2012 IEEE 12th International Conference on Data Mining

Student-t based Robust Spatio-Temporal Prediction

Yang Chen*, Feng Chen*, Jing Dai†, T. Charles Clancy‡ and Yao-Jan Wu§
*Department of Computer Science, Virginia Tech, VA 22043
†Google Inc. New York, NY 10011
‡Bradley Electrical and Computer Engineering, Virginia Tech, VA 22203
§Department of Civil Engineering, Saint Louis University, St. Louis, MO 63103
{yangc10*, chenf*, tcc‡}@vt.edu, jddai@google.com†, yaojan@slu.edu§

***Abstract*—This paper describes an efficient and effective design of Robust Spatio-Temporal Prediction based on Student's t distribution, namely, St-RSTP, to provide estimations based on observations over spatio-temporal neighbors. The proposed St-RSTP is more resilient to outliers or other small departures from model assumptions than its ancestor, the Spatio-Temporal Random Effects (STRE) model. STRE is a state-of-the-art statistical model with linear order complexity for large scale processing. However, it assumes Gaussian observations, which has the well-known limitation of non-robustness. In our St-RSTP design, the measurement error follows Student's t distribution, instead of a traditional Gaussian distribution. This design reduces the influence of outliers, improves prediction quality, and keeps the problem analytically intractable. We propose a novel approximate inference approach, which approximates the model into the form that separates the high dimensional latent variables into groups, and then estimates the posterior distributions of different groups of variables separately in the framework of Expectation Propagation. As a good property, our approximate approach degeneralizes to the standard STRE based prediction, when the degree of freedom of the Student's t distribution is set to infinite. Extensive experimental evaluations based on both simulation and real-life data sets demonstrated the robustness and the efficiency of our Student-t prediction model. The proposed approach provides critical functionality for stochastic processes on spatio-temporal data.**

***Keywords*-Spatio-Temporal Process; Expectation Propagation; Student's t Distribution.**

I. Introduction

Predicting spatial and temporal data is an essential component in many emerging applications in geographical information systems, medical imaging, urban planning, economy study, and climate forecasting. In the real world, most physical, biological, or social processes involve some degree of spatial and temporal variability [1]. It is suggested that any application that requires dynamic and stochastic process as a component should take spatial and temporal dependencies into account [2]. In these processes, an efficient and robust spatiotemporal prediction approach helps identify the causalities due to environmental effects, and forecast the impact of changes. Applications of such an approach include predicting traffic of an unsensored road segment using nearby traffic sensors, and estimating average income using known samples in similar geographic locations.

There have been two paradigms for spatio-temporal prediction, Kriging based and dynamical (mechanic or probabilistic) specification based. The Kriging based paradigm basically extends spatial dimensions (d) with an extra time dimension and focuses on the modeling of the variance-covariance structure between the observations in the $(d+1)$-dimensional space. The dynamic specification based paradigm considers spatio-temporal processes through a dynamical-statistical (or state space based) framework. In this framework the observations in the current state are dependent on its previous states through dynamic mechanical (or probabilistic) relationships. Our work focuses on the dynamic statistical paradigm, which can be explicitly specified based on the knowledge of the phenomenon under study. It always leads to a valid variance-covariance structure, and allows fast filtering, smoothing, and forecasting [3].

One emerging research challenge for spatio-temporal prediction is to efficiently model massive spatio-temporal data that have been collected by using advanced remote sensing technologies. For example, NASA collects data on the order of 100,000 observations per day from satellites. Big data challenges from smartphone usages have recently attracted a lot of research efforts [4]. Given the large data volume, most traditional spatio-temporal statistical models fail to process in either memory space or execution time, even in supercomputing environments. Although recent progresses have been made [5], the preceding works are still unable to achieve near-real-time performance and thus not suitable for processing massive streaming spatial data.

As the most recent advancement, [2] presents a spatio-temporal random effects (STRE) model that reduces the problem into a fixed dimension problem and makes it possible to do fast filtering, smoothing, and forecasting with a linear order time complexity. The STRE model assumes that 1) the spatial dependence can be captured by a predefined set of basis functions; 2) the temporal dependence can be modeled by a latent first-order Gaussian autoregressive process; and 3) the measurement error can be modeled by a Gaussian distribution. These assumptions make the STRE model mainly applicable to linear dynamic environments

However, the spatio-temporal dynamics of real applications are usually nonlinear, and some of the STRE's distribution assumptions are often violated. For example, the

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.135

data may have a number of outliers, such as random hardware failures in digital control systems [6], sensor faults in aerospace applications [7], cochannel fading and interference in wireless communications [4], and traffic incidents and malfunctioning detectors in urban traffic networks [17]. This paper presents a robust spatio-temporal prediction approach for applications in nonlinear dynamic environments where some of the STRE assumptions are violated.

In recent years, robust methods have received much attention for a variety of learning problems(e.g., [8], [9], [10], [11], [6], [12]). The majority of these methods can be summarized using a probabilistic framework [8] in which the measurement error is modeled by a heavy tailed distribution, instead of the traditional Gaussian distribution. However, employing heavy tailed distributions makes the prediction process analytically intractable. Although stochastic simulation methods have been applied to estimate an approximate posterior distribution, for example via MCMC or particle filtering [9], they are very computationally intensive. An efficient expectation propagation algorithm [10] was presented for robust Gaussian process regression based on the Student's t distribution. Similar efforts include a variational inference approach [11] for robust Student's t mixture clustering, a robust Kalman filter [6] based on the Huber distribution, and a Kalman smoother [12] based on the Laplace distribution.

This paper focuses on robust prediction in a probabilistic framework. We propose an observation model for spatio-temporal prediction based on Student's t. Because of its good robustness properties, the Student's t can be altered continuously from a very heavy tailed distribution to the Gaussian model with the degrees of freedom parameter. Further more, this work resolves the main challenge of the student-t based model, which is the analytically intractable inference of high dimensional latent variables. The main contributions of our study can be summarized as follows.

- We formalize an innovative robust prediction model for spatio-temporal data in a systematical framework;
- We approximate the robust prediction model such that the high-dimensional latent variables can be separated into groups that can be optimized iteratively.
- We present novel implementations of Expectation Propagation (EP) in order to efficiently estimate the posterior distributions of latent variables.
- We validate the robustness and the efficiency of the proposed St-RSTP model compared with the regular STRE model by an extensive simulation study and experiments on two real data sets.

The rest of the paper is organized as follows. Preliminaries on the formulation and inference algorithms of the regular STRE model is reviewed in Section II. Section III presents the robust spatio-temporal prediction model, St-RSTP, followed by the detailed approximation prediction techniques based on EP in Section IV. Simulation study and evaluation of our proposed robust smoothing algorithm on two real world data sets are illustrated in Section V. Finally, we conclude our work in Section VI.

II. THEORETICAL BACKGROUNDS

This section reviews the Spatio-Temporal Random Effects (STRE) model and STRE-based spatio-temporal prediction.

A. *Spatio-Temporal Random Effects Model*

The STRE model is a recently proposed statistical model for processing large spatio-temporal data in linear order time complexity [2]. The STRE model is used to model a spatial random process that evolves over time, $\{Y_t(\mathbf{s}) \in \Re : s \in D \subset \Re^2, t = 1, 2, \cdots\}$, where D is the spatial domain under study, and $Y_t(\mathbf{s})$ is the nonspatial measurement (e.g., temperature) at location $\mathbf{s}$ and time t.

