial^2 l_c}{\partial b_i^2} &= -\frac{n}{\sigma_A^2}-\frac{n\rho^2}{\sigma_B^2}-\frac{1}{\sigma_b^2}\\
\frac{\partial^2 l_c}{\partial a_i^2} &= -\frac{n}{\sigma_B^2}-\frac{1}{\sigma_a^2}\\
\frac{\partial^2 l_c}{\partial b_i\partial a_i} &= \frac{\partial^2 l_c}{\partial a_i\partial b_i}=-\frac{n\rho}{\sigma_B^2}\\
\frac{\partial^2 l_c}{\partial b_i\partial a_j} &= 0\text{ (where } i\neq j).
\end{aligned}$$

Thus, $\partial^2 l_c/\partial(a,b)^2$ becomes a $2k \times 2k$ matrix:

$$-\left(\begin{array}{cccccccc}
\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2} & 0 & \dots & 0 & \frac{n\rho}{\sigma_B^2} & 0 & \dots & 0 \\
0 & \frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2} & \dots & 0 & 0 & \frac{n\rho}{\sigma_B^2} & \dots & 0 \\
 & \dots & & & & \dots & & \\
0 & 0 & \dots & \frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2} & 0 & 0 & \dots & \frac{n\rho}{\sigma_B^2} \\
\frac{n\rho}{\sigma_B^2} & 0 & \dots & 0 & \frac{n}{\sigma_B^2}+\frac{1}{\sigma_a^2} & 0 & \dots & 0 \\
0 & \frac{n\rho}{\sigma_B^2} & \dots & 0 & 0 & \frac{n}{\sigma_B^2}+\frac{1}{\sigma_a^2} & \dots & 0 \\
 & \dots & & & & \dots & & \\
0 & 0 & \dots & \frac{n\rho}{\sigma_B^2} & 0 & 0 & \dots & \frac{n}{\sigma_B^2}+\frac{1}{\sigma_a^2}
\end{array}\right)_{2k\times 2k}$$

The "observed data" log likelihood function is :

$$\begin{aligned}
\tilde{l}_o(\theta;\mathbf{y},\tilde{a},\tilde{b}) &= l_c(\theta;\mathbf{y},\tilde{a},\tilde{b}) \\
&\quad -\frac{1}{2}\log\left\{\left(\frac{1}{2\pi}\right)^{2k}\left[\left(\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2}\right)\left(\frac{n}{\sigma_B^2}+\frac{1}{\sigma_a^2}\right)-\left(\frac{n\rho}{\sigma_B^2}\right)^2\right]^k\right\}.
\end{aligned}$$

A similar iteration procedure as described in Chapter 3 is employed to obtain estimates. The variances of the estimators can be calculated by deriving diagonal terms of the variance-covariance matrix:

$$Cov(\hat{\theta}) = -\left[\frac{\partial^2 l_o(\theta;\mathbf{y},a^*,b^*)}{\partial\theta\partial\theta^T}\right]^{-1}_{\theta=\hat{\theta}}.$$

6.3 Joint Outcome Modeling of Zero Heavy Data

In many intermittent stream flow studies, many zero values of daily flow may be observed during dry periods in the summer. In this situation, a two-part model which accommodates zeros is helpful. We utilize a mixture of a log-normal and a zero-heavy component to account for the zeros. In this case, conditional on the annual random effect b_i described above, suppose that response variable Y_{ij}, representing streamflow at a specific station, is distributed as

$$Y_{ij}|z_{ij} = \begin{cases} 0 & \text{if } z_{ij} = 1 \\ lognormal(\mu_{ij}, \sigma^2) & \text{if } z_{ij} = 0 \end{cases}$$

The variable z_{ij} is a latent Bernoulli indicator for the zero-heavy component with mean function π_{ij}, whereas $lognormal(\mu_{ij}, \sigma^2)$ represents an independent log-normal random variable with mean μ_{ij} and variance component σ^2.

