Spatial Auto-regressive Dependency Interpretable Learning Based on Spatial Topological Constraints

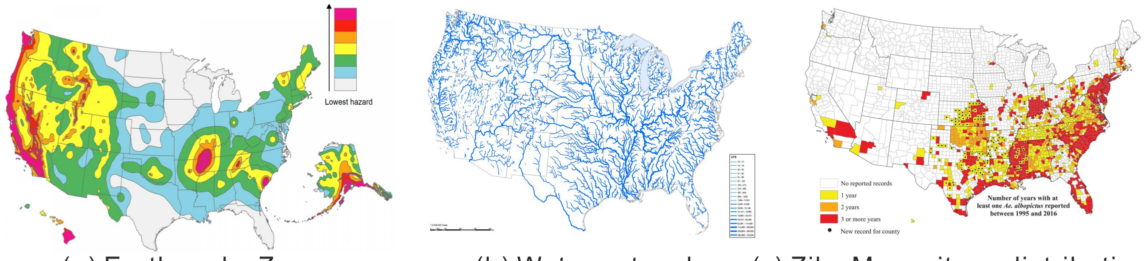
LIANG ZHAO, George Mason University, USA
OLGA GKOUNTOUNA, George Mason University, USA
DIETER PFOSER, George Mason University, USA

Spatial regression models are widely used in numerous areas, including detecting and predicting traffic volume, air pollution, and housing prices. Unlike conventional regression models, which commonly assume independent and identically distributions among observations, existing spatial regression requires the prior knowledge on spatial dependency among the observations in different spatial locations. Such spatial dependency is typically predefined by domain experts or heuristics. However, without sufficient consideration on the context of the specific prediction task, it is prohibitively difficult for human to pre-define the numerical values of the spatial dependency without bias. More importantly, in many situations, the techniques are insufficient to sense the complete connectivity and topological patterns among spatial locations (e.g., in underground water network and human brain network). Until now, these issues are still extremely difficult to address and little attention has been paid to the automatic optimization of spatial dependency in relation to a prediction task, due to three challenges: 1) necessity and complexity of modeling the spatial topological constraints; 2) incomplete prior spatial knowledge; and 3) difficulty in optimizing under spatial topological constraints which are usually discrete or nonconvex. To address these challenges, this paper proposes a novel convex framework that jointly learns the prediction mapping and spatial dependency automatically based on spatial topological constraints. There are two different scenarios to be addressed. First, when the prior knowledge on existence of conditional independence among spatial locations is known (e.g., via spatial contiguity), we propose the first model named Spatial-Autoregressive Dependency Learning I (SADL-I) to further quantify such spatial dependency. However, when the knowledge on the conditional independence is unknown or incomplete, our second model named Spatial-Autoregressive Dependency Learning II (SADL-II) is proposed to automatically learn the conditional independence pattern as well as quantify the numerical values of the spatial dependency, based on spatial topological constraints. Topological constraints are usually discrete and nonconvex which is extremely difficult to be optimized together with continuous optimization problem of spatial regression. To address this, we propose convex and continuous equivalence of the original discrete topological constraints with theoretical guarantee. The proposed models are then transferred to convex problems which can be iteratively optimized by our new efficient algorithms until convergence to a global optimal solution. Extensive experimentation using several real-world datasets demonstrates the outstanding performance of the proposed models.

Additional Key Words and Phrases: Spatial auto-regressive, spatial topological constraints, Alternating Direction Method of Multipliers, graphical LASSO

Authors' addresses: Liang Zhao, George Mason University, 4400 University Drive, Fairfax, VA, 22030, USA, lzhao9@gmu.edu; Olga Gkountouna, George Mason University, 4400 University Drive, Fairfax, VA, 22030, USA, beranger@inria.fr; Dieter Pfoser, George Mason University, 4400 University Drive, Fairfax, VA, 22030, USA.