A discretized version of the process can be represented as

$$\{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_t, \mathbf{Y}_{t+1}, \cdots\}, \tag{1}$$

where $\mathbf{Y}_t = [Y_t(\mathbf{s}_{1,t}), Y_t(\mathbf{s}_{2,t}), \cdots, Y_t(\mathbf{s}_{m_t,t})]^T$. The sample locations $\{\mathbf{s}_{1,t}, \mathbf{s}_{2,t}, \cdots, \mathbf{s}_{m_t,t}\}$ can be different spatial locations at different time t. Observations $\mathbf{Z}_t$ and latent observations $\mathbf{Y}_t$are given by the data process,

$$\mathbf{Z}_t = \mathbf{O}_t\mathbf{Y}_t + \varepsilon_t, t = 1, 2, \cdots, \tag{2}$$

where $\mathbf{Z}_t$ is an n_t-dimenstional vector ($n_t \leq m_t$), $\mathbf{O}_t$ is an $n_t \times m_t$ incidence matrix, used to handle missing values that are related to locations where no observations are available, and $\varepsilon_t = [\varepsilon_t(\mathbf{s}_{1,t}), \cdots, \varepsilon_t(\mathbf{s}_{n_t,t})]^T \sim \mathcal{N}_{n_t}(\mathbf{0}, \sigma^2_{\varepsilon,t}\mathbf{V}_{\varepsilon,t})$ is a vector of white noise Gaussian processes, with $\mathbf{V}_{\varepsilon,t} = diag(v_{\varepsilon,t}(\mathbf{s}_{1,t}), \cdots, v_{\varepsilon,t}(\mathbf{s}_{n_t,t}))$. Particularly, $var(\varepsilon_t(\mathbf{s})) = \sigma^2_{\varepsilon,t}v(\mathbf{s}) > 0, \sigma^2_{\varepsilon,t}$ is a parameter to be estimated, and $v(\mathbf{s})$ is known. The white noise assumption implies that $cov(\varepsilon_t(\mathbf{s}), \varepsilon_u(\mathbf{r})) = 0$, for $t \neq u$ and $\mathbf{s} \neq \mathbf{r}$.

The vector $\mathbf{Y}_t$ is given by the spatial process:

$$\mathbf{Y}_t = \mathbf{X}_t\beta_t + \nu_t, t = 1, 2, \cdots, \tag{3}$$

where $\mathbf{X}_t = [\mathbf{x}_t(\mathbf{s}_{1,t}), \cdots, \mathbf{x}_t(\mathbf{s}_{m_t,t})]^T$, $\mathbf{x}_t(\mathbf{s}_{i,t}) \in \Re^p, 1 \leq i \leq m_t$, represents a vector of covariates, and the coefficients $\beta_t = (\beta_{1,t}, \cdots, \beta_{p,t})^T$ are general unknown. The random process ν_t captures the small scale variations. For traditional spatio-temporal Kalman filtering models, a large number of parameters need to be estimated with high computational costs due to the high data dimensionality during the filtering, smoothing, and forecasting processes. As a key advantage of the STRE model, it models the small scale variation ν_t as a vector of spatial random effects (SRE) processes

$$\nu_t = \mathbf{S}_t^T\eta_t + \xi_t, t = 1, 2, \cdots, \tag{4}$$

where $\mathbf{S}_t = [S_t(\mathbf{s}_{1,t}), \cdots, S_t(\mathbf{s}_{m_t,t})]$, $S_t(\mathbf{s}_{i,t}) = [S_{1,t}(\mathbf{s}_{i,t}), \cdots, S_{r,t}(\mathbf{s}_{i,t})]^T$, $1 \leq i \leq m_t$, is a vector of r predefined spatial basis functions, such as wavelet and bisquare basis functions, and η_t is an r-dimensional zero-mean Gaussian random vector with an $r \times r$ covaraince matrix given by $\mathbf{K}_t$. The first component in Equation (4) denotes a smoothed small-scale variation at time t, captured by the set of basis functions $\mathbf{S}_t$.

The second component in Equation (4) captures the micro-scale variability similar to the nugget effect as defined in geostatistics [2]. It is assumed that $\xi_t \sim \mathcal{N}_{m_t}(\mathbf{0}, \sigma^2_{\xi,t}\mathbf{V}_{\xi,t})$, $\mathbf{V}_{\xi,t} = diag(v_{\xi,t}(\mathbf{s}_{1,t}), \cdots, v_{\xi,t}(\mathbf{s}_{m_t,t}))$, and $v_{\xi,t}(\cdot)$ describes the variance of the micro-scale variation and is typically considered known. Note that the component ξ_t is important, since it can be used to capture the extra uncertainty due to the dimension reduction in replacing ν_t by $\mathbf{S}_t^T\eta_t$. The coefficient vector η_t is assumed to follow a vector-autoregressive process of order one,

$$\eta_t = \mathbf{H}_t\eta_{t-1} + \zeta_t, t = 1, 2, \cdots, \quad (5)$$

where $\mathbf{H}_t$ refers to the so-called propagator matrix, $\zeta_t \sim \mathcal{N}(0, \mathbf{U}_t)$ is an r-dimensional innovation vector, and $\mathbf{U}_t$ is named as the innovation matrix. The initial state $\eta_0 \sim \mathcal{N}_r(\mathbf{0}, \mathbf{K}_0)$ and $\mathbf{K}_0$ is in general unknown.

Combining Equations (2), (3), and (4), the (discretized) data process can be represented as

$$\mathbf{Z}_t = \mathbf{O}_t\mu_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\xi_t + \varepsilon_t, t = 1, 2, \cdots, \quad (6)$$

where $\mu_t = \mathbf{X}_t\beta_t$ is deterministic and the other components are stochastic.

B. *STRE based Spatio-Temporal Prediction*

Given a set of observations $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_T\}$, the spatio-temporal prediction problem is to predict the latent (or de-noised) values $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_t\}$. As discussed in Subsection II-A, the incidence matrix $\mathbf{O}_t$ allows for the specification of missing observations, which makes it possible to concurrently predict the latent Y values for both observed and unobserved locations. This is a smoothing problem if $t < T$; and a filtering problem if $t = T$; and a forecasting problem if $t > T$. Readers are referred to [2] for the detailed STRE based prediction equations.

III. Problem Formulation

This section introduces the new Robust Spatio-Temporal Prediction model based on Student's t, St-RSTP, and describes the problem of estimating the posterior distributions $p(\mathbf{Y}_t|\mathbf{Z}_{1:t})$ and $p(\mathbf{Y}_t|\mathbf{Z}_{1:T})$ for spatial prediction.

A. *Robust Spatio-Temporal Prediction Model*

The Robust Spatio-Temporal Prediction model based on Student-t (St-RSTP) considers Student's t distribution to model the measurement error, instead of the traditional Gaussian distribution. Student's t distribution has a heavier tail than Gaussian distribution. The tail heaviness is controlled by setting the degrees of freedom (ν). When the degree of freedom approaches infinity, Student's t distribution becomes equivalent to Gaussian distribution. Student's t distribution has been used in a number of statistical models, and has been shown effective for a variety of robust processes [1], [10].

We use the same symbols and definitions as in subsection II-A. The St-RSTP model can be formalized as

$$\mathbf{Z}_t = \mathbf{O}_t\mathbf{Y}_t + \varepsilon_t, \quad (7)$$

$$\mathbf{Y}_t = \mathbf{X}_t\beta_t + \mathbf{S}_t^T\eta_t + \xi_t, \quad (8)$$

$$\eta_t = \mathbf{H}_t\eta_{t-1} + \zeta_t. \quad (9)$$

As a key difference from the STRE model, the measurement error ε_{tn} now follows a Student's t distribution $Student\text{-}t(0, \nu, \sigma)$ with the probability density function as

$$p(\varepsilon_{tn}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}(\frac{1}{\pi\nu\sigma})^{\frac{1}{2}}(1 + \frac{\varepsilon_{tn}^2}{\nu\sigma})^{-\frac{\nu}{2}-\frac{1}{2}}, \quad (10)$$

where ν is the degrees of freedom and σ is the scale parameter.

B. *Problem Formulation for Robust Prediction*

Given the observations $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_T\}$, the predictive process is to estimate the latent variables $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_t\}$ at sampled and unsampled locations, where $t = 1, 2, \cdots$ The estimation of $\mathbf{Y}$ variables at unsampled locations is realized by using the incidence matrix $\mathbf{O}_t$ in the St-RSTP model, where $\mathbf{O}_t \in \mathcal{R}^{n_t \times m_t}$, n_t refers to the number of observations at sampled locations, and $m_t - n_t$ refers to the number of unsampled locations that are of interest for prediction.