The parameters μ_{ij} and π_{ij} may depend on random effects b_i and d_i as follows:

$$\begin{aligned} \mu_{ij}^A &= X_{1ij}^A \beta^A + b_i, \text{logit}(\pi_{ij}^A) = X_{2ij}^A \alpha^A + d_i \\ \mu_{ij}^B &= X_{1ij}^B \beta^B + \rho b_i, \text{logit}(\pi_{ij}^B) = X_{2ij}^B \alpha^B + \gamma d_i \end{aligned}$$

where ρ and γ are two link parameters for the two components of the models, characterizing shared common factors in both components; b_i is the same annual effect specified in our previously proposed model, d_i is another shared annual random effect in the zero-heavy component. Note that X_{1ij}^A, X_{1ij}^B, X_{2ij}^A, X_{2ij}^B may contain the same covariates. The probability density function of Y_{ij} is:

$$f(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & \text{if } y_{ij} = 0 \\ (1 - \pi_{ij}) \frac{1}{y_{ij}\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(\log y_{ij} - \mu_{ij})^2}{2\sigma^2} \right\} & \text{if } y_{ij} = 1 \end{cases}$$

The “complete-data" likelihood function $L(\theta; y, b_i, d_i)$ can then be specified as:

$$\prod_{i=1}^{k}\prod_{j=1}^{n}\left[I(z_{ij}^{A}=1)\{\pi_{ij}^{A}\}+I(z_{ij}^{A}=0)\left\{(1-\pi_{ij}^{A})\frac{1}{y_{ij}^{A}\sigma_A\sqrt{2\pi^{A}}}\exp\left\{-\frac{(\log y_{ij}^{A}-\mu_{ij}^{A})^2}{2\sigma_A^2}\right\}\right\}\right]$$

$$\left[I(z_{ij}^{B}=1)\{\pi_{ij}^{B}\}+I(z_{ij}^{B}=0)\left\{(1-\pi_{ij}^{B})\frac{1}{y_{ij}^{B}\sigma_B\sqrt{2\pi^{B}}}\exp\left\{-\frac{(\log y_{ij}^{B}-\mu_{ij}^{B})^2}{2\sigma_B^2}\right\}\right\}\right] f(b_i)f(d_i)$$

As well in intermittent streamflow studies, quite often, the problem of autocorrelation arises, for example, dry days are often serially correlated. We may extend our model to account for autocorrelation in each of the components. This section defines important next steps in model development for streamflow data.

APPENDIX A

VARIANCE-COVARIANCE MATRICES FOR THE LAPLACE APPROXIMATION

In section 3.2 in Chapter 3, the variance-covariance matrix is:

$$\begin{pmatrix}
\frac{\partial^2 l_o}{\partial(\beta_1^A)^2} & \frac{\partial^2 l_0}{\partial\beta_1^A\partial\beta_2^A} & \cdots & \frac{\partial^2 l_o}{\partial\beta_1^A\partial\beta_{10}^B} & \frac{\partial^2 l_0}{\partial\beta_1^A\partial\sigma_A} & \frac{\partial^2 l_o}{\partial\beta_1^A\partial\sigma_B} & \frac{\partial^2 l_0}{\partial\beta_1^A\partial\rho} & \frac{\partial^2 l_o}{\partial\beta_1^A\partial\sigma_b} \\
\cdots & \frac{\partial^2 l_o}{\partial(\beta_2^A)^2} & \cdots & \frac{\partial^2 l_o}{\partial\beta_2^A\partial\beta_{10}^B} & \frac{\partial^2 l_o}{\partial\beta_2^A\partial\sigma_A} & \frac{\partial^2 l_o}{\partial\beta_2^A\partial\sigma_B} & \frac{\partial^2 l_o}{\partial\beta_2^A\partial\rho} & \frac{\partial^2 l_0}{\partial\beta_2^A\partial\sigma_b} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \frac{\partial^2 l_o}{\partial(\beta_{10}^B)^2} & \frac{\partial^2 l_o}{\partial\beta_{10}^B\partial\sigma_A} & \frac{\partial^2 l_o}{\partial\beta_{10}^B\partial\sigma_B} & \frac{\partial^2 l_0}{\partial\beta_{10}^B\partial\rho} & \frac{\partial^2 l_0}{\partial\beta_{10}^B\partial\sigma_b} \\
\cdots & \cdots & \cdots & \cdots & \frac{\partial^2 l_o}{\partial\sigma_A^2} & \frac{\partial^2 l_o}{\partial\sigma_A\partial\sigma_B} & \frac{\partial^2 l_o}{\partial\sigma_A\partial\rho} & \frac{\partial^2 l_0}{\partial\sigma_A\partial\sigma_b} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 l_o}{\partial\sigma_B^2} & \frac{\partial^2 l_o}{\partial\sigma_B\partial\rho} & \frac{\partial^2 l_o}{\partial\sigma_B\partial\sigma_b} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 l_o}{\partial\rho^2} & \frac{\partial^2 l_o}{\partial\rho\partial\sigma_b} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 l_o}{\partial\sigma_b^2}
\end{pmatrix}_{24\times 24}$$

