This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

Multi-Task Bayesian Spatiotemporal Gaussian Processes for Short-term Load Forecasting

Mostafa Gilanifar, *Student Member, IEEE,* Hui Wang, Lalitha Madhavi Konila Sriram, *Student Member, IEEE,* Eren Erman Ozguven, *Member, IEEE,* Reza Arghandeh, *Senior Member, IEEE,*

***Abstract*—Short-term household electricity load forecasting is important for utility companies to ensure reliable power supplies. Traditional methods for load forecasting relied on historical records from one single data source and have limitations with insufficient or missing data. Recently, an emerging family of machine learning algorithms, multi-task learning (MTL), have been developed and have the potential for load forecasting. By MTL, the electricity consumption data from multiple communities can be fused to improve forecasting accuracy. However, appropriate modeling of the relatedness to enable the between-community knowledge transfer remains a challenge. This paper proposes an improved MTL algorithm for a Bayesian spatiotemporal Gaussian process model (BSGP) to characterize the relatedness among the different residential communities. It hypothesizes on the similar impacts of environmental and traffic conditions on electricity consumption in improving short-term load forecasting. Furthermore, the paper proposes a low ranked dirty model (LRDM) along with an iterative algorithm to improve the learning of model parameters under an MTL framework. This paper used a real-world case study from two residential communities in Tallahassee, Florida, to demonstrate the method effectiveness. The proposed method significantly outperforms state-of-the-art forecasting methods and effectively capture the relatedness to provide between-community knowledge transfer compared with other MTL methods.**

***Index Terms*—Gaussian Process, Low Ranked Model, Multi-task Learning, Electric Load Forecasting, Power Distribution, Transportation Network.**

NOMENCLATURE

ABBREVIATIONS

BSGP	Bayesian spatiotemporal Gaussian Process
GP	Gaussian Process
LRDM	Low ranked dirty model
MT-BSGP	Multi-task Bayesian spatiotemporal Gaussian Process
MTL	Multi-task learning
STL	Single-task learning

LIST OF VARIABLES

$\boldsymbol{B}$	Basis vectors
$\boldsymbol{P}$	Group sparse component
$\boldsymbol{Q}$	Low-rank component
$\boldsymbol{\beta}$	Coefficients of input variables for all communities
η	Correlated stochastic process
λ	Coefficients of the norms
μ	Mean shift or trending patterns
ϕ	Decay parameter at each time stamp
σ_{ϵ}^2	Variance of the noise
σ_{GP}^2	House-invariant common variance at each time stamp
τ	Number of time points
H	Humidity
l	Index for the community
M	Number of houses
p	Time lags
S	Solar radiation
T	Temperature
Tr	Traffic counts near the residential community
Z	Number of communities
ϵ	Noise
$\vec{\beta}$	Coefficients of input variables
E	Electricity consumption
s	Index for houses
t	Time
X	Input variables

Manuscript received March 12, 2019; revised May 17, 2019; accepted June 17, 2019.

M. Gilanifar was a Ph.D. candidate of IME department at Florida State University when the work was completed and is currently a postdoctoral research associate in ECE department at the University of Utah; L.M.K. Sriram is a Ph.D. candidate with the ECE Department, Florida State University.

H. Wang and E.E. Ozguven are Assistant Professors in the Departments of IME and CEE, Florida State University; R. Arghandeh is a Professor in Western Norway University of Applied Science, Norway.

This research is partly supported by U.S. NSF 1640587, CMMI-1744131, and HRD-1646897.

I. INTRODUCTION

LOAD forecasting is essential for balancing the power supply and demand to avoid instabilities in the grid. The short-term load forecasting focuses on forecasting the electrical load for one hour to one week can guide utility company and power plants to regulate the electricity generation to meet market demands by producing as-needed energy. For example, it was reported that a 1% reduction in forecasting error reduced 10 million pounds in the operating cost per year for one utility in the U.K. [1]. The forecasting also allows the utilities to adopt dynamic pricing schemes in electricity markets. However, load forecasting is still challenging due to the complexity of the electric grid and uncertainty in the electricity consumption [1]. Improving the load forecasting has been possible due to the implementation of advanced data acquisition systems such as smart meters along with the development of new data analytic techniques. Recently, machine learning-based methods achieve more attention in the load forecasting. For example, the autoregressive integrated moving average (ARIMA) models are among the most used techniques, as shown in [2]-[3]. Literature shows that the multilinear and Gaussian Process regression are proper approaches for load forecasting [4]. Some researchers utilized fuzzy logic for the short-term load forecasting [5]. In the field of machine learning, support vector regression (SVR) [6], artificial neural

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

networks (ANN) [7]-[8], and deep neural networks (DNN) [9] are highly used methods of load forecasting.

It is worthwhile mentioning that the majority of available machine learning based load forecasting works including the ones that are mentioned so far fall under the category of single-task learning (STL) methods, in which one model is trained over one data source. Usually, in the STL methods, the forecasting performance is affected by missing data or erroneous measurements over a certain period or the low measurement resolution.

Recently, a new opportunity emerges to improve the load forecasting by fusing the electricity consumption data from multiple data sources in the multi-task learning (MTL) framework, in which multiple models are jointly learned over multiple data sources. There exist a small number of papers using the MTL framework for load forecasting. Zhang et al. [10] focused on the utility-level load forecasting using multi-task Gaussian Process. More recently, Fiot and Dinuzzo [11] used a kernel-based MTL method called Low-Rank Output Kernel Learning (LR-OKL) for mid-term load forecasting at the distribution substation level. In the field of machine learning, a regularization-based strategy is commonly adopted to extract the similar knowledge or variation patterns among multiple data sources to improve learning performance (as called relatedness). The method introduces a certain regularization term, which applies a weighted penalty term to the learning of objective function such as the minimization of the least square of forecasting errors [12]-[14]. MTL has also been developed to deal with the MTL of Gaussian Process (GP) models [15]. Although the prior research has demonstrated the potential of the MTL methods, finding the relatedness among the data sets from multiple tasks is still a huge challenge. A summary of the load forecasting methods with their strengths and weaknesses is shown in Table I.

Regarding the application point of view, a significant body of load forecasting studies use only historical electricity consumption data to perform forecasting [16]. Later on, electricity load forecasting studies went beyond the methods that only consider historical electricity consumption data and included weather conditions to increase the accuracy [17]-[20]. The emergence of the system of systems and multi-network theories plus the advent of "smart city" concept, encouraged new studies that consider the interdependency and interconnectivity of electricity networks to the other infrastructure networks [21]. The model postulation in this paper is inspired by [2] and [22] that presented a causal inference framework to characterize the relationship between traffic load and electricity consumption in power distribution networks. The [2], [22], [23] performed combined electricity and traffic load forecasting and showed that adding traffic data and mobility information as a predictor improves the accuracy of electricity load forecasting. Therefore, in this paper, the combination of electricity consumption data, traffic data, and weather parameters is adopted as the predictor for short-term electricity load forecasting. In the context of smart city solutions, there are some recent works on the household's electricity load forecasting, e.g., [24]-[25]. These methods are under the STL framework, which is more sensitive to missing data or measurement resolution.