The objective of this paper is to estimate the expectation and variance-covariance of the posterior distributions $p(\mathbf{Y}_t|\mathbf{Z}_{1:T}), t = 1, 2, \cdots$, denoted as $\mathbf{Y}_{t|T}$ and $\mathbf{\Sigma}_{t|T}$, respectively. $\mathbf{Y}_{t|T}$ will be regarded as the prediction values, and $\mathbf{\Sigma}_{t|T}$ will be applied to estimate confidence intervals. Specifically, if $t < T$, the predictive process is called smoothing; if $t = T$, the predictive process is called filtering; and if $t = T + k$, $k > 0$, the predictive process is called k-step forecasting.

According to the STRE model decomposition as shown in Equation 6, we can first estimate the mean and variance-covaraince matrix of the joint posterior $p(\eta_t, \xi_t|\mathbf{Z}_{1:T})$. The components $\mathbf{Y}_{t|T}$ and $\mathbf{\Sigma}_{t|T}$ can then be estimated by linear transformations. However, the total dimension of η_t and ξ_t is "$r + m_t$". This high dimensionality makes the estimation process computationally expensive even using advanced convex optimization techniques.

IV. Approximate Spatio-Temporal Prediction

In this section, we first present an approximate St-RSTP model, such that the posterior distributions of latent variables $\{\eta_t, \xi_t\}_{t=1}^T$ can be estimated iteratively. EP based approximate algorithms are then designed in order to efficiently infer the posterior distributions $p(\eta_t|\mathbf{Z}_{1:T})$ and $p(\xi_t|\mathbf{Z}_{1:T})$.

A. *Approximate St-RSTP Model*

Let $\eta_{t|T} \equiv E[p(\eta_t|\mathbf{Z}_{1:T})]$, $\mathbf{P}_{t|T} \equiv Var[p(\eta_t|\mathbf{Z}_{1:T})]$, $\xi_{t|T} \equiv E[p(\xi_t|\mathbf{Z}_{1:T})]$, and $\mathbf{R}_{t|T} \equiv Var[p(\xi_t|\mathbf{Z}_{1:T})]$. It follows that

$$\mathbf{Y}_{t|T} = \mathbf{X}_t\beta_t + \mathbf{S}_t^T\eta_{t|T} + \xi_{t|T}, .$$

In order to efficiently estimate the variance-covariance matrix $\mathbf{\Sigma}_{t|T}$, we make the approximation as

$$\mathbf{\Sigma}_{t|T} \approx \mathbf{S}_t^T \mathbf{P}_{t|T} \mathbf{S}_t + \mathbf{R}_{t|T}. \quad (11)$$

Based on the above strategy, the major task is to conduct Gaussian approximations to $p(\eta_t|\mathbf{Z}_{1:T})$ and $p(\xi_t|\mathbf{Z}_{1:T})$:

$$p(\eta_t|\mathbf{Z}_{1:T}) \sim_G \mathcal{N}(\eta_{t|T}, \mathbf{P}_{t|T}) \quad (12)$$
$$p(\xi_t|\mathbf{Z}_{1:T}) \sim_G \mathcal{N}(\xi_{t|T}, \mathbf{R}_{t|T}). \quad (13)$$

A popular strategy is to calculate the maximum-a-posterior (MAP) estimations of the above posteriors using numerical optimization techniques (e.g., gradient decent, interior point algorithms), and then calculate the corresponding Hessian matrices at the MAP locations. However, there exist no analytical forms of the posteriors

$$p(\eta_t|\mathbf{Z}_{1:T}) = \int p(\xi_t, \eta_t|\mathbf{Z}_T) d\xi_t, \quad (14)$$
$$p(\xi_t|\mathbf{Z}_{1:T}) = \int p(\xi_t, \eta_t|\mathbf{Z}_T) d\eta_t, \quad (15)$$

and the application of numerical optimizations is difficult, because no analytical forms of gradient and Hessian matrix can be calculated. The following presents several approximations to make the estimation of the posteriors tractable.

Phase I: Approximate Estimation of $\eta_{t|T}$ and $\mathbf{P}_{t|T}$

The St-RSTP model can be reformulated as follows

$$\mathbf{Z}_t = \mathbf{O}_t \mathbf{X}_t \beta_t + \mathbf{O}_t \mathbf{S}_t^T \eta_t + \mathbf{O}_t \xi_t + \varepsilon_t, \quad (16)$$
$$\eta_t = \mathbf{H}_t \eta_{t-1} + \zeta_t \quad (17)$$

The component ξ_{tn} captures a micro-scale variation and is modeled by a white noise Gaussian process with mean zero and variance $var(\xi(\mathbf{s};t)) = \sigma_\xi^2 v_t'(\mathbf{s})$. The component ε_{tn} is a Student's t process with mean zero and variance $var(\varepsilon(\mathbf{s};t)) = \sigma_\varepsilon^2 v_t(\mathbf{s})$. An approximation is made as

$$\tilde{\xi}_t = \mathbf{O}_t \xi_t + \varepsilon_t, \quad (18)$$
$$\tilde{\xi}_{tn} \sim \textit{Student-t}(0, \nu, \tilde{\sigma}). \quad (19)$$

The approximate St-RSTP model can be reformulated as

$$\mathbf{Z}_t = \mathbf{O}_t \mathbf{X}_t \beta_t + \mathbf{O}_t \mathbf{S}_t^T \eta_t + \tilde{\xi}_t, \quad (20)$$
$$\eta_t = \mathbf{H}_t \eta_{t-1} + \zeta_t. \quad (21)$$

Figure 1 shows the graph model representation about the statistical relationships between observation Z_t and latent variables η_t and ξ_t.

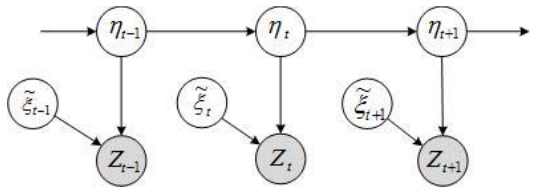


Figure 1: Approximate St-RSTP Graphic Model

Based on the above approximate St-RSTP model, the subsequent subsection (IV-B) presents an efficient EP-based algorithm to conduct Gaussian approximation of $p(\eta_t|\mathbf{Z}_{1:T})$.

Phase II: Approximate Estimation of $\xi_{t|T}$ and $\mathbf{R}_{t|T}$

In order to estimate $\xi_{t|T}$ and $\mathbf{R}_{t|T}$, we need to first conduct Gaussian approximation of the posterior $p(\xi_t|\mathbf{Z}_{1:t})$. The joint posterior distribution

$$p(\xi_t, \eta_t|\mathbf{Z}_{1:t}) = p(\xi_t, \eta_t, \mathbf{Z}_t|\mathbf{Z}_{1:t-1}) \quad (22)$$
$$= p(\mathbf{Z}_t|\eta_t, \xi_t) p(\xi_t|\eta_t) p(\eta_t|\mathbf{Z}_{1:t-1}).$$

Given $\hat{p}(\eta_{t-1}|\mathbf{Z}_{1:t-1}) \sim \mathcal{N}(\eta_{t-1|t-1}, \mathbf{P}_{t-1|t-1})$ estimated in Phase I, it follows that

$$\hat{p}(\eta_t|\mathbf{Z}_{1:t-1}) \sim \mathcal{N}(\eta_{t|t-1}, \mathbf{P}_{t|t-1}), \quad (23)$$
$$\text{where } \eta_{t|t-1} = \mathbf{H}_t \eta_{t-1|t-1},$$
$$\mathbf{P}_{t|t-1} = \mathbf{H}_t \mathbf{P}_{t-1|t-1} \mathbf{H}_t^T + \mathbf{U}_t.$$

The posterior $p(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ can be approximated as

$$\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t}) = p(\mathbf{Z}_t|\eta_t, \xi_t) p(\xi_t) \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}). \quad (24)$$

Integrating out η_t, we obtain

$$p(\xi_t|\mathbf{Z}_{1:t}) = \int p(\xi_t, \eta_t|\mathbf{Z}_t) d\eta_t$$
$$\approx \int p(\mathbf{Z}_t|\eta_t, \xi_t) p(\xi_t) \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}) d\eta_t$$
$$\approx \int \hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t}) d\eta_t. \quad (25)$$

Notice that the components $p(\xi_t)$ and $\hat{p}(\eta_t|\mathbf{Z}_{1:t-1})$ are Gaussian. By applying Gaussian approximation to $p(\mathbf{Z}_t|\eta_t, \xi_t)$, the posterior $\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ is hence approximated as Gaussian as well, and the analytical form of the above integration (25) can be obtained. An efficient EP-based algorithm is presented in subsection IV-C.