Where

$$
\begin{aligned}
\frac{\partial^2 l_o}{\partial(\beta_1^A)^2} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(x_{ij(1)})^2}{\sigma_A^2} \\
\frac{\partial^2 l_o}{\partial(\beta_2^A)^2} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(x_{ij(2)})^2}{\sigma_A^2} \\
\frac{\partial^2 l_o}{\partial\beta_1^A\partial\beta_2^A} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{x_{ij(1)}x_{ij(2)}}{\sigma_A^2} \\
&\dots \\
\frac{\partial^2 l_o}{\partial(\beta_{10}^B)^2} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(x_{ij(10)})^2}{\sigma_B^2} \\
\frac{\partial^2 l_o}{\partial\beta_1^B\partial\beta_2^B} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{x_{ij(1)}x_{ij(2)}}{\sigma_B^2} \\
&\dots \\
\frac{\partial^2 l_o}{\partial\beta_1^A\partial\beta_{10}^B} &= 0 \\
&\dots \\
\frac{\partial^2 l_o}{\partial\sigma_A^2} &= \sum_{i=1}^{k}\sum_{j=1}^{n}\left[\frac{1}{\sigma_A^2}-\frac{3(\ln y_{ij}-X\beta-b_i)^2}{\sigma_A^4}\right]-\frac{kn[n+(\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2})3\sigma_A^2]}{(n\sigma_A+[\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2})\sigma_A^3]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_B^2} &= \sum_{i=1}^{k}\sum_{j=1}^{n}\left[\frac{1}{\sigma_B^2}-\frac{3(\ln y_{ij}-X\beta-\rho b_i)^2}{\sigma_B^4}\right]-\frac{kn\rho^2[n\rho^2+(\frac{n\rho^2}{\sigma_A^2}+\frac{1}{\sigma_b^2})3\sigma_B^2]}{[n\rho^2\sigma_B+(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^3]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_b^2} &= \sum_{i=1}^{k}\left[\frac{1}{\sigma_b^2}-\frac{3b_i^2}{\sigma_b^4}\right]-k\frac{3(\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2})\sigma_b^2+1}{[(\frac{n}{\sigma_A^2}+\frac{n\rho^2}{\sigma_B^2})\sigma_b^3+\sigma_b]^2} \\
\frac{\partial^2 l_o}{\partial\rho^2} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{b_i^2}{\sigma_B^2}+kn\frac{n\rho^2-(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^2}{[n\rho^2+(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^2]^2} \\
&\dots
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l_o}{\partial\sigma_A\partial\sigma_B} &= kn\frac{\frac{2n\rho^2\sigma_A^3}{\sigma_B^3}}{[n\sigma_A+(\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2})\sigma_A^3]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_A\partial\rho} &= -kn\frac{\frac{2n\rho\sigma_A^3}{\sigma_B^2}}{[n\sigma_A+(\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2})\sigma_A^3]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_A\partial\sigma_b} &= kn\frac{\frac{2\sigma_A^3}{\sigma_b^3}}{[n\sigma_A+(\frac{n\rho^2}{\sigma_B^2}+\frac{1}{\sigma_b^2})\sigma_A^3]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_B\partial\rho} &= -2\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(\ln y_{ij}^B - X\beta^B - \rho b_i)b_i}{\sigma_B^3} + kn\frac{2\rho(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^3}{[n\rho^2\sigma_B+(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^3]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_B\partial\sigma_b} &= kn\frac{2\rho^2\frac{\sigma_B^3}{\sigma_b^3}}{[n\rho^2\sigma_B+(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^3]^2} \\
\frac{\partial^2 l_o}{\partial\rho\partial\sigma_b} &= -kn\frac{2\rho\frac{\sigma_B^2}{\sigma_b^3}}{[n\rho^2+(\frac{n}{\sigma_A^2}+\frac{1}{\sigma_b^2})\sigma_B^2]^2} \\
\frac{\partial^2 l_o}{\partial\sigma_A\partial\beta_1^A} &= -2\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(\ln y_{ij}^A - X\beta^A - b_i)x_{ij(1)}}{\sigma_A^3} \\
\frac{\partial^2 l_o}{\partial\sigma_A\partial\beta_{10}^A} &= -2\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(\ln y_{ij}^A - X\beta^A - b_i)x_{ij(10)}}{\sigma_A^3} \\
\frac{\partial^2 l_o}{\partial\sigma_B\partial\beta_1^B} &= -2\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(\ln y_{ij}^B - X\beta^B - \rho b_i)x_{ij(1)}}{\sigma_B^3} \\
\frac{\partial^2 l_o}{\partial\sigma_B\partial\beta_{10}^B} &= -2\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(\ln y_{ij}^B - X\beta^B - \rho b_i)x_{ij(10)}}{\sigma_B^3} \\
\frac{\partial^2 l_o}{\partial\beta_1^B\partial\rho} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{x_{ij(1)}b_i}{\sigma_B^2} \\
& \ldots \\
\frac{\partial^2 l_0}{\partial\beta_{10}^B\partial\rho} &= -\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{x_{ij(10)}b_i}{\sigma_B^2}
\end{aligned}
$$