This paper aims to improve the short-term load forecasting based on multi-task learning (MTL) framework, in which one "task" refers to the learning of load forecasting model for each residential community, and "multi-task" means that the load forecasting models for multiple communities are jointly learned by fusing the data from these communities. Specifically, the training data from multiple residential communities in a city under similar conditions or setups are fused to jointly learn an inter-community relatedness by exploring similarities in the data patterns across communities. As such, it is feasible to utilize the relatedness concept to characterize the similar effects of environmental and traffic conditions on the electricity consumptions for multiple residential communities. In the mornings when residents leave for work, there are heavier traffic counts near the residential communities, resulting in less household electricity consumptions until the afternoon when people return home. Furthermore, the electricity consumptions vary with the ambient temperature due to air conditioning loads. The effects of environmental and traffic data on electricity consumptions share similarities across different residential communities in a city. In addition, there may exist some community-specific information such as local spatiotemporal variations within a community that is not shared across communities. The **contributions** of this paper is summarized below by addressing the following **research gaps**, i.e.,

Table I
LOAD FORECASTING METHODS

Paper	Method	Type	Strengths	Weaknesses
[2]-[3]	ARIMA	Single-task	Solid underlying theory, Usable for different time-series	Needs enough data
[4]	Multilinear regression	Single-task	straightforward to understand	performs poorly with non-linear relationship and lack of data
[5]	Fuzzy Logic	Single-task	Interpretability and simplicity	non-robustness and arbitrary of rules
[6]	SVR	Single-task	Model non-linear decision boundaries, have many kernels	Memory intensive, tricky to tune
[7]-[9]	NN	Single-task	Detect nonlinear relationships, easily update and adapt with new data	Determination of proper network structure, computationally intensive to train
[10]	MTL of GP	Multi-task	Performs well with small data, straightforward to use	Merging the shared and non-shared information, Gaussian (kernel) assumption
[11]	LR-OKL	Multi-task	Performs well with small data	Dependent on the kernel, non-separating shared and non-shared information

1) From the **methodological** point of view: (a) The traditional MTL methods, including regularization-based methods [12]-[14] and kernel-based methods [11],[15], do not distinguish between the general trend, which can be commonly shared across communities, and the local spatiotemporal variations; (b) The regularization-based MTL methods [12]-[14] characterize the relatedness by introducing some regularization terms in their objective function based on either a "common set of features" or "shared low-rank" structures, which only partially capture the relatedness among different tasks. Therefore, this paper **proposes** a Low Ranked Dirty Model (LRDM) for the Multi-Task Bayesian Spatiotemporal Gaussian Process (MT-BSGP) to capture more inter-community relatedness using both a "common set of features" and a "shared low-rank structure" simulta-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

neously. Furthermore, this paper **proposes** an iterative algorithm to simultaneously estimate parameters of the trend model and Gaussian process.

2) From the **application** point of view, there is a lack of research on exploring the value of using similar effects of time-varying environmental and traffic conditions on the electricity consumption to model the inter-community relatedness for short-term load forecasting. This paper **proposes** a load forecasting model solved by MTL through characterizing such similar effects as the inter-community relatedness.

The remainder of this paper is organized as follows. Section II explains the proposed MT-BSGP and learning algorithm in detail. Section III describes a real-world case study based on load forecasting for the City of Tallahassee. Section IV discusses the results, and Section V concludes the paper.

II. Multi-Task Bayesian Spatiotemporal Gaussian Process (MT-BSGP)

This section proposes an MT-BSGP for load forecasting. The structure of a forecasting model fusing environmental and traffic data for a community is presented in Section II-A, and Section II-B further formulates a learning problem for estimating the MT-BSGP. Section II-C overviews existing regularization-based MTL methods. In Section II-D, the LRDM is developed to improve the learning performance for MT-BSGP. The LRDM is implemented by an iterative algorithm to learn the model parameters under the MTL framework in Section II-E.

A. Overview of the Forecasting Model For a community

The electricity consumption (E) can be forecasted for a community as follows:

$$E = \mu + \eta + \epsilon, \tag{1}$$

where μ captures the mean shift or trending patterns in electricity consumption data and η characterizes a correlated stochastic process in the data, which reflect a system-level correlation among the data that can be used to improve the forecasting accuracy.

There are a number of ways to expand η. One common method is to cluster the consumers according to their electricity consumption patterns by grouping similar or correlated behaviors within one cluster. As such, η can be expanded as a predictor based on the consumers' data within the same cluster [26]. More recently, literature have reported that η can be potentially characterized by a spatially correlated process if consumers exhibit spatially clustered patterns. The reasons include (1) the topological layout of the electricity networks and feeders' laterals connectivity. The loads at nearby feeders are likely to be more similar than those that are farther apart [27] and (2) the demographic and topographic characteristics of residential neighborhoods including land use, type of buildings, size of buildings, landscape design, the income level of households, leading to a spatially correlated process in the household electricity consumption patterns. For example, most of the buildings in a neighborhood follow similar designs and footprints, and were built by the same developer and similar construction materials [28] that make building to have similar insulation, building envelope, etc. which leads to similar electricity consumption. The correlation in the consumer activities further leads to statistically correlated electricity consumption behaviors [27] since the consumer activities as a result of similar environments and traffic/road conditions in the adjacent neighborhood tend to be correlated.

The aforementioned reasons reflect system-level causes in the human-power system contributing to the spatial dependencies, which can be characterized by a combination of global trend, a spatially correlated process (usually by Gaussian process), and independent variation [29]. Thus, the model above can be simplified into

$$E = \mu + GP + \epsilon, \tag{2}$$

which is a Bayesian spatiotemporal Gaussian Process (BSGP) model. One motivation for adopting this formulation is that we may have to forecast the electricity load for a household (location) with limited historical data or even without data to predict. Most of the forecasting methods in the literature need a certain number of historical recordings for the household (location) of interest. BSGP model can perform forecasting for the household without sufficient historical data by leveraging the information from its neighborhood.

The mean μ is commonly assumed as a linear model of variables that are correlated to the electricity consumption. Denote s as the index for the location of each house and t for time then, the μ can be expanded by: $\mu(s,t) = \sum_i \sum_j X_{ij}(s,t)\beta_j$. The model postulation in this paper is also inspired by the cause and effect relationship between transportation and electricity networks provided in [2] and [22] by using these two variables for load forecasting. Therefore, in this paper, X can include the electricity consumption data, traffic near the residential community (Tr), and weather parameters such as temperature (T), humidity (H), and solar radiation (S).

Furthermore, the socio-economic factors such as the number of households and income level can also play an influential role in load forecasting and can be potentially included in the forecasting model. When such input data can become available, they can be included in the model as a linear term. Furthermore, the living standards can have an effect on the average power consumption as well as the temporal fluctuations in power consumption patterns. Thus, the average living standard can also be included in the model postulation. However, the living standards for all the households in the communities in the short-term horizons are relatively constant over time. Therefore, it does not affect the performance of the proposed load forecasting method.

The vector $\vec{\beta}$ includes the coefficients of input variables. For example, if there are M houses and τ different time points, the μ can be expanded as follows, i.e.,

$$\begin{aligned}\mu(s,t) = \sum_{i=1}^{n} \Bigg[\Bigg(\sum_{j=1}^{p} \beta_j E_i(s,t-p) \Bigg) + \beta_{p+1} T_i(s,t) \\ + \beta_{p+2} H_i(s,t) + \beta_{p+3} S_i(s,t) + \beta_{p+4} Tr_i(s,t) \Bigg], \\ s = 1, 2, \cdots, M, \qquad t = 1, 2, \cdots, \tau,\end{aligned} \tag{3}$$

where n is the number of observations; the time lag p in the time series model in Eq. 3 is determined by the Autocorrelation Function (ACF), and Partial Autocorrelation Function

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

(PACF) to find how many previous observations are needed in the model; and $p+4$ is the number of variables. Moreover, in Eq. 3, $\vec{\beta} = [\beta_1, \cdots, \beta_{p+4}]$ reflects the impacts of input variables on electricity consumption.

The term "GP" in Eq. 2 is normally distributed with a zero mean and variance-covariance matrix of Σ_{GP} as $\Sigma_{GP} = (\sigma^2_{GP}|t) exp(-(\phi|t)||s_i - s_j||^2), \phi > 0$, where Σ_{GP} is characterized by a house-invariant common variance at each time stamp ($\sigma^2_{GP}|t$) and a spatial correlation function ($\kappa(s_i, s_j; \phi)$). A squared exponentially correlated function is usually chosen for $\kappa(s_i, s_j; \phi)$, which includes a decay parameter at each time stamp ($\phi|t$) and the squared distance between two houses ($||s_i - s_j||^2$). For more details, please refer to [30]. Furthermore, the term ϵ in Eq. 1 is noise, which is independent and identically distributed (i.i.d.) of a normal distribution with a zero mean and a variance of σ^2_ϵ.

In summary, the model postulation utilized the existing conclusion from prior research [2], [22]. This paper focuses on formulating a multi-task learning problem for load forecasting based on this model and developing an effective algorithm to solve the problem with improved accuracy as explained in the following subsections.

Based on the model postulation, the overall framework of the proposed MT-BSGP is illustrated in Fig.1. The MT-BSGP includes the proposed LRDM for estimating overall mean shift or trending patterns that are similar across communities, a Gaussian Process that is community-specific and captures the local variation and spatial dependency, and an iterative estimation procedure between the LRDM and GP as indicated by the circular arrow. The following three subsections describe the procedures in more details.