Note that in Phase I and Phase II, it is required that $t \leq T$. That means, the results are only suitable for smoothing and filtering. Given the filtering estimations $\eta_{t|T}$ and $\mathbf{P}_{t|T}$ by Phase I, the forecasting estimations $\eta_{t|T}$, $\mathbf{P}_{t|T}$, $\xi_{t|T}$, and $\mathbf{R}_{t|T}$, where $t = T + k$ and $k > 0$, can be obtained based on the regular STRE model [2], because it is unnecessary to consider outliers in future "observations".

$$\eta_{T+k|T} = \left(\prod_{i=T+1}^{T+k} \mathbf{H}_i \right) \eta_{t|T}, \quad (26)$$

$$\mathbf{P}_{T+k|T} = \sum_{i=T+1}^{T+k-1} \left\{ \left(\prod_{j=i+1}^{T+k} \mathbf{H}_j \right) \mathbf{U}_i \left(\prod_{j=i+1}^{T+k} \mathbf{H}_j \right)^T \right\} +$$
$$\left(\prod_{i=T+1}^{T+k} \mathbf{H}_i \right) \mathbf{P}_{T|T} \left(\prod_{i=T+1}^{T+k} \mathbf{H}_i \right)^T + \mathbf{U}_{T+k},$$

$$\xi_{T+k|T} = \mathbf{0}, \mathbf{R}_{T+k|T} = \mathbf{0}.$$

Theorem 1. *If the degree-of-freedom parameter of the Student's t distribution used in the St-RSTP model is set*

to infinite, then the estimation results of $p(\eta_t|\mathbf{Z}_{1:T})$ and $p(\xi_t|\mathbf{Z}_{1:T})$ by Phase I and II, as well as the prediction results by Equations (10) and (11), are equivalent to the exact estimation and prediction results of the standard STRE model.

Proof: The proof is removed due to space limit. ■

The above theorem presents a pleasant theoretical property of our proposed St-RSTP model. It shows that the standard STRE is a special case of our robust model.

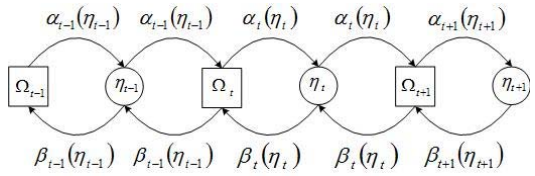


Figure 2: Factor Graph Presentation of St-RSTP

B. EP-Based Estimation of $\eta_{t|T}$ and $\mathbf{P}_{t|T}$

In order to apply EP to the estimation problem, we first present the factor graph [13] representation in the framework of dynamic Bayesian networks as shown in Figure 2. From Figure 2, the joint distribution of latent variables and observations, forward and backward message passing components $\alpha(\cdot)$ and $\beta(\cdot)$ can be derived from literature [14], as showed below:

$$\begin{aligned} p(\eta_{1:T}, \mathbf{Z}_{1:T}) &= p(\eta_1)p(\mathbf{Z}_1|\eta_1)\prod_{t=2}^{T} p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t), \\ \alpha_t(\eta_t) &= p(\mathbf{Z}_t|\eta_t)\int p(\eta_t|\eta_{t-1})\alpha_{t-1}(\eta_{t-1})d\eta_{t-1}, \\ \beta_{t-1}(\eta_{t-1}) &= \int p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\beta_t(\eta_t)d\eta_t. \end{aligned} \quad (27)$$

The posterior distribution of latent variable can be reformalized as the production of factor functions:

$$p(\eta_{1:T}|\mathbf{Z}_{1:T}) \propto \prod_t \Omega_t(\eta_{t-1}, \eta_t), \quad (28)$$

where each factor function is represented as

$$\Omega_t(\eta_{t-1}, \eta_t) := p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t),$$

and $\Omega_t(\eta_0, \eta_1) := p(\eta_1)p(\mathbf{Z}_1|\eta_1)$, when $t = 1$.

Recall that $p(\mathbf{Z}_t|\eta_t)$ follows a Student's t distribution, the estimation of Equation (27) is intractable. It can be further approximated as the following factorized form

$$q(\eta) = \prod_t q_t(\eta_{t-1}, \eta_t) \propto \prod_t \hat{\Omega}(\eta_{t-1}, \eta_t), \quad (29)$$

where $\hat{}$ indicates an approximation of the corresponding symbol.

Combining Equations (27) and (28), the smoothing latent variable can be estimated by

$$\begin{aligned} p(\eta_t|\mathbf{Z}_{1:T}) &\approx q_t(\eta_t) \propto \hat{\alpha}_t(\eta_t)\hat{\beta}_t(\eta_t) && (30) \\ p(\eta_{t-1}, \eta_t|\mathbf{Z}_{1:T}) &\approx \hat{p}_t(\eta_{t-1}, \eta_t), && (31) \\ &\propto \hat{\alpha}_{t-1}(\eta_{t-1})p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\hat{\beta}_t(\eta_t) \\ &= \hat{\alpha}_{t-1}(\eta_{t-1})\Omega_t(\eta_{t-1}, \eta_t)\hat{\beta}_t(\eta_t). \end{aligned}$$

Furthermore, given that from the factorial form,

$$\hat{p}_t(\eta_{t-1}, \eta_t) = q_{t-1}(\eta_{t-1})q_t(\eta_t), \quad (32)$$

plugging Equations (29), (30), (31) into Equation (32) leads to the simplified approximation form:

$$\hat{\Omega}_t(\eta_{t-1}, \eta_t) = \hat{\beta}_{t-1}(\eta_{t-1})\hat{\alpha}_t(\eta_t). \quad (33)$$

The EP algorithm refines the approximate posterior $q(\eta)$ iteratively by recomputing passing messages. As indicated in Equation (33), in order to estimate the approximate factor $\hat{\Omega}_t^{new}(\eta_{t-1}, \eta_t)$, we need to estimate $\hat{\beta}_{t-1}^{new}(\eta_{t-1})$ and $\hat{\alpha}_t^{new}(\eta_t)$. One-slice posterior distribution can be acquired by integrating one latent variable from two-slice posterior distribution. When we compute the one-slice posterior, the corresponding message can be calculated by Equation (30). Hence, by combining Equations (31) and (32), these two messages can be obtained by following two steps: 1) approximating $\hat{p}_t(\eta_{t-1}, \eta_t) \propto \hat{\alpha}_{t-1}(\eta_{t-1})p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\hat{\beta}_t(\eta_t)$ as a Gaussian distribution by Laplace Approximation

$$\hat{p}_t(\eta_{t-1}, \eta_t) \approx_{LA} \mathcal{N}(\eta_{t-1}, \eta_t \mid \mu, \mathbf{\Sigma}), \quad (34)$$

where μ and $\mathbf{\Sigma}$ match the first and second moments of $\hat{p}_t(\eta_{t-1}, \eta_t)$; 2) integrating out η_{t-1} (or η_t) to obtain $\hat{\alpha}_t^{new}(\eta_t)$ (or $\hat{\beta}_{t-1}^{new}(\eta_{t-1})$):

$$\hat{\alpha}_t^{new}(\eta_t) \propto \frac{\int \mathcal{N}(\eta_{t-1}, \eta_t \mid \mu, \mathbf{\Sigma})d\eta_{t-1}}{\hat{\beta}_t(\eta_t)}, \quad (35)$$

$$\hat{\beta}_{t-1}^{new}(\eta_{t-1}) \propto \frac{\int \mathcal{N}(\eta_{t-1}, \eta_t \mid \mu, \mathbf{\Sigma})d\eta_t}{\hat{\alpha}_{t-1}(\eta_{t-1})}. \quad (36)$$