Manuscript submitted to ACM


Lowest hazard
No reported records
1 year
2 years
3 or more years
New record for county
Number of years with at least one *Ae. albopictus* reported between 1995 and 2016

(a) Earthquake Zones (b) Water network (c) Zika Mosquitoes distribution

Fig. 1. Spatial dependency varies across different application backgrounds.

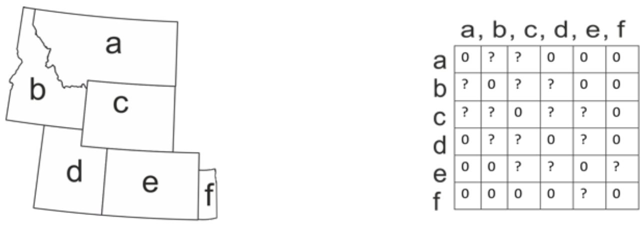
ACM Reference Format:

Liang Zhao, Olga Gkountouna, and Dieter Pfoser. 2019. Spatial Auto-regressive Dependency Interpretable Learning Based on Spatial Topological Constraints. 1, 1 (May 2019), 27 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Spatial regression is an important research area that has applications in domains such as predicting traffic volume, home prices, and the pollution index [38]. A core characteristic of spatial regression relates to Tobler's first law of geography, which states that "Everything is related to everything else, but things that are nearby are more related than distant things" [43]. This aspect is contradicting to the "independent and identically distributed (i.i.d.)" assumption, which is typically applied to conventional regression problems. In spatial data mining and spatio-temporal statistics, various spatial regression models have been proposed to address this type of non-i.i.d. problems. The most-commonly used models tend to be the spatio-autoregressive model and its variants, which enforce the smoothness of data values within geographical neighborhoods. Over the years, numerous spatial autoregressive models have been developed and widely utilized, such as spatial Durbin, geographically weighted models, spatial X, and spatial panel models [38]. All these prediction models require prior knowledge of the spatial dependency, which is usually pre-defined by domain experts or estimated by heuristic distributions [11].

Contiguity matrix [3] is widely used to define the dependency among spatial locations which are considered as nodes and their connections are determined by the existence of contiguity. Contiguity matrix assumes that contiguous locations that share boundaries have the same strength of dependency. Unfortunately, this is usually not the case, as the connectivity strength among locations could be different due to the various lengths of the shared boundary, the distance between different locations, and even the sizes and shapes of the spatial regions. To account for such aspects, some spatial statistics methods consider the boundary length, the distance between locations, or the spatial autocorrelation statistics [11] as heuristics for estimating the actual spatial dependency. However, the actual strength of the spatial dependency for a set of locations can neither be completely determined solely by the boundary length, nor by the distance or autocorrelation statistics, but tends to be a comprehensive combination of all the (explicit and implicit) relevant factors. Two locations can be correlated in various ways, for example by sharing the same earthquake zone or being passed through by the same highway. Two nearby and contiguous locations can also be less correlated if they are separated by mountains or bodies of water. More importantly, the spatial dependency normally comes with context, so the strength of their spatial dependency could vary across different prediction tasks. As shown in Figure 1, the dependency between the states of Louisiana and Arkansas is weak in terms of earthquakes but very strong in terms of their shared water network. Similarly, the correlation between Florida and South Carolina is high in terms of their



(a) spatial locations

	a	b	c	d	e	f
a	0	?	?	0	0	0
b	?	0	?	?	0	0
c	?	?	0	?	?	0
d	0	?	?	0	?	0
e	0	0	?	?	0	?
f	0	0	0	0	?	0

(b) topological constraint

Fig. 2. Example of prior knowledge on the spatial conditional independency via contiguity correlation

risk of Zika virus outbreaks, but low in terms of earthquake risk. Therefore, it could be prohibitive for the conventional expertise and heuristics to comprehensively consider and precisely quantify the context-based dependencies.