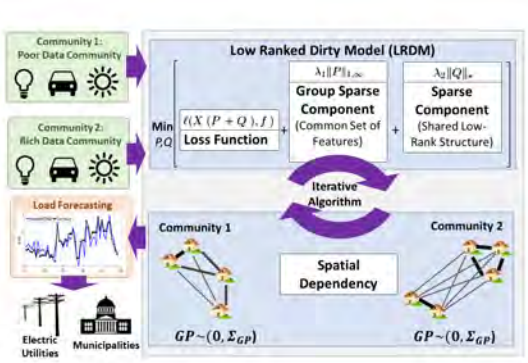


Fig. 1. Schematic overview of the proposed MT-BSGP

B. The Proposed Multi-Task BSGP Structure

The estimation of BSGP can be improved by fusing the data from multiple communities under an MTL framework. This subsection presents a modeling structure of the MT-BSGP by exploring the relatedness across different communities. The inter-community relatedness is reflected by the similar relationships between environmental factors/ traffic counts and electricity consumptions ($\vec{\beta}$). Therefore, the coefficients $\vec{\beta}$ in the BSGP (in μ) for different communities can be "similarly related" under the MTL framework. Mathematical characterization of such relatedness will be presented in Section C. Also, the GP term captures the spatiotemporal dependencies among household electricity consumption data that are community-specific. To characterize the spatial dependency, Gaussian Process has been employed by prior research [31] for load forecasting, and it provides an example platform to develop the algorithm to solve multi-task learning for load forecasting problems.

Assume that there are Z communities, and for community l, $l = 1, 2, ...Z$, the proposed MT-BSGP has the following model structure:

$$\begin{aligned} E_1 &= X_1 \overbrace{\vec{\beta_1}}^{\text{similarly related}} + GP_1 + \epsilon_1, \\ &\vdots \\ E_Z &= X_Z \quad \vec{\beta_Z} \quad + GP_Z + \epsilon_Z. \end{aligned} \tag{4}$$

The MT-BSGP is estimated by two learning objectives including (1) MTL of the "μ" and (2) learning of the "GP" separately within each community. The objective of MTL for the μ is to estimate the $\vec{\beta_1}, \cdots, \vec{\beta_Z}$ simultaneously given the data X on the environment, traffic, and historical electricity consumption from each community as well as the assumed relatedness between $\vec{\beta_1}, \cdots, \vec{\beta_Z}$. The objective for the GP is to estimate the GP parameters given the data in each community and the associated $\vec{\beta_l}$, $l = 1, \cdots, Z$, which is estimated by the MTL. The learning objectives are summarized in Eq. 5, where a hat is placed on the top of each variable to be estimated from data.

$$\begin{aligned} &\text{Obj. 1 } (\mu)\text{: } To\ Jointly\ Estimate\ \hat{\vec{\beta}}_1, \cdots, \hat{\vec{\beta}}_Z | (X_1, \cdots, X_Z) \\ &\text{Obj. 2 } (GP)\text{: } To\ Estimate \begin{cases} \text{Community 1: } \hat{\sigma}^2_{GP_1}, \hat{\phi}_1 | (X_1, \hat{\vec{\beta}}_1) \\ \vdots \\ \text{Community Z: } \hat{\sigma}^2_{GP_Z}, \hat{\phi}_Z | (X_Z, \hat{\vec{\beta}}_Z) \end{cases} \end{aligned} \tag{5}$$

C. Review of Regularization-based MTL methods

This subsection reviews the formulation of the regularization-based MTL methods for a linear regression model such as μ and summarizes two common norms as the regularization terms in the learning objective. The general formulation of the regularization-based MTL is to solve the following problem, i.e.,

$$\min_{\boldsymbol{\beta}} \quad \ell(X\boldsymbol{\beta}, \mu) + \lambda(\text{norm}(\boldsymbol{\beta})), \tag{6}$$

where $\boldsymbol{\beta} = [\vec{\beta_1}, \cdots, \vec{\beta_Z}] \in \mathbb{R}^{(p+4)\times Z}$ is a group of coefficients in the μ model for all communities, and λ is a regularizer coefficient. Also, ℓ denotes a least square loss function over all the communities as follows:

$$\ell(X\boldsymbol{\beta}, \mu) = \sum_{l=1}^{Z} \sum_{j=1}^{n_l} \frac{1}{Z n_l} (X_l{}^j \vec{\beta_l} - \mu_l{}^j)^2, \tag{7}$$

where n_l is the number of observations for community l and X_l^j is the j^{th} observation of X in community l. The norm($\boldsymbol{\beta}$) in Eq. 6 is introduced to capture the relatedness of μ model among all communities. State-of-the-art research presented different norms to characterize the relatedness from the following two perspectives:

1) Common Set of Features for All Communities: The norm($\boldsymbol{\beta}$) $= \|\boldsymbol{\beta}\|_{1,\infty}$ (which is called $l_{1,\infty}$) is introduced as

$$\|\boldsymbol{\beta}\|_{1,\infty} = \sum_j max(|\beta_{j1}|, \cdots, |\beta_{jZ}|), \tag{8}$$

where β_{jk} is the the j^{th} row and k^{th} column of $\boldsymbol{\beta} \in \mathbb{R}^{(p+4)\times Z}$. The $l_{1,\infty}$ regularizer introduces a group-sparse structure to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

conduct variable selection for the μ model among all communities. It penalizes the sum of the maximum absolute values of each row to encourage each row of $\boldsymbol{\beta}$ to have zero elements. Therefore, the $l_{1,\infty}$ regularizer can find a common set of input variables that have an effect on the electricity consumption data among different communities.

2) Shared Low-Rank Structure for All Communities: fu f

$$\|\boldsymbol{\beta}\|_* = \sum_{i=1}^{rank(\boldsymbol{\beta})} \sigma_i(\boldsymbol{\beta}), \tag{9}$$

where σ_i's are singular values of the matrix $\boldsymbol{\beta}$ obtained by a singular value decomposition.

The two types of norms capture the information on the inter-community relatedness from two different perspectives. By combining the two types of relatedness information, this paper develops a new method to improve the MT-BSGP modeling in the next subsection.

D. The Proposed LRDM for Capturing Relatedness

More recently, Jalali et.al. [12] proposed a Dirty model by decomposing the variables coefficients ($\vec{\beta}_l$ in Eq. 4) into a group sparse component ($\vec{P}_l$) and a sparse component ($\vec{Q}_l$) as:

$$\vec{\beta}_l = \vec{P}_l + \vec{Q}_l, \quad \vec{P}_l, \vec{Q}_l \in \mathbb{R}^{p+4}, \tag{10}$$

by which the penalization using the sparse component $\vec{Q}_l$ deals with community-specific variations while that using the group sparse component $\vec{P}_l$ aims to capture the relatedness among the model estimations for different communities.

Inspired by the "Dirty" model and two types of norms to capture the relatedness information, this paper further develops a low-rank structure for the matrix $\vec{Q}_l$ to incorporate more inter-community relatedness information that could not be captured by the group sparse component, leading to a low-ranked version of dirty model (LRDM). The objective of the LRDM is proposed as follows:

$$\min_{\boldsymbol{P},\boldsymbol{Q}} \quad \ell(X(\boldsymbol{P}+\boldsymbol{Q}),\mu) + \lambda_1\|\boldsymbol{P}\|_{1,\infty} + \lambda_2\|\boldsymbol{Q}\|_*, \tag{11}$$

where each column of $\boldsymbol{P}$ and $\boldsymbol{Q}$ corresponds to a community, i.e., $\boldsymbol{P} = [\vec{P}_1, \cdots, \vec{P}_Z]$ and $\boldsymbol{Q} = [\vec{Q}_1, \cdots, \vec{Q}_Z]$, and $\boldsymbol{\beta} = \boldsymbol{P}+\boldsymbol{Q}$. Moreover, λ_1 and λ_2 are the coefficients of the norms.

The structure of $\boldsymbol{\beta}$ in the proposed LRDM can be illustrated in Fig. 2. The $l_{1,\infty}$ norm of the group sparse component $\boldsymbol{P}$ attempts to find those input variables that have similar effects on the electricity consumptions across communities by encouraging the entire rows of those variables that do not have the similar effects on the electricity consumption to have zero elements. Thus, the group sparse component is to constrain all models to share a common set of features (input variables).