The above strategy outputs the estimated messages $\hat{\alpha}_t(\eta_t)$ and $\hat{\beta}_t(\eta_t)$, $t = 1, \cdots, T$, each of which follows a Gaussian distribution, with known parameters. The posterior distributions of $p(\eta_t|\mathbf{Z}_{1:t})$, $p(\eta_t|\mathbf{Z}_{1:T})$ can be estimated as

$$\begin{aligned} \hat{p}(\eta_t|\mathbf{Z}_{1:t}) &= \frac{1}{\mathcal{Z}_{1:t}}\hat{\alpha}_t(\eta_t), \\ \hat{p}(\eta_t|\mathbf{Z}_{1:T}) &= \frac{1}{\mathcal{Z}_{1:T}}\hat{\alpha}_t(\eta_t)\hat{\beta}_t(\eta_t), \end{aligned} \quad (37)$$

where $\mathcal{Z}_{1:t}$ and $\mathcal{Z}_{1:T}$ are the normalization factors. The mean and variance-covariance matrix $\eta_{t|T}$ and $\mathbf{P}_{t|T}$ can be estimated readily from (37).

C. EP-Based Estimation of $\xi_{t|T}$ and $\mathbf{R}_{t|T}$

As illustrated in the above Phase II, this subsection focuses on the EP-based Gaussian approximation of the posterior $p(\xi_t|\mathbf{Z}_{1:t})$:

$$p(\xi_t|\mathbf{Z}_{1:t}) \sim_G \mathcal{N}(\xi_{t|t}, \mathbf{R}_{t|t}). \quad (38)$$

Based on the above Gaussian approximation, as well as the Gaussian approximations $p(\eta_t|\mathbf{Z}_{1:t}) \sim_G \mathcal{N}(\eta_{t|t}, \mathbf{P}_{t|t})$ and $p(\eta_t|\mathbf{Z}_{1:T}) \sim_G \mathcal{N}(\eta_{t|T}, \mathbf{P}_{t|T})$ conducted in the subsection IV-B, the parameters $\xi_{t|T}$ and $\mathbf{R}_{t|T}$ can be conveniently estimated by the regular STRE model as shown in [2].

The joint distribution $\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ comprises a product of factors in the form

$$\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t}) = \prod_{n=1}^{N_t} \{p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn})p(\xi_{tn})\}\, \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}). \quad (39)$$

We approximate $\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ as a product of factors

$$q(\xi_t, \eta_t) = \prod_{n=1}^{N_t} \left\{ q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})p(\xi_{tn}) \right\} \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}), \quad (40)$$

where $p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn})$ is approximated by the Gaussian function

$$q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}) \sim \mathcal{N}(\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}), \quad (41)$$

and $\hat{\mu}_{tn}$ and $\hat{\mathbf{\Sigma}}_{tn}$ are unknown parameters to be estimated. Notice that, given the estimated $\hat{p}(\eta_t|\mathbf{Z}_{1:t-1})$, Equation (24) indicates that the sets of variables $\{\xi_t, \eta_t\}$ and $\{\xi_s, \eta_s\}$ are independent when $t \neq s$. Different from the EP algorithm in Section IV-B, which needs to propagate the messages backward and forward to the variables at different time stamps, the EP algorithm for estimating $\hat{q}(\xi_t, \eta_t)$ can be conducted separately for different time stamps. The detailed EP algorithm for estimating $p(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ can be described as follows:

1) Estimate the approximate factors $\hat{p}(\eta_{t-1}|\mathbf{Z}_{1:t-1})$ by the EP algorithm proposed in Section IV-B. Estimate $\hat{p}(\eta_t|\mathbf{Z}_{1:t-1})$ by Equation (23).
2) Initialize the factors $q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})$, $n = 1, \cdots, N_t$, by setting $\hat{\mu}_{tn} = [\mathbf{0}]$ and
$$\hat{\mathbf{\Sigma}}_{tn} = \left| \begin{array}{cc} 1 & -\mathbf{S}_{tn}^T \\ -\mathbf{S}_{tn} & \mathbf{S}_{tn}\mathbf{S}_{tn}^T \end{array} \right| \sigma_\xi^2.$$
3) Until convergence (iterate on $n = 1, \cdots, N_t$):
   a) Remove the factor $q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})$ from $q(\xi_t, \eta_t)$ by division
   $$q^{\setminus n}(\xi_t, \eta_t) \propto \frac{q(\xi_t, \eta_t)}{q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})}. \quad (42)$$
   b) Estimate the new posterior $q^{new}(\xi_t, \eta_t)$ by matching the first and second moments of
   $$q^{\setminus n}(\xi_t, \eta_t)p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn}).$$
   c) Update the new factor
   $$q^{new}(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}) = \frac{q^{new}(\xi_t, \eta_t)}{q^{\setminus n}(\xi_t, \eta_t)}. \quad (43)$$

Evaluating the above EP algorithm, the number of required iterations is greater than N_t, which is the size of locations at time stamp t. For each iteration, it needs to evaluate the new posterior $q^{new}(\xi_t, \eta_t)$ by setting the first and second order moments of $q^{new}(\xi_t, \eta_t)$ equal to those of $q^{\setminus n}(\xi_t, \eta_t)p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn})$. An efficient strategy is to explore the special structure of the factorized forms (39) and (40). The dependency between ξ_{tn} and $\{\xi_{ts}, s \neq n\}$ is realized only through η_t, and the joint distribution of η_t and $\{\xi_{ts}, s \neq n\}$ is Gaussian. Hence, we are able to obtain the analytical form ($\tilde{q}_n(\xi_{tn}, \eta_t)$) by marginalization over $\{\xi_{ts}, s \neq n\}$. The factor $\tilde{q}(\xi_{tn}, \eta_t)$ can be efficiently approximated as a Gaussian form $\tilde{f}(\xi_{tn}, \eta_t)$ by matching the first and second order moments using iterative reweighted least squares (IRLS) [15].

V. Experiments

This section evaluates the effectiveness and efficiency of our proposed St-RSTP prediction algorithms based on a simulation study and comprehensive experiments on two real data sets, including an Aerosol Optical Depth (AOD) data set collected by NASA and a region-wide traffic volume (TV) data set collected in the City of Bellevue, WA.

A. Experiment Design

Given the raw data, we first conducted a preprocess to generate original observations $\mathbf{Z}_{1:T}$ by cleaning the data set, converting the observations into a close-to-symmetric distribution, and selecting a study region. The second step was to estimate model parameters based on the clean data set by applying the EM estimation method proposed by [16]. The third step was to run the STRE smoothing on the clean data set to obtain the set of smoothed values $\hat{\mathbf{Y}}_{1:T}$ as the ground truth for evaluation. The fourth step was to randomly add isolated or region (cluster of) outliers into the clean data to obtain the contaminated data set $\tilde{\mathbf{Z}}_{1:T}$ (except for TV). The fifth step was to apply the STRE prediction algorithm and the proposed St-RSTP prediction algorithm to estimate $\mathbf{Y}_{1:T}^{(s)}$ and $\mathbf{Y}_{1:T}^{(sr)}$, respectively. The final step was to calculate the mean absolute percentage error (MAPE) and root mean squared error (RMSE) by comparing $\mathbf{Y}_{1:T}^{(s)}$ and $\mathbf{Y}_{1:T}^{(sr)}$ with $\mathbf{Y}_{1:T}$.