all the other terms are zero.

REFERENCES

Bigelow, J. L. and Dunson, D. B. (2009), “Bayesian Semiparametric Joint Models for Functional Predictors,” *Journal of the American Statistical Association*, 104, 26–36.

Catalano, P. J. (1997), “Bivariate modelling of clustered continuous and ordered categorical outcomes,” *Statistics in Medicine*, 16, 883–900.

Catalano, P. J. and Ryan, L. M. (1992), “Bivariate latent variable models for clustered discrete and continuous outcomes,” *Journal of the American Statistical Association*, 87, 651–658.

Cox, D. R. and Wermuth, N. (1992), “Response models for mixed binary and quantitative variables.” *Biomatrika*, 79, 441–461.

Dunson, D., Chen, Z., and Harry, J. (2003), “A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes,” *Biometrics*, 59, 521–530.

Dunson, D. B. (2003), “Dynamic latent trait models for multidimensional longitudinal data,” *Journal of the American Statistical Association*, 98, 555–563.

Faucet, C. J. and Thomas, D. (1996), “Simultaneously modeling censored survival data and repeatedly measured covariates: A Gibbs sampling approach.” *Statistics in Medicine*, 15, 1663–1685.

Feng, C. and Dean, C. (2012), “Joint Analysis of Multivariate Spatial Count and Zero-Heavy Count Outcomes Using Common Spatial Factor Models.” *Environmetrics*, 23, 493–508.

Fieuws, S., Verbeke, G., Maes, B., and Vanrenterghem, Y. (2008), “Predicting renal graft failure using multivariate longitudinal profiles.” *Biostatistics*, 9, 419.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008), *Longitudinal Data Analysis*, Chapman & Hall/CRC.