The l_* norm of component $\boldsymbol{Q}$ provides a low-rank structure that has common basis vectors shared across multiple communities. Suppose that $rank(\boldsymbol{\beta}) = \nu$. The component $\boldsymbol{Q}$ can then be represented on a basis vector multiplied with a coefficient matrix as $\boldsymbol{Q} = \boldsymbol{B}\boldsymbol{C}^T$ where $\boldsymbol{B} = \left[\vec{b_1}, \cdots, \vec{b_\nu}\right] \in \mathbb{R}^{(p+4)\times\nu}$ and $\boldsymbol{C} = \left[c_{ij}\right], i = 1, \cdots, Z$ and $j = 1, \cdots, \nu$. The basis vectors $\boldsymbol{B}$ span a low-dimensional subspace of matrix $\boldsymbol{Q}$ and capture the inter-community relatedness. The coefficient matrix $\boldsymbol{C}$ can be different for different communities.

Therefore, the proposed LRDM aims to combine the two types of norms as described above. Specifically, we propose that integrating of $l_{1,\infty}$ and l_* norms can increase the chance of capturing more shared information among multiple communities and can outperform the existing MTL methods which only utilizing either $l_{1,\infty}$ or l_* for capturing the relatedness. A real-world case study will be conducted in Section III, and the results in Section IV-C will validate the effectiveness of combining both norm structures in improving load forecasting accuracy.

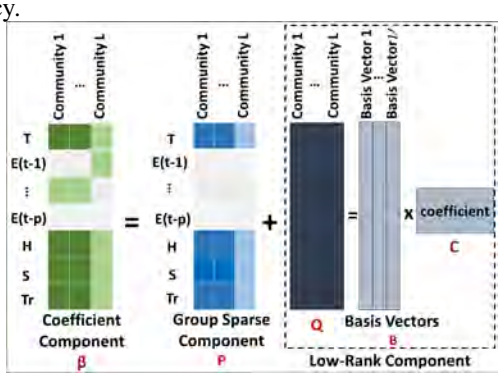


Fig. 2. Structure of the model coefficients $\boldsymbol{\beta}$ in the proposed LRDM

The proposed LRDM in Eq. 11 is an unconstrained convex optimization problem whose function is non-smooth. This non-smoothness exists in the l_* and $l_{1,\infty}$, creating a challenge in solving the LRDM problem in Eq. 11. One method is to use the Accelerated Proximal Method (APM) [14] due to its optimal convergence rate and its capability to deal with large-scale non-smooth optimization problems [33]. For more details on the APM procedures, please refer to [14] and [33].

Accuracy of the LRDM: This section also briefly estimates a theoretical boundary for characterizing the accuracy of the LRDM. Consider the optimization problem of the Eq. 11 for $l \geq 2$ and $n \geq 1$ and the same sizes of training data for all the tasks. Also define the following parameters, i.e.,

$$\frac{\lambda_1}{\sqrt{Z}}; \lambda_2 \geq \lambda; \lambda = \frac{2\sigma_\epsilon}{nZ}\sqrt{(p+4)+t} \tag{12}$$

where $t > 0$ is a universal constant. Then, with a probability of at least $1 - Z\exp(-\frac{1}{2}(t-(p+4)\log(1+\frac{t}{(p+4)})))$ for a global minimizer $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{P}}$ in Eq. 11 and any $\boldsymbol{P}, \boldsymbol{Q} \in R^{(p+4)\times Z}$; a theoretical boundary can be derived as:

$$\sum_{l=1}^{Z}\frac{1}{Zn}\|X_l{}^T(\hat{\vec{P}}_l + \hat{\vec{Q}}_l) - \mu_l\|_F^2 \leq (1+\gamma)\inf_{\vec{P}_l,\vec{Q}_l}\sum_{l=1}^{Z}\frac{1}{Zn}\|X_l{}^T(\vec{P}_l+\vec{Q}_l)-\mu_l\|_F^2 + \gamma(\frac{1}{2}+\frac{1}{\gamma})^2\Big(\frac{\lambda_1^2}{\kappa_1^2(c)} + \frac{\lambda_2^2}{\kappa_2^2(2r)}\Big), \tag{13}$$

where $\sum_{l=1}^{Z}\frac{1}{Zn}\|X_l{}^T(\hat{\vec{P}}_l + \hat{\vec{Q}}_l) - \mu_l\|_F^2$ is the loss function of the estimated model and $\inf_{\vec{P}_l,\vec{Q}_l}\sum_{l=1}^{Z}\frac{1}{Zn}\|X_l{}^T(\vec{P}_l+\vec{Q}_l) - \mu_l\|_F^2$ reflects the performance of the theoretically best model (by infimum of the loss function) and $\gamma(\frac{1}{2}+\frac{1}{\gamma})^2\big(\frac{\lambda_1^2}{\kappa_1^2(c)} + \frac{\lambda_2^2}{\kappa_2^2(2r)}\big)$ is an adjustment term with a constant $\gamma \geq 0$ and $rank(\boldsymbol{Q}) \leq r$ and $|C(\boldsymbol{P})| \leq c$. Also, $C(\boldsymbol{P})$ is defined as the set of indices corresponding to the non-zero rows of the matrix $\boldsymbol{P}$, and $|C|$ denotes the number of elements in C. The parameters $\kappa_1(c)$ and $\kappa_2(2r)$ are defined as:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

ALGORITHM 1: 10-fold Cross Validation for finding the best values of λ_1 and λ_2

1) **Input** $Trr = \{X, Y\}$: Training data; $Tss = \{X_s, Y_s\}$: Testing data; $\lambda_1 = [\lambda_{11}, \cdots, \lambda_{1m}]$; $\lambda_2 = [\lambda_{21}, \cdots, \lambda_{2m}]$.
2) Find an equal partition of $[Trr_1, \cdots, Trr_{10}]$ of all observations in Trr
3) **For** $i = 1 : m$ (number of suggested values for λ_1 and λ_2)
   10-fold cross validation:
   a) **For** $k \in \{1, \cdots, 10\}$
      - Define $S_k = \{S_X, S_Y\}$ where $S_k = Trr \setminus Trr_k$,
      - n = length(Trr_k),
      - $\hat{\vec{\beta}}_l^{(ik)}$ = LRDM.Fit($\lambda_{1i}, \lambda_{2i}; S_X, S_Y$),
      - $pred_l^{(ik)}$ = predict.LRDM ($\hat{\vec{\beta}}_l^{(ik)}, Trr_k$),
      - $Er_k = \sqrt{(\sum_{z=1}^{n}(pred_l^{(ik)}(z) - Trr_k(z))^2)/n}$,

   b) **End For**
   c) Compute $\bar{Er}^{(i)} = \underset{k}{average}(Er_k)$,
4) **End For**
5) Obtain λ_{1i^*} and λ_{2i^*} by $i^* = \underset{i}{\operatorname{argmin}} \{\bar{Er}^{(i)}\}$,

$$\kappa_1(c) = \min_{(\hat{\boldsymbol{P}}-\boldsymbol{P}),(\hat{\boldsymbol{Q}}-\boldsymbol{Q})\in\mathcal{R}(2r,c)} \frac{\|X\big((\hat{\boldsymbol{P}}-\boldsymbol{P})+(\hat{\boldsymbol{Q}}-\boldsymbol{Q})\big)\|_F}{\sqrt{Zn}\|(\hat{\boldsymbol{P}}-\boldsymbol{P})_{C(\boldsymbol{P})}\|_{1,\infty}}, \tag{14}$$

$$\kappa_2(2r) = \min_{(\hat{\boldsymbol{P}}-\boldsymbol{P}),(\hat{\boldsymbol{Q}}-\boldsymbol{Q})\in\mathcal{R}(2r,c)} \frac{\|X\big((\hat{\boldsymbol{P}}-\boldsymbol{P})+(\hat{\boldsymbol{Q}}-\boldsymbol{Q})\big)\|_F}{\sqrt{Zn}\|\hat{\boldsymbol{Q}}-\boldsymbol{Q}\|_*}, \tag{15}$$

where $X \in \mathbb{R}^{Z(p+4)\times Zn}$ is a block-diagonal matrix with its l_{th} block formed by $X_l \in \mathbb{R}^{(p+4)\times n}$ and the set $\mathcal{R}(2r,c)$ is defined as:

$$\mathcal{R}(2r,c) = \left\{ \begin{matrix} \forall(\hat{\boldsymbol{P}}-\boldsymbol{P})\&(\hat{\boldsymbol{Q}}-\boldsymbol{Q}) \in \mathbb{R}^{(p+4)\times Z}, \ |\ (\hat{\boldsymbol{P}}-\boldsymbol{P}) \neq 0, (\hat{\boldsymbol{Q}}-\boldsymbol{Q}) \neq 0, \\ rank((\hat{\boldsymbol{Q}}-\boldsymbol{Q})) \leq 2r, \ \ |C((\hat{\boldsymbol{P}}-\boldsymbol{P}))| \leq c. \end{matrix} \right\} \tag{16}$$