The superscripts (sr) and (s) of MAPE and RMSE refer to the St-RSTP processing and the STRE processing, respectively. If MAPE$^{(s)}$ (or RMSE$^{(s)}$) is larger than MAPE$^{(sr)}$ (or RMSE$^{(sr)}$), we can conclude that our proposed algorithm is more robust than the STRE algorithm.

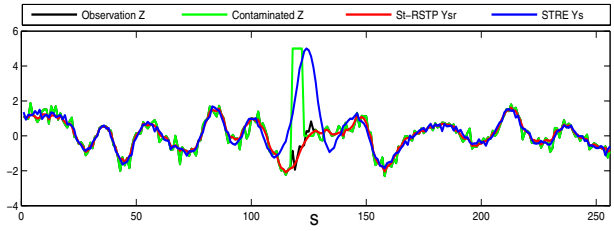
B. Simulation Study

This section presents a simulation study on the robustness of the proposed St-RSTP prediction algorithm, compared with that of the STRE algorithm. In this work, we considered the same simulation model as employed in recent STRE related papers [2], [16], to generate spatio-temporal simulation data.

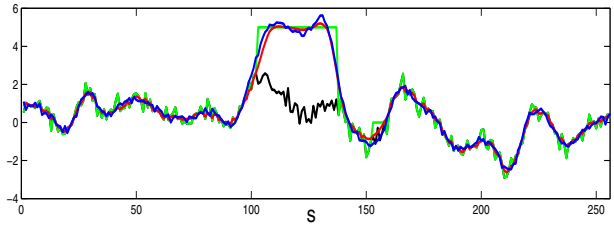
The spatial domain was designed as one dimension and had the observation locations, $D = \{s : s = 1, \cdots, 256\}$. The temporal domain had the observation timestamps $t = 1, 2, \cdots, 50$. We assumed that the trend component $\mu(\mathbf{s}; t)$ was zero and simulated the processes $Y(\mathbf{s}; t)$ and $Z(\mathbf{s}; t)$

according to Equations (2) and (3). The small scale (autoregressive) process $\{\eta_t\}$ was generated by the matrix parameters $\mathbf{H}$ and $\mathbf{U}$. The spatial basis functions $\mathbf{S}$ were defined by 30 W-wavelets from the first four resolutions.

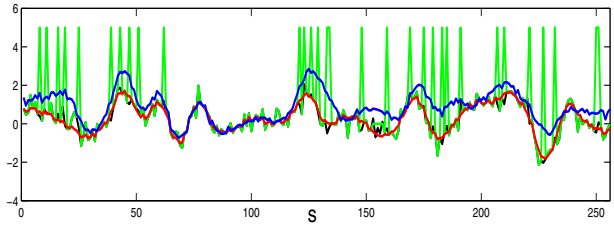
We considered two types of outliers, isolated outliers and regional (cluster of) outliers. For isolated outliers, we randomly picked locations and timestamps, and then shifted the observation to a larger value 5. We generated cases with 5, 15, and 35 random outliers. For regional outliers, we fixed the center of the region and set region sizes (number of outliers) to 5, 15, and 35. The temporal dimension of the region was fixed to a 6 units window. Note that, other combinations of the time and spatial locations had also been tested and similar patterns were observed.



(a) t = 4, 5 regional outliers



(b) t = 7, 35 regional outliers



(c) t = 44, 35 isolated outliers

Figure 3: STRE vs. St-RSTP using simulation data

*1) **Simulation Results**:* We conducted both St-RSTP and STRE smoothing, filtering, and forecasting in a variety of simulated scenarios with isolated and regional outliers. Several case studies are discussed as follows. Figure 3 illustrates the impacts of isolated and regional outliers on the filtering algorithms at three different timestamps with various number of outliers. Each sub-figure has four curves that are related to the original observations $\mathbf{Z}_t$, the contaminated observations $\tilde{\mathbf{Z}}_t$, the filtered values $\mathbf{Y}^{(s)}$ via the regular STRE algorithm, and the filtered values $\mathbf{Y}^{(sr)}$ via our proposed St-RSTP algorithm, respectively. The X-axis refers to location index, with totally 256 distinct locations. The Y-axis denotes the $\mathbf{Z}$ values. The symbol t refers to time stamp. As shown in the figures, with the increasing number of outliers, the STRE curve was clearly distorted at an increasing degree. On the contrary, our proposed robust filtering algorithm demonstrated strong resilience to outlier effects. Even in the situation of high rate contaminations (35 isolated outliers, around 13% percentage in Figure 3(c)), our proposed algorithm could still recover the latent random variables $\mathbf{Y}_t$ very well.

Figure 3 (a) and (b) illustrates the impacts of regional outliers on the two filtering algorithms with different outlier region size. When the outlier region size was small (5 adjacent outliers), our proposed robust filtering algorithm performed very well, whereas the STRE filtering algorithm was already misguided by the outliers and the filtered curve segment around the outlier region was clearly distorted. On the other hand, outside the outlier region, the filtered curve generated by St-RSTP was almost identical to the filtered curve generated from the STRE filtering algorithm. This indicates that when there are no outliers, our algorithm performs similarly as the regular STRE model, but when outliers appeared, our algorithm tends to be more resilient to the outliers.

However, we also observed that large region outliers have significant impacts on both the STRE and St-RSTP, in Figure 3 (b). When we increased the region size to 35, both St-RSTP and STRE filtering algorithms were misguided and the filtered values around the outlier region were close to outlier values. This could be interpreted by the STRE model assumptions (See Section II-A) that define spatio-temporal dependencies between $Z(\mathbf{s}_i; u)$ and $Z(\mathbf{s}_j; t)$, with $i \neq j$ or $u \neq t$. Particularly, the STRE model assumes a Markov Gaussian process to model spatial dependencies between $Z(\mathbf{s}_i; t)$ and $Z(\mathbf{s}_j; t)$, $i \neq j$. Observations will have a high spatial correlation if they are spatially close. For temporal dependency, the STRE model assumes a first order Markov process. That is, except for the dependence on the other locations at the current time t, $Z(\mathbf{s}; t)$ is also dependent on its previous time stamp observations $\mathbf{Z}_{t-1}$. To conclude, the STRE model considers spatial Gaussian process, log-1 temporal autocorrelation, and white noise (Gaussian distribution) to model the whole data variation. Spatio-temporal outliers can be interpreted as the observations that have low correlations with their spatio-temporal neighbors and can not be regarded as the normal measurement error (white noise). When a data set has outliers, for the standard STRE model the additional variations due to outliers will be captured by distorting the spatio-temporal dependencies. The white noise component can not handle large deviations due to the non-heavy tail distribution characteristics. This explains the distorted STRE curves as shown in Figures 3. A specific spatio-temporal autocorrelation pattern is associated with certain degree of sharpness of the resulting filtered curves. In comparison, our St-RSTP model uses Student's

Table I: Model Robustness Comparison using Different Simulation Settings

Outlier Type	Size	MAPE$^{(sr)}$ (O)	MAPE$^{(s)}$ (O)	RMSE$^{(sr)}$ (O)	RMSE$^{(s)}$ (O)	MAPE$^{(sr)}$ (R)	MAPE$^{(s)}$ (R)	RMSE$^{(sr)}$ (R)	RMSE$^{(s)}$ (R)
Isolated Outliers	5	1.2554	2.1253	0.3534	0.7105	6.5712	10.656	0.2468	0.3341
	15	1.3436	4.8988	0.3313	0.8400	6.6204	20.061	0.2466	0.3575
	35	1.6939	7.7223	0.3497	1.2457	6.7262	11.337	0.2498	0.4085
Regional Outliers	5	2.1965	14.047	0.5454	3.4465	6.6423	11.050	0.2536	0.3938
	35	132.14	138.94	4.5852	4.7571	7.2288	10.824	0.2537	0.3301

Table II: Model Robustness Comparison use the AOD Data

	MAPE (O)	MAPE (R)	MAPE (A)	RMSE (O)	RMSE (R)	RMSE (A)
STRE	3.3475	3.9940	3.9682	0.4309	0.3799	0.3821
St-RSTP	2.2176	2.1619	2.1641	0.3322	0.3244	0.3247
improve	33.8%	45.9%	45.5%	22.9%	14.6%	15.0%

t distributions to model white noise (or the measurement error). When outliers appear, our St-RSTP model directly captures the additional large variations due to outliers as white noise. When the outlier region becomes large, however, it becomes possible to directly use the spatio-temporal autocorrelations to capture the outlier variations. Intuitively, we are able to use a smooth curve to fit the observations well. This potentially explains why the St-RSTP model could not recover the true Y values around the outlier region, when the outlier region size was large.