Gueorguieva, R. and Agresti, A. (2001), “A corrlated probit model for joint modeling of cluster binary and contunuous reponses.” *Journal of the American Statistical Association*, 96, 1102–1112.

Johnson, R. A. and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis, 5th Edition*, Prentice Hall.

Krzanowski, W. J. (1988), *Principles of Multivariate Analysis.*, Claredon Press.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006), *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.*, Chapman & Hall/CRC.

McCulloch, C. (2008), "Joint modelling of mixed outcome types using latent variables," *Statistical Methods in Medical Research*, 17, 53–73.

Millar, R. (2011), *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB (Statistics in Practice)*, Wiley.

Molenberghs, G. and Ryan, L. (2002), "An exponential family model for clustered mutivariate binary data." *Environmetrics*, 10, 279–300.

Price, C., Kimmel, C., Tyl, R., and Marr, M. (1985), "The developmental toxicity of thylene glycol in rats and mice," *Toxicology and Applied Pharmacology*, 81, 113–127.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 319–392.

Skaug, H. and Fournierb, D. (2006), "Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models," *Computational Statistics & Data Analysis*, 51, 699–709.

Taylor, J. M. G. and Wang, Y. (2002), "Surrogate markers and joint models for longitudinal and survival data," *Controlled Clinical Trials*, 23, 626–634.

Tsiatis, A. A. and Davidian, M. (2004), "Joint modeling of longitudinal and time-to-event data: An overview." *Statistica Sinica*, 14, 809–834.

Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. (2002), "Conditional second order generalized estimating equations for generalized linear and nonlinear mixed-effects models," *Journal of the American Statistical Association*, 97, 271–283.

Wood, S. (2006), *Generalized Addtive Models: An Introduction with R.*, Chapman & Hall/CRC.

Wu, H. and Ding, A. (1999), "Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from aids clinical trials," *Biometrics*, 55, 410–418.

Wu, L., Liu, W., Y, G., and Huang, Y. (2012), "Analysis of longitudinal and survival data: joint modeling, inference methods, and issues," *Journal of Probability and Statistics*, 2012, 1–17.

Wu, M. C. and Carroll, R. J. (1988), "Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process." *Biometrics*, 44, 175–188.

Wulfsohn, M. and Tsiatis, A. A. (1997), “A joint model for survival and longitudinal data measured with error,” *Biometrics*, 53, 330–339.

Curriculum Vitae

Name:	Lihua Li
Post-Secondary Education and Degrees:	Wuhan University Wuhan, Hubei, China 1997 - 2001 B.A.
	Wuhan University Wuhan, Hubei, China 2004 - 2007 Ph.D.
	Simon Fraser University Burnaby, BC, Canada 2006 - 2008 M.A.
	The University of Western Ontario London, ON, Canada 2011 - 2013 M.Sc.
Honours and Awards:	Queen Elizabeth II Scholarship 2012-2013
Related Work Experience:	Teaching Assistant The University of Western Ontario 2011 - 2013

Publications:

Weir MA, Jain AK, Gomes T, Juurlink DN, Mamdani M, **Li L**, Garg AX, Sevelamer prescriptions after reporting of the Dialysis Clinical Outcomes Revisited (DCOR) trial findings: An analysis of 5,495 patients receiving maintenance dialysis in Ontario, Canada, *Am J Kidney Dis.* 2011 Feb; 57(2): 357-9.

Siddiqui N, DO S, Devereaux PJ, Jain AK, **Li L**, Luo J, Parikh C, Paterson M, Thiessen-

Philbrook H , Wald R, Walsh M, Whitlock R , Garg AX, Secular trends in acute dialysis following major elective surgery, 1995 to 2009, *CMAJ.* 2012 Aug 7; 184(11): 1237-45.

Doshi MD, Goggins MO, **Li L**, Garg AX, Medical outcomes in African American live kidney donors: a matched cohort study, *Am J Transplant.* 2013 Jan; 13(1): 111-8.