The proof the theoretical boundary is similar to a derivation procedure in [14]. Due to the page limitation, the details are omitted. Please refer to [12] and [14]. The numerical estimation of the accuracy boundary is presented in Fig. 8, Section IV.C

E. Iterative Algorithm to Estimate Parameters of the MT-BSGP

For establishing the MT-BSGP model, two types of parameters should be estimated for each community including $\vec{\beta}_l$ in the μ_l and $\sigma^2_{GP_l}$ and ϕ_l at each time stamp in GP_l. The challenge for the learning is that $\vec{\beta}_l$ should be jointly estimated by the data from all communities under the MTL framework, whereas the GP_l is community-specific and should be estimated from the data from community l. Any change in the μ_l estimation directly impacts the GP_l and vice versa. This paper develops an iterative algorithm to jointly estimate all the parameters in Eq. 5 for all communities.

The flowchart of MT-BSGP is summarized in Fig. 3, where the superscript shows the number of iterations. According to Fig. 3, in the initialization step, the parameters of GP and $\vec{\beta}_l$ at the first iteration are assigned with zero values. Moreover, a big number is specified for the $conv$ at first iteration, and a threshold for the $conv$ is determined depending on the desired accuracy. Also, some initial values are specified for the λ_1 (coefficient of the group sparse component) and λ_2 (coefficient of the low-rank component). In the training stage, after partitioning the training data (Trr) into ten equal subsets ($Trr_k, k = 1, \cdots, 10$) and for any given λ_1 and λ_2 pair values, a 10-fold cross validation is performed. The 10-fold cross validation is explained in Algorithm 1. According to that, the LRDM is trained on nine subsets and then is tested on the tenth subset. The error Er_k is estimated by the root mean square error (RMSE) for each value of λ_1 and λ_2 and for each of those ten parts. The $\bar{Er}^{(i)}$ is the average errors of the Er_k over those ten parts (iterations) which is the average error of the LRDM with the chosen λ_1 and λ_2. This process is implemented for all the m suggested values of λ_1 and λ_2. Finally, the minimum $\bar{Er}^{(i)}$ is selected, and it determines the best values for the λ_1 and λ_2.

The iterative procedure for learning the model parameters begins after determining the λ_1 and λ_2 in the training part. In iteration j, the μ model is updated by subtracting the Gaussian Process obtained in iteration $j-1$ from the electricity consumption data, and the coefficients $\vec{\beta}_l$'s are estimated by using the proposed LRDM. A convergence test is then run to check the convergence of the μ. The convergence is judged based on whether the changes of the coefficient vectors are within a predefined threshold (ϵ). Otherwise, the algorithm updates GP by subtracting the estimated μ at iteration j from the electricity consumption data (Y_l) before proceeding to the next iteration. This process is run until the convergence test is passed. Finally, the error is obtained by comparing the forecasted electricity consumption with true values.

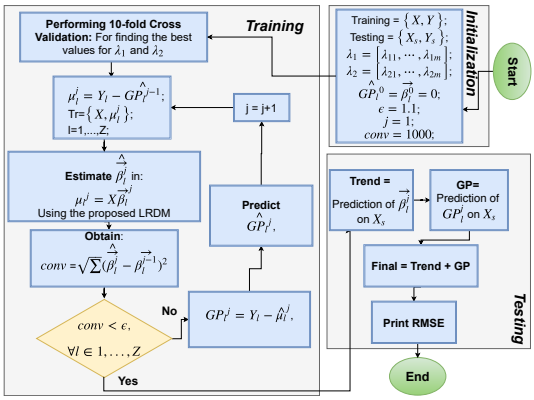


Fig. 3. Flowchart of the proposed MT-BSGP

III. Case Study

The proposed MT-BSGP model is validated by the real-world data from two residential communities in the City of Tallahassee, Florida. The electricity consumption data (kWh) were measured remotely every half an hour and stored by the Meter Data Management System (MDMS) of the City of Tallahassee Utilities. Traffic data were obtained from the Florida Department of Transportation (FDOT) and the City of Tallahassee. The FDOT has six different Telemetered Traffic Monitoring Sites (TTMS) in the city storing traffic counts every half an hour continuously, and the City of Tallahassee collects traffic data on local roadways only for specific locations and days. Also, the environmental data including

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

temperature, humidity, and solar radiation [35] are available through public records for the selected communities.

For the sake of data anonymity and privacy protection, the two selected residential communities in Tallahassee is referred to as the Northeast community and Southeast community in this paper. The measurements in the Northeast and Southeast community include the electricity consumption and environmental data that were collected every half an hour during October, November, and December of 2015 for 50 houses. Also, traffic counts sampled every hour was collected from the same area and used as a predictor variable. The Northeast community has a relatively less amount of data that were collected only at locations with traffic detectors on certain days of 2015. The Northeast only has the data for 18 weekdays in October, November, and December of 2015 for ten houses. Furthermore, all the available data for the Southeast community (50 houses$\times$92 days$\times$48 half-hour increments = 220800 observations) and 17 weekdays for the Northeast community (10 houses$\times$17 days$\times$48 half-hour increments = 8160 observations) are chosen as the training data. The last remaining weekday (the 18th day which is Monday 14th Dec. 2015) for the Northeast community (10 houses$\times$1 day $\times$48 half-hour increments = 480 observations) is chosen as the testing data, which is outside the training data.

Fig. 4 shows the two communities chosen for the case study. It should be pointed out that the type of loads in both selected communities are townhouses with air-conditioning which is the majority of the residential loads category in the State of Florida.

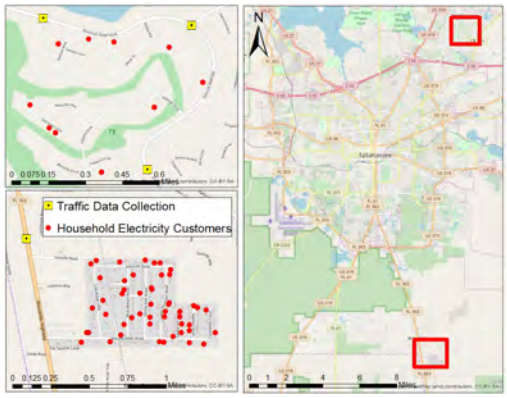


Fig. 4. Two selected communities in the city of Tallahassee, FL. Note that the Northeast community has less measurement locations (red dots) compared with the Southeast community

IV. Results and Discussions

This section discusses the short-term load forecasting results obtained from the proposed MT-BSGP. First, to validate the spatial correlation of the electricity consumption in the Northeast community, we have calculated two statistics of Moran's I [36] and Geary's C [37]. The Moran's I p-value for the Northeast community is 0.04007 and the Geary's C p-value is 0.04607. These results ensure that a spatial dependency exists at a 5% significance level in the electrical consumption data in the case study. Furthermore, after clustering the electrical consumption at the household level based on customer's consumption behaviors, the results show that there still exists spatial dependency within some clusters, thus demonstrating the effectiveness of adopting a Gaussian process model.

The model inputs for this case study are historical electricity consumptions, temperature, humidity, solar radiation, and traffic counts. A time lag two ($p = 2$) was determined for the time series modeling by examining ACF and PACF plots at a 5% significance level. In our case study, after performing the 10-fold cross validation, $\lambda_1 = 100$ and $\lambda_2 = 300$ are selected. Moreover, the iterative algorithm is converged after 6 iterations. In this paper, we use the root mean square error (RMSE), and symmetric mean absolute percentage error (SMAPE) for error indexes as follows: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y(i) - \hat{y}(i)]^2}$, $SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}(i)-y(i)|}{|y(i)|+|\hat{y}(i)|}$, where $y(i)$ is the observation, $\hat{y}(i)$ is the forecasted value, and n is the total number of observations. It should be highlighted that the RMSE and SMAPE reported in this paper are obtained by aggregating the forecasting errors for all the houses in the Northeast community.