Table I illustrates the robustness of the filtering algorithms based on different settings of outliers. In this table, (O) refers to outliers, and (R) refers to non-outliers. It can be observed that St-RSTP algorithm always outperformed the STRE filtering algorithm in all the scenarios we have experimented. Although we observed the similar results for 1-step forecasting, we only present the forecasting results for the real data sets due to the space limit.

C. *Aerosol Optical Depth Data Experiments*

The AOD data set was collected by NASA's Terra satellite with MISR (Multi-angle Imaging Spectro Radiometer) on board. Because the AOD data are heavily right-skewed, we applied log transformation $\log(AOD)$ to convert the 40-day level-3 data (with spatial resolution ($0.5^\circ \times 0.5^\circ$) and temporal resolution (1 day)) into a close-to symmetric distribution. Each time unit is defined as an exclusive eight-day period. We focus on the data collected in a rectangle region D between longitudes 14° and 46° and between latitudes 14° and 30°, as shown in Figure 4(a). The number of level-3 observations (pixels) in the region is $32 \times 64 = 2048$. Other geographical regions had also been studied and similar patterns were obtained.

In order to evaluate the robustness of different filtering and forecasting algorithms on the AOD data, we randomly set 5% locations in every timestamp and replaced the observations with value 5, which is outside the normal range of the observations (-0.0843 ± 0.4958).

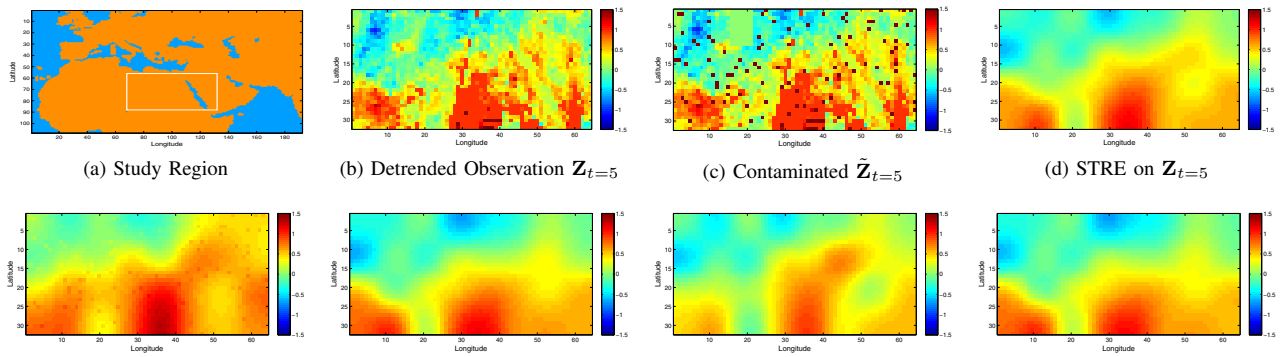
A similar STRE model specification as used in [2] was applied in this simulation. We detrended the observations $\mathbf{Z}_t$ by the residuals $\mathbf{Z}_t - \mathbf{X}_t\beta$ to $\mathbf{Z}_t$. After this process, the observations $\mathbf{Z}$ no longer had trend components and could be called as detrended observations. The unknown parameters σ_ε^2, $\mathbf{K}_1$, and $\{\mathbf{H}_t, \mathbf{U}_t\}, t = 1, \cdots, 5$, in basis functions $\mathbf{S}$ were estimated by using the EM estimation algorithm proposed by [16].

Figures 4 illustrate the robustness of our St-RSTP filtering and forecasting algorithms compared with that of the regular STRE algorithms at timestamp $t = 5$. Figure 4(a) shows our study region, which was within the white box on the map. Figure 4(b) shows the heatmap of the detrended observations $\mathbf{Z}_{t=5}$. Figure 4(c) displays the contaminated observations $\tilde{\mathbf{Z}}_{t=5}$, in which we injected an red-color outlier dots in the image. Figure 4(d) shows the STRE filtering results on the clean detrended observations $\mathbf{Z}_{t=5}$, and Figure 4(e) displays the STRE filtering results on the contaminated observations $\tilde{\mathbf{Z}}_{t=5}$. Figure 4(f) shows the St-RSTP filtering results on $\tilde{\mathbf{Z}}_{t=5}$. By comparing Figure 4(e) and (f) with the original filtering results shown in Figure 4(d), we can observe that the regular STRE filtering results were clearly distorted by the region outliers round the neighborhood area. However, our St-RSTP filtering results in Figure 4(f) were still very close to the original filtering results in Figure 4(e). Similarly, the 1-step forecasting results in Figure 4(g) and 4(h) showed that the St-RSTP produced more accurate prediction than the STRE.

To demonstrate the results in a more comprehensive way, Table II presents the average results on all the five time units, where (O) refers to outlier region, (R) refers to non-outlier region, and (A) refers to all the region. It can be clearly observed that the St-RSTP achieved much lower MAPE and RMSE than the STRE filtering algorithm in both outlier and nonoutlier regions.

D. *Case Study on Traffic Volume Data*

The traffic volume data were collected in the City of Bellevue, WA. The data was managed by the Smart Transportation Application and Research Laboratory (STAR Lab) at the University of Washington, Seattle. In this set of experiments, 17 detectors located in NE 8th Ave was selected as the test route because it's a major city corridor, with annual average weekday traffic of 37,700 (veh/day). Weekday data (Tuesday, Wednesday and Thursday) collected from first two weeks of July, 2007 were used for training and the last two

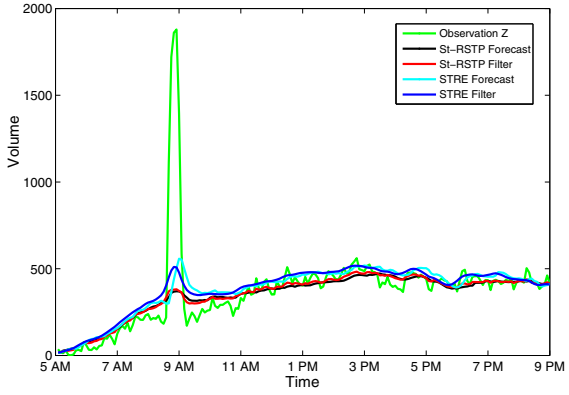


(a) Study Region (b) Detrended Observation $\mathbf{Z}_{t=5}$ (c) Contaminated $\tilde{\mathbf{Z}}_{t=5}$ (d) STRE on $\mathbf{Z}_{t=5}$

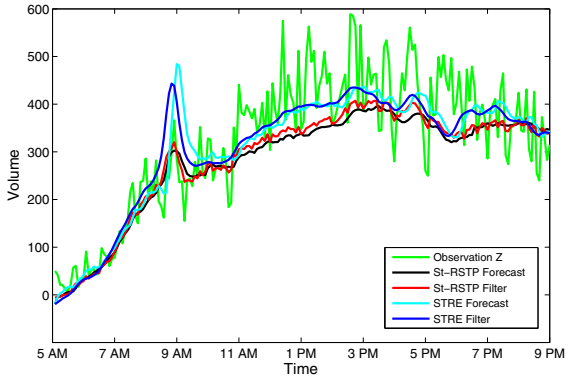
(e) STRE Filter on $\tilde{\mathbf{Z}}_{t=5}$ (f) St-RSTP Filter on $\tilde{\mathbf{Z}}_{t=5}$ (g) STRE Forecast on $\tilde{\mathbf{Z}}_{t=5}$ (h) St-RSTP Forecast on $\tilde{\mathbf{Z}}_{t=5}$

Figure 4: STRE vs. St-RSTP on AOD data sets at time unit 5

weeks of June, 2007 were used for cross validation. The verification data were collected during the first week of July in 2008. In this study, all data were aggregated into 5-minute intervals to reduce the effect of random noise. In total, the detector data collected on on 17 detectors within 5376 time intervals were evaluated.



(a) t = 5th day, detector #3



(b) t = 5th day, detector #16

Figure 5: STRE vs. St-RSTP using the TV data on 5th day

Figure 5 shows the comparison results on two detectors with different real-world outlier rates. The X-axis refers to the 192 timestamps from 5 am to 9 pm, and the Y-axis refers to the traffic volume, aggregated at 5 minute intervals. Figure 5(a) shows the traffic volume from detector #3 with one significant spike reached 1900 around 9 am, which was probably caused by malfunctioning. On this detector, the STRE filtering algorithm had a spike over 500 triggered by the outlier, and its 1-step forecasting had a even higher spike right after the real one. On the other hand, the St-RSTP smoothed the spike to around 300, which is closer to their spatial neighbors. The St-RSTP 1-step prediction produced the volumes very similar to its smoothed curve. Figure 5(b) shows the results on detector 16 with vibrating volumes throughout the day. Because this detector was located close to detector #3 on the same route, the outlier on detector #3 affected the STRE process on detector #16. As can be observed from the figure, the STRE approach had a significant spike on the filtering curve at exactly the same time when the outlier appeared on detector #3; and a higher spike on the forecasting curve right after the outlier appeared. On the contrary, although the St-RSTP did filtering and forecasting by considering spatial and temporal neighbors as well, its process successfully resisted the impact from the spatially neighboring outlier. Besides that, one can also notice that the St-RSTP handled the vibrations on the original volume more smoothly than the STRE. More specifically, the St-RSTP forecasting gave smoother volumes than its filtering. This suggested that both St-RSTP filtering and forecasting are robust on the temporal domain. These patterns are consistent with what we observed from the simulation study and the AOD results.

E. Time Cost

Table III presents the execution time comparisons between our St-RSTP model and regular STRE model. The comparisons are under Windows 7 Professional 64-bit operating system, Intel core i7-Q740, 1.73GHz (CPU), 8.00 GB (RAM). We compare all the scenarios in simulation data and the whole set in AOD data. The result shows that the St-RSTP can reach ten times in execution time comparing to that of STRE algorithms under all tested simulation data scenarios. But in the AOD dataset, St-RSTP outperformed

the regular STRE algorithms in the all 5 time units. Our St-RSTP algorithm estimated small-scale and micro-scale variation separately. The estimation went through all the timestamps one by one, so it would cost less time and outperform STRE in a dataset with fewer timestamps.

Table III: Comparison of Time Cost using the Simulated and AOD Data

Dataset		Outliers (#)	STRE (Sec)	St-RSTP (Sec)
Simulation Data	Isolated Outliers	5	2.95	29.10
		15	3.03	29.19
		35	3.14	29.64
	Regional Outliers	5	2.72	28.28
		15	2.87	28.54
		35	2.88	28.28
AOD Data		5%	69.07	26.58

Note: The simulated data has 256 locations and 50 time units. The AOD data has 2048 locations and 5 time units.

On the other hand, time costs of St-RSTP and STRE on the simulation data with various location sizes are illustrated in Figure 6, where the X-axis shows the number of locations in log scale, and the Y-axis represents the execution time in seconds. As can be clearly observed, both St-RSTP and STRE had increased time costs when the number of locations grew up. Although the St-RSTP took longer to execute when the number of locations changed from 32 to 1024, the St-RSTP has shown better scalability than the STRE as the time differences reduced from tens of times to about 30%.

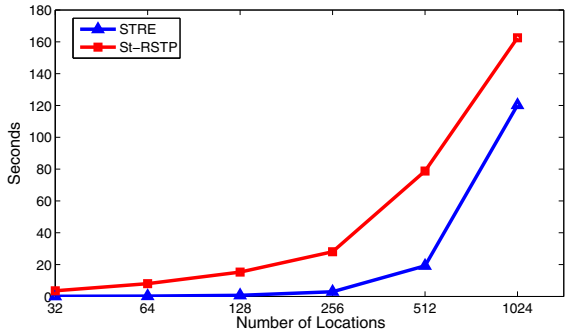


Figure 6: Time Cost vs. Number of Locations

VI. Conclusion

This paper proposes a robust and effective design of spatio-temporal prediction based on Student's t distribution, St-RSTP. This prediction model inherits the ability of processing large scale spatio-temporal data with linear time complexity from STRE, and provides enhanced tolerance to outliers or other small departures. An approximate inference approach in the framework of Expectation Propagation is proposed to support the analytical intractable inference of Student's t model in near linear time. The robustness and the efficiency of our Student-t based prediction model have been demonstrated in extensive experiments evaluations based on both simulation and real-life data sets. The proposed approach provides critical functionality for stochastic processes on spatio-temporal data.

Acknowledgment

The authors would like to thank the City of Bellevue, Washington for providing arterial traffic data. The authors are also grateful to the STAR Lab for maintaining the arterial database and providing the online portal to access the data.

References

[1] N. Cressie and C. Wikle, *Statistics for Spatio-Temporal Data*. Wiley, 2011, iSBN 978-0471692744.

[2] N. Cressie, T. Shi, and E. L.Kang, "Fixed rank filtering for spatial-temporal data," *Journal of Computational and Graphical Statistics*, vol. 19, no. 3, pp. 724–745, 2010.

[3] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications With R Examples*. Springer, 2006.

[4] I. A. J. B. Juha K. Laurila, Daniel Gatica-Perez and O. Bornet, "The mobile data challenge: Big data for mobile computing research," *In Proc. Mobile Data Challenge by Nokia Workshop, in conj. with Int. Conf.. on Pervasive Computing*, 2012.

[5] N. Cressie and C. Wikle, "Space-time kalman filter," *Encyclopedia of Environmetrics*, vol. 4, pp. 2045–2049, 2002.

[6] M. Gandhi and L. Mili, "Robust kalman filter based on a generalized maximum-likelihood-type estimator," *IEEE Transactions on Signal Processing*, vol. 58, pp. 2509–2520, 2010.

[7] Y. Ruan and P. Willett, "Practical fusion of quantized measurements via particle filtering," *IEEE Aerosp. Conf.*, 2003.

[8] R. Maronna, R. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*. John Wiley Sons, Ltd, 2006.

[9] J. Durbin and S. J. Koopman, "Monte carlo maximum likelihood estimation for non-gaussian state space models," *Biometrika,*, vol. 84, pp. 669–684, 1997.

[10] P. Jylanki, J. Vanhatalo, and A. Vehtari, "Gaussian process regression with a student-t likelihood," *Journal of Machine Learning Research*, p. Accept for Publication, 2011.

[11] C. M. Bishop and M. Svensen, "Robust bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[12] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, "An 1 -laplace robust kalman smoother." *IEEE Trans. Automat. Contr.*, vol. 56, no. 12, pp. 2898–2911, 2011.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[14] A. Ypma and T. Heskes, "Novel approximations for inference in nonlinear dynamical systems using expectation propagation." *Neurocomputing*, vol. 69, pp. 85–99, 2005.

[15] P. J. Green, "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives," *J ROY STAT SOC B*, vol. 46, no. 2, 1984.

[16] M. Katzfuss and N. Cressie, "Spatio-temporal smoothing and em estimation for massive remote-sensing data sets," *Journal of Time Series Analysis*, vol. 32, no. 4, pp. 430–446, 2010.

[17] V J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, 2004.