A. Multi-task Learning vs. Single-Task Learning

This subsection first compares the MT-BSGP with the BSGP that only relies on the data from one community, i.e., single-task learning. Moreover, this subsection further compares the proposed MT-BSGP with state-of-the-art electricity forecasting methods including the autoregressive integrated moving average with explanatory variable (ARIMAX) [2]-[3], support vector machine (SVM) [6], random forest (RF) [38], and neural network (NN) Regression [7]-[8]. All these methods were implemented only in the Northeast community using the same input data including electricity, weather, and traffic counts. By contrast, the MT-BSGP explores the similar data patterns from the Southeast community to supplement more information to forecast electricity in the Northeast community. RMSE and SMAPE obtained from these methods are shown in Table II. Results show that the MT-BSGP reduced the error by 34% compared with ARIMAX, 22% compared with BSGP in single task structure, 20% compared with SVM, 17% compared with RF, and 15% compared with the neural network. It should be noted that these improvements are not only attributed to a combination of multiple data sources. Next section will discuss this issue.

In this comparison, all the methods were tuned, and the best parameters were selected for them. For example, for ARIMAX, the best model was selected by searching for possible models in the *auto* function in the *forecast* package in *R* programming language. For the SVM, the *radial basis function* kernel was used, and the best parameters were obtained after parameter tuning with 10-fold cross-validation. RF was implemented in the *randomForest* package in the R programming language after tuning the parameters. For the NN, a multilayer perceptron approach (backpropagation) was used, and the best initial values were determined after parameter tuning for the hidden layers, the number of neurons in each hidden layer, and the threshold for the partial derivatives of the error function as stopping criteria. The best NN with three neurons in a hidden layer, with a threshold equal to 0.01, and 1×10^7 as the maximum number of steps is chosen.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

Table II
COMPARING LOAD FORECASTING ERROR (RMSE AND SMAPE) FOR DIFFERENT SINGLE-TASK LEARNING METHODS

Method	Type	RMSE	SMAPE
ARIMAX	STL	0.6050	1.8553
SVM	STL	0.4998	1.1613
RF	STL	0.4815	1.3163
BSGP	STL	0.5106	1.1438
NN	STL	0.4697	1.2844
MT-BSGP	MTL	0.3996	1.0772

For visualization, Fig. 5 compares the forecasted load profiles for a house on a weekday by using MT-BSGP. The key finding of this figure is that the MT-BSGP can effectively capture trends of the load profiles of the households even for some sudden jump in the load profile.

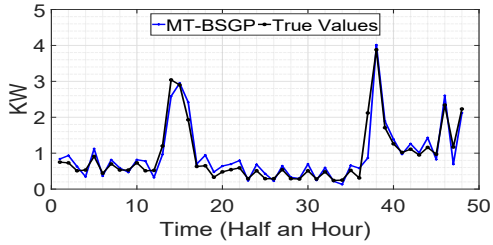


Fig. 5. Forecasted load profile of a weekday (14 Dec. 2015) for a random house in the Northeast community

B. Multi-task Learning vs. Learning Based on Merged Data

The MTL is not as simple as combining the data from multiple communities. Simply merging the multi-community data ignores the community-unique information and between-community difference, thus introducing the information that may negatively affect the model learning process. The MTL, however, explores the similar pattern in the data and distinguishes the shared information across communities from community-specific information, thereby improving the learning accuracy. When two datasets share a certain amount of similarities but also present significant dissimilarities, merging those datasets in a single-task learning framework usually can perform worse than treating them separately in a multi-task learning framework [39]. This result was also reported by [40] and [41] based on human activity and vehicle sensor datasets.

Fig. 6 compares the proposed MT-BSGP with the load forecasting methods based on the merged data from two communities. It can be seen that the MT-BSGP significantly outperforms the competitive methods including SVM, RF, and NN applied to the merged data. It is also noticed that the learning by using the merged data did not significantly reduce RMSE by comparing Table II and Fig. 6. The results demonstrate the values of the MTL in exploring the inter-community relatedness and knowledge transfer among different communities to improve the forecasting accuracy.

C. MT-BSGP vs. State-of-the-Art MTL Methods

The proposed MT-BSGP learned by the LRDM and the iterative algorithm was compared with three regularization-based MTL methods including dirty model (Dirty) [12], sparse-low rank (SLR) [13], and robust MTL (Robust) [14], and a kernel-based MTL method called low-rank output kernel learning (LR-OKL) [11]. The difference between the proposed LRDM

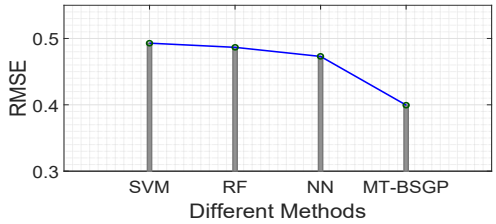


Fig. 6. Comparing the MT-BSGP with state-of-the-art forecasting methods using the merged data from two communities (number of samples: 480)

and the aforementioned regularization-based MTL methods is shown in Table III. Also, the results obtained by using the aforementioned MTL methods for the case with and without GP are presented in Table IV. It is observed that adding the GP to the regularization-based MTL methods can significantly improve the forecasting accuracy i.e., MT-BSGP reduced the RMSE by 22% compared with SLR+GP. Comparing the Dirty model vs. the proposed LRDM shows that LRDM outperforms the Dirty model by 11%. This improvement implies that adding the shared low-rank structure (l_*) to the Dirty model can significantly improve the accuracy, thus demonstrating the contribution from the low-rank structure in improving the forecasting accuracy. On the other hand, the comparisons between the proposed LRDM and SLR/Robust methods, which lead to 24% and 10% reductions in RMSE error, respectively, indicate the significant contribution from the sparse structure.

Table III
DIFFERENCE BETWEEN THE PROPOSED LRDM AND COMPARED METHODS

Methods	Perspective	Norms
Dirty Model	Common Set of Features	$l_{1,\infty}$
SLR and Robust	Shared Low-Rank	l_*
Proposed LRDM	Common Set of Features and Shared Low-Rank	$l_{1,\infty}$ and l_*

Table IV
COMPARING LOAD FORECASTING ERROR (RMSE AND SMAPE) FOR DIFFERENT MULTI-TASK LEARNING METHODS

Method	Type	RMSE	SMAPE
SLR	MTL	2.077	2.9513
Dirty	MTL	1.7635	2.7432
Robust	MTL	1.7523	2.6819
LRDM	MTL	1.5783	2.2364
SLR+GP	MTL	0.5127	1.3384
Dirty+GP	MTL	0.4728	1.1747
Robust+GP	MTL	0.4635	1.1414
MT-BSGP (LRDM+GP)	MTL	0.3996	1.0772

Fig. 7 compares the forecasted load profiles of a randomly selected house on a weekday by using MT-BSGP, BSGP, with LR-OKL. Although LR-OKL demonstrates its significant reduction of error compared with other methods as shown in Table IV, the proposed MT-BSGP further reduces the RMSE by 14% and better captures the temporal data variations. Given the same data sources for MTL, the proposed MT-BSGP also takes the advantages of the model structure, which is decomposed into a μ that captures the shared temporal correlations among communities and a GP that models the community-specific local variations. The model based on the decomposition ($\mu - GP$) outperforms the integrated kernel method adopted by LR-OKL to capture the spatiotemporal correlations among different tasks since it is more challenging to select a proper format for the kernel function in the LR-OKL method. The comparison also shows that the highest error

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics

occurred in the peak hour (33rd half an hour). This result is not surprising since the peak load is a relatively rare event that happens once a day with a short duration time. Therefore, load forecasting usually has more errors for the peak load compared to non-peak loads. Fig. 7 also shows that MT-BSGP overall performance is better than other methods throughout the day. Even for the peak load, the MT-BSGP forecasting (blue line or V style) is the closest to the true peak load curve (black line or o style) compared with the other methods. Fig. 8 further compared the standard deviation of the proposed MT-BSGP vs. other MTL methods. It can be seen that MT-BSGP has the smallest mean RMSE and standard deviation.

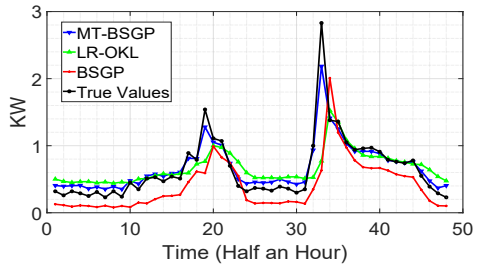


Fig. 7. Foretasted load profile of a weekday (14 Dec. 2015) for a random house in the Northeast community

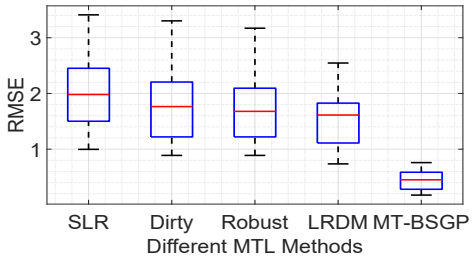


Fig. 8. RMSE obtained from different MTL methods after 15 repetitions

D. MT-BSGP vs. Clustering-based Method

In this subsection, in order to exhibit the superiority of the proposed MT-BSGP over clustering-based methods, we compare the BSGP and MT-BSGP with one of the most recent paper [26] in the clustering-based load forecasting models. According to [26], the method is run for different k (e.g., varying from $k = 1$ to 10) and the best k is chosen based on the least error obtained. After applying K-means clustering for a given k, LS-SVM will be utilized to each cluster. The forecasting result for each cluster is then obtained and combined to obtain the final estimation result. The k value that gives the highest accuracy should be selected to determine the number of classes.

The above procedure is run for both Southeast and Northeast communities. The houses in the training data (50 houses for the Southeast community and 10 houses for the Northeast community) are clustered by the k-means algorithm based on their load consumption patterns for different specified "k" ranging from 1 to 10. As such, for each "k" specified, all houses are clustered to k clusters. Then, for each cluster, LS-SVM is trained and is tested for the next-day forecasting and the RMSE is recorded. Finally, the overall error is aggregated for each "k" and the "k" that leads to a minimum error (obtained by LS-SVM) is selected. Based on this procedure similar to [26], $k = 5$ is selected for the Southeast community and $k = 4$ is selected for the Northeast community. Then, we have checked the spatial dependencies of all the clusters obtained from the Southeast and Northeast communities by Moran's I and Geary's C. The results show that some clusters including clusters 4 and 5 in the Southeast and clusters 1 and 3 in the Northeast community did exhibit spatial dependency within the clusters. Therefore, for those clusters that have the spatial dependency in the Southeast and Northeast communities, we compare the k-means based LS-SVM with the BSGP and our proposed MT-BSGP in Table V.

Table V
ERRORS OBTAINED FROM THE BSGP AND MT-BSGP WITH THE K-MEANS BASED LS-SVM

Community	Clusters	K-means Based LS-SVM		BSGP		MT-BSGP	
		RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Southeast	4	0.5137	0.2618	0.5076	0.2033	N/A	N/A
	5	0.3359	0.2525	0.3312	0.2021	N/A	N/A
Northeast	1	0.2251	0.5639	0.2119	0.5122	0.1893	0.4513
	3	0.4764	0.8375	0.4614	0.7790	0.3335	0.6464

Note: N/A means not applicable since the Southeast community is the source data source and the target of interest for forecasting is the Northeast community with limited data.

The results shown in Table V indicate that BSGP can improve the accuracy by 22% and 20% for clusters 4 and 5 in the Southeast community based on the SMAPE. Moreover, according to the SMAPE, BSGP outperforms the K-means based LS-SVM by 9% and 7% for clusters 1 and 3 in the Northeast community. The results highlight that in some clusters where spatial dependencies still exist, considering the GP and spatial dependencies can improve the prediction compared with the K-means based LS-SVM.

V. CONCLUSION

This paper proposes a multi-task Bayesian spatiotemporal Gaussian Process (MT-BSGP) to capture the relatedness across different residential communities in a city and the local spatial variations between households to improve the short-term load forecasting. To better capture the relatedness, a novel multi-task learning approach, which is called low ranked dirty model (LRDM) is proposed to improve the learning of the MT-BSGP by employing the "common set of features" ($l_{1,\infty}$ norm) and "shared low-rank" (l_* norm) structures. Moreover, to overcome the challenge in jointly estimating the parameters of the LRDM and the Gaussian Process (local spatial variations) an iterative algorithm is proposed. Based on the real-world data from the City of Tallahassee, Florida, USA, the key findings of this paper are as follows:

- The proposed MT-BSGP has a better performance than traditional single-task learning methods including ARIMAX, BSGP, SVM, RF, and NN by 34%, 22%, 20%, 17%, and 15% respectively.
- The proposed MT-BSGP significantly outperforms state-of-the-art load forecasting methods including SVM, RF, and NN based on the merged multi-community data by almost 20%, 18%, and 17% respectively.
- The proposed MT-BSGP outperforms the state-of-the-art MTL methods (SLR, Dirty, and Robust) and a more recently developed LR-OKL since it employs the "common set of features" and "shared low-rank structure" simultaneously in the μ which can be commonly shared across

communities and by employing a μ-GP decomposition structure.

- The proposed MT-BSGP outperforms the clustering-based methods such as K-means based LS-SVM since the spatial dependencies may still exist in some clusters, which only can be captured by GP in the MT-BSGP.

Future work will include the generalization of the proposed MT-BSGP method to incorporate more environmental variables under different scenarios of multi-community relatedness. In addition, the model-based strategy will be developed by using the MT-BSGP to guide utility company and power plants to regulate the electricity generation to meet market demands by producing as-needed energy. In addition, dynamic pricing schemes in electricity markets based on the proposed load forecasting will be considered to improve energy utilization and cost-effectiveness.

REFERENCES

[1] G. Cerne, D. Dovzan, I. Skrjanc, "Short-term load forecasting by separating daily profiles and using a single fuzzy model across the entire domain," *IEEE Transaction on Industrial Electronics*, Vol.65, No.9, 2018.

[2] K. S. L. Madhavi, J. Cordova, M. B. Ulak, M. Ohlsen, E. E. Ozguven, R. Arghandeh, and A. Kocatepe, "Advanced electricity load forecasting combining electricity and transportation network," *in 2017 NorthAmerican Power Symposium (NAPS)*, Sep. 2017, pp. 1–6.

[3] W. Jian-jun, N. Dong-Xiao, and L. Li, "An arma cooperate with artificial neural network approach in short-term load forecasting," *in 2009 Fifth International Conference on Natural Computation*, vol. 1, Aug 2009.

[4] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-term solar power forecasting based on weighted Gaussian Process regression," *IEEE Transaction on Industrial Electronics*, vol. 65, no. 1, pp. 300-308, 2018.

[5] S.H. Ling, F.H.F. Leung, H.K. Lam, P.K.S. Tam, "Short-term electric load forecasting based on a neural fuzzy network," *IEEE Transaction on Industrial Electronics*, vol.50, no.6, pp. 1305-1316, 2003.

[6] B.-J. Chen, M.-W. Chang, and C.-J. lin, "Load forecasting using support vector machines: a study on eunite competition 2001," *IEEE Transactions on Power Systems*, vol. 19, no. 4, Nov 2004.

[7] C. Cecati, J. Kolbusz, P. Rozycky, P. Siano, and B.M. Wilamowski, "A novel RBF training algorithm for short-term electric load forecasting and comparative studies," *IEEE Transaction on Industrial Electronics*, vol. 62, no. 10, pp. 6519-6529, Apr. 2015.

[8] S.H. Ling, FHF. Leung, H.K. Lam, Y.S. Lee, P.K.S. Tam, "A novel genetic-algorithm-based neural network for short-term load forecasting," *IEEE Transaction on Industrial Electronics*, vol.50, no.4, 2003.

[9] K. S. L. Madhavi, M. Gilanifar, Y. Zhou, E. E. Ozguven and R. Arghandeh, "Multivariate Deep Causal Network for Time Series Forecasting in Interdependent Networks," 2018 IEEE Conference on Decision and Control (CDC), Miami Beach, FL, 2018, pp. 6476-6481.

[10] Y. Zhang, G. Luo, F. Pu, "Power Load Forecasting based on Multi-task Gaussian Process," *Proceedings of the 19th world congress the international federation of automatic control*, South Africa, August 2014.

[11] J.-B. Fiot, F. Dinuzzo, "Electricity Demand Forecasting by Multi-Task Learning," *IEEE Transactions on Smart Grid*, Vol. PP, No. 99, 2016.

[12] A. Jalali, P. Ravikumar, S. Sanghavi, "A Dirty Model for Multiple Sparse Regression," *Neural Information Processing Systems (NIPS)*, 2010.

[13] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1179-1188, 2010.

[14] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," *in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 42–50, 2011.

[15] E.V. Bonilla, F.V. Agakov, C.K.I. Williams, "Kernel Multi-task Learning using Task-specific Features," *Artificial Intelligence and Statistics*, 2007.

[16] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behavior learning," *IEEE Transactions on Power Systems*, vol. PP, no. 99, 2017.

[17] Y.-M. Wi, S.K. Joo, K.-B. Song, "Holiday Load Forecasting Using Fuzzy Polynomial Regression With Weather Feature Selection and Adjustment," *IEEE Transactions on Power Systems*, Vol.27, pp.596-603, 2012.

[18] D. Pinzan, A. Kocatepe, M. Gilanifar, M. B. Ulak, E. E. Ozguven, R. Arghandeh, "Data-Driven and Hurricane-Focused Metrics for Combined Transportation and Power Networks Resilience" (No. 18-05714), 2018.

[19] D. Fay, J.V. Ringwood, "On the Influence of Weather Forecast Errors in Short-Term Load Forecasting Models," *IEEE Transactions on Power Systems*, Vol. 25, pp. 1751-1758, 2010.

[20] L. M. Konila Sriram, M. B. Ulak, E. E. Ozguven and R. Arghandeh, "Multi-Network Vulnerability Causal Model for Infrastructure Co-Resilience," in IEEE Access, vol. 7, pp. 35344-35358, 2019. doi: 10.1109/ACCESS.2019.2904457

[21] J. Aparicio, J. Rosca, M. Mediger, A. Essl, K. Arzig, C. Develder, "Exploiting road traffic data for Very short-term load forecasting in Smart Grids," *Innovative Smart Grid Technologies Conference (ISGT), IEEE PES*, Washington, DC, 2014.

[22] L. M. Konila Sriram, M. Gilanifar, Y. Zhou, E. E. Ozguven, R. Arghandeh, "Causal Markov Elman Network for Load Forecasting in Multinetwork Systems," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1434-1442, 2019.

[23] J. Cordova, L. M. K. Sriram, A. Kocatepe, Y. Zhou, E. E. Ozguven and R. Arghandeh, "Combined Electricity and Traffic Short-Term Load Forecasting Using Bundled Causality Engine," *IEEE Transactions on Intelligent Transportation Systems*. doi: 10.1109/TITS.2018.2876871

[24] A. Jindal, G. S. Aujla, N. Kumar, R. Prodan and M. S. Obaidat, "DRUMS: Demand Response Management in a Smart City Using Deep Learning and SVR," 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-6. doi: 10.1109/GLOCOM.2018.8647926

[25] O. Valgaev, F. Kupzog and H. Schmeck, "Building power demand forecasting using K-nearest neighbours model – practical application in Smart City Demo Aspern project," in CIRED - Open Access Proceedings Journal, vol. 2017, no. 1, pp. 1601-1604, 10 2017. doi: 10.1049/oap-cired.2017.0419

[26] X. Wang, W. Lee, H. Huang, R. L. Szabados, D. Y. Wang and P. Van Olinda, "Factors that Impact the Accuracy of Clustering-Based Load Forecasting," *in IEEE Transactions on Industry Applications,* vol. 52, no. 5, pp. 3625-3630, Sept.-Oct. 2016.

[27] J. Shi, Y. Liu and N. Yu, "Spatio-temporal modeling of electric loads," *2017 North American Power Symposium (NAPS), Morgantown, WV*, 2017, pp. 1-6.

[28] C. Gu and P. Jirutitijaroen, "Dynamic State Estimation Under Communication Failure Using Kriging Based Bus Load Forecasting," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 2831-2840, Nov. 2015.

[29] H. Xia, Y. Ding, J. Wang, "Gaussian process method for form error assessment using coordinate measurements," *IIE Transactions*, 40:10, pp. 931-946, 2008.

[30] K. S. Bakar, "Bayesian Analysis of Daily Maximum Ozone Levels," *Ph.D. Thesis*, University of Southampton, pagination, 2012.

[31] G. Xie, X. Chen and Y. Weng, "An Integrated Gaussian Process Modeling Framework for Residential Load Prediction," *in IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7238-7248, Nov. 2018.

[32] M. Fazel, H. Hindi, and S.P. Boyd. "A rank minimization heuristic with application to minimum order system approximation," *In Proceedings American Control Conference*, Vol. 6, 2001.

[33] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal of Imaging Science*, vol. 2, pp.183-202, 2009.

[34] B. Recht, M. Fazel, P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, 52(3), 471-501, 2010.

[35] WeatherSTEM, https://leon.weatherstem.com.

[36] H. Li, C. A. Calder, N. Cressie, "Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model," *Geographical Analysis*, 39.4, pp. 357-375, 2007.

[37] M.-J. Fortin, M.R.T. Dale, J.M. Ver Hoef, "Spatial Analysis in Ecology," *Wiley StatsRef: Statistics Reference Online*, 2014.

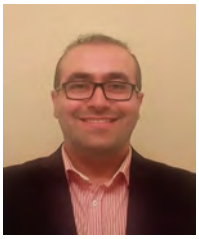
[38] J. Moon, K.-H. Kim, Y. Kim, and E. Hwang, "A short-term electric load forecasting scheme using 2-stage predictive analytics," *IEEE International Conference on Big Data and Smart Computing (BigComp)*, China, 2018.

[39] V. Smith, C.K. Chiang, M. Sanjabi, A.S. Talwalkar, "Federated multi-task learning," *In Advances in Neural Information Processing Systems*, pp. 4424, 2017.

[40] J. Konecny, H.B. McMahan, D. Ramage, "Federated Optimization: Distributed Optimization Beyond the Datacenter," arXiv preprint arXiv:1511.03575, 2015.

[41] McMahan, H. Brendan, et al. "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2928275, IEEE Transactions on Industrial Electronics



Mostafa Gilanifar Mostafa Gilanifar, Ph.D. is a postdoctoral research associate in the ECE department at the University of Utah. He received his Ph.D. degree in Industrial & Manufacturing Engineering from Florida State University in May 2019. He also holds M.Sc. and B.Sc. degrees in Industrial Engineering. His research interests include data analytics for complex systems, machine learning algorithms, and data fusion & multi-task learning models for heterogeneous spatiotemporal data obtained from smart cities and smart grids.

Hui Wang Hui Wang received his Ph.D. degree in industrial engineering from the University of South Florida in 2007. He is currently an assistant professor of Industrial and Manufacturing Engineering at Florida State University. His recent research focuses on cloud data fusion to support artificial intelligence (AI) in Internet-of-Things (IoT) applications including inter-connected flexible cybermanufacturing systems, supply chain network for on-site/on-demand production, and smart power systems. His research has been mostly sponsored by the US National Science Foundation.

Lalitha Madhavi Konila Sriram Lalitha Madhavi K.S is a Ph.D. candidate at Florida State University, carrying out research for the Center for Advanced Power Systems (CAPS). She received her BS degree in Electrical Engineering in 2013. Her research interests include internet of things, machine learning, and data analysis for decision support in smart cities and smart grids.

Eren E. Ozguven Eren E. Ozguven, Ph.D. is an Assistant Professor at the Dept. of Civil & Environmental Eng. at Florida A&M University-Florida State University. Dr. Ozguven holds a Ph.D. degree in Civil and Environmental Engineering from the Rutgers University (New Brunswick, NJ, USA) with concentration in emergency supply transportation operations. His research interests include smart cities, urban mobility, traffic safety and reliability, emergency transportation, and intelligent transportation systems.

Reza Arghandeh Dr. Reza Arghandeh is a Full Professor in the Dept of Computing, Mathematics, and Physics at the Western Norway University of Applied Sciences, Norway. He is also a Senior Data Scientist at StormGeo Company in Bergen, Norway. He has been an Assistant Professor in the Electrical and Computer Engineering Dept at Florida State University, USA during 2015-2018. He completed his Ph.D. in Electrical Engineering at Virginia Tech (2013) and spent two years as a post-doctorate fellow at the University of California-Berkeley. His research interests include data analysis and decision support for smart grids and smart cities.