In order to address the above problem, it is preferable to learn the numerical strength of spatial dependency from a given application context. However, up until now there has been little work reported on this issue due to the serious technical challenges. **1) Incomplete prior knowledge on the conditional independency among spatial regions.** The conditional independency among spatial locations signifies whether there exists or not connectivity between any pairs of them. However, due to the limitation of sensing techniques, in many important domains, it is too complex and expensive to investigate the prior knowledge on the fine-grained connectivity among different locations. Such domains include environment science, epidemiology, sociology, criminology, neuroscience, and chemistry, especially for those open and crucial domains where the spatial patterns have not yet been completely figured out. For example, the connectivity of the underground water network could be expensive and typically prohibitive to be investigated. In epidemiology, it is prohibitively difficult to comprehensively figure out the spatial transmissibility correlation among all the locations, as one needs to consider transportations, flight connections, climates, and even the migration of animal species for some pest-borne diseases. In neuroscience, the specific fiber connections among different voxels have not yet been fully determined. Hence, it is very common that the prior knowledge on the existence (or absence) of connections may only be available for a portion or even none among all the locations. **2) the difficulty in optimizing under spatial topological constraints.** Typically, although the detailed knowledge on the spatial connectivity among all the locations is not fully provided, the higher lever spatial topological prior knowledge could be available. For example, although we do not know the detailed connection structure for underground water network, it is relatively easy to obtain knowledge on which subset of locations form connected components and the topological type of each connected component is typically tree structured. This requires us to automatically learn the spatial dependency among all the locations under the spatial topological constraints, which are typically discrete or nonconvex. So adding such discrete constraints over the spatial autoregressive which is typically continuous optimization problem leads to the simultaneous optimization of both continuous and discrete variables, which is extremely challenging for current techniques in a unified framework. A solution that is scalable to large number of locations with theoretical guarantee on the optimality of the solution is highly imperative . **3) the difficulty in balancing the amount of exogenous and endogenous information.** Although too much exogenous information and constraints will make a model rigid and far from being optimal, optimizing spatial dependency based solely on endogenous information on the observation data itself without enough prior knowledge could also cause over-fitting. This is because the number of parameters involved in determining the spatial dependency is quadratic to the number of locations. Thus, the model parameters could easily overwhelm the historical data for even a modest number of locations (e.g., 30). The model needs to find an optimal balance between being "over-constrained" and "under-constrained" by the prior knowledge.

In order to address all the above challenges, this paper proposes a generic Spatial Autoregressive Dependency Learning (SADL) framework to solve a general spatial regression problem with incomplete prior knowledge on spatial connectivity and topological constraints. The SADL framework jointly learns the prediction mapping from inputs to outputs and the spatial dependency, under various amounts of available prior knowledge. When the spatial conditional independency is available which provides the binary information on the existence (or not) of connection between any pair of locations, we propose the model SADL-I that further optimizes the quantitative strength of the spatial dependency among all the locations. Take Figure 2 as an example, here the six locations are treated as six nodes, two nodes have connection if they are contiguous. So the connectivity only tells us which values in adjacency matrix are nonzeros as shown in Figure 2(b). For example, there is no connection between 'a' and 'd' while there is connection between 'a' and 'b'. But although we know the existence of such connection, we still do not know the strength of it. For example, the strength of connectivity between 'a' and 'b' might be different from that between 'c' and 'd'. Therefore, instead of using heuristics to define it agnostic to the prediction tasks, SADL-I is developed to optimize the strength of the connectivity, namely the specific numerical values of those nonzero entries marked by '?' in Figure 2(b). For the cases where the prior information on conditional independency is incomplete and only the higher-level spatial topological constraints are available, we propose another convex model, SADL-II, which learns both the conditional independency as well as the spatial dependency under the given spatial constraints. To make the discrete spatial constraints computationally feasible to solve, we propose their convex and continuous equivalence with theoretical proofs. Effective optimization methods based on Alternating Direction Methods of Multipliers (ADMM) are proposed to obtain the global optimal solutions for both models efficiently with theoretical guarantees. The major contributions of this paper are as follows:

- Propose a novel generic convex framework for simultaneous spatial dependency optimization and spatial regression. To avoid the bias in the existing methods based on predefined spatial dependence, the proposed framework instead optimizes the spatial dependence strength among locations based on spatial conditional independency under spatial constraints. The outputs of the framework provide both high performance and interpretability of the data and locations.
- Propose two novel models SADL-I and SADL-II that address different types of prior knowledge. The first model SADL-I optimizes the strength of spatial dependency based on prior knowledge on the existence (or not) of spatial connections among locations. When such prior knowledge is incomplete or unavailable, the second model SADL-II further embeds higher-level spatial topological constraints to directly learn the conditional independency as well as spatial dependency strength.
- Develop efficient algorithms for optimizing SADL-I and SADL-II. Algorithms based on ADMM are developed to decompose the large optimization problems into smaller subproblems, which are then solved efficiently using their closed forms or projected gradient descent. The newly developed algorithms for SADL-I and SADL-II are theoretically guaranteed to obtain global optimal solutions after convergence.
- Conduct extensive experiments on three real-world data sets against several state-of-the-art methods. The proposed SADL-I and SADL-II methods outperform the best competitor methods by around 10% on average. In-depth discussions on the parameter sensitivity, convergence, and the discovered spatial dependency conclude the study.

2 RELATED WORK

Spatial regression based on spatial correlation. Spatial prediction has long been of interest to academics as a branch of spatio-temporal statistics [11]. It characterizes the correlation among the patterns in different geo-locations and requires the removal of the conventional "independent and identically distributed" assumption on the data. One way to achieve this is through dependency modeling, where the best known models are spatio-autoregressive (SAR) [8, 34, 38, 41], spatial Durbin [29], geographically weighted [7], and spatial X [18]. Researchers have generally assumed that the status of each location is not only determined by its own input, but also by the status of other locations through their spatial auto-correlation patterns. For example, based on the fixed prior knowledge on the spatial dependency such as spatial contiguity, feature weights and trade-off parameter were estimated in the SAR models [8, 38]. In addition to modeling the spatial dependency directly, researchers have also added extra regularization terms to enforce a "smoothing" across different locations. Regularization terms such as spatial entropy [9] and spatial information gain are typically utilized for such classification problems. One common way for the regularization of spatial regression is through graph Laplacian approaches [14], which penalizes the divergence among the parameters of the models for neighboring locations, and thus the spatial correlation of the models for different locations will be enforced. All the above methods typically require the existence of some prior knowledge as a heuristic surrogate of the true spatial correlation strength, such as the spatial adjacency relationship and geographical distance between locations. However, the real spatial correlation strength is typically not the same as the heuristic estimation and will vary across different prediction tasks [39]. Farber et al. [15] focuses on performing statistical analysis on how much two specific network properties influence the performance of three existing models: namely logistic regression (LR), spatial autoregressive (SAR), and Lagrange multiplier spatial lag dependence model (LM-LAG). Most recently, Ziat et al. [46] proposed to infer the spatial correlation in time series data, but only focus on predicting the endogenous input variables in future points. Moreover it cannot consider the topological constraints from spatial prior knowledge when learning the spatial correlation. Qu et al. [36] present model specification and estimation of the SAR model with an endogenous spatial weight matrix which however requires strong assumption on their probabilistic distribution. Kelejian and Prucha [22] estimate the spatial disturbance term that is spatially autoregressive while Li et al. [27] develop a new heuristic to estimate the spatial dependence, as a suitable alternative to Moran's statistic. Otto and Steinert [33] as well as Bhattacharjee and Jensen-Butler [4] both propose two-step lasso estimation approaches to estimate spatial weights matrix, whereas the joint global optimality cannot be guaranteed due to the separate optimization of each step. In addition, Lam and Souza [24] focus on a spatiotemporal model setup with exogenous regressors, where the spatial weight matrix has a block diagonal structure. Lam et al. [25] as well as Ahrens and Bhattacharjee [1] both purely use LASSO for sparsifying spatial dependency matrix without sufficiently utilizing the exogenous information such as the spatial conditional independency or spatial topological constraints.

Graph structure learning. There are basically two types of graph structure learning problems. The first seeks to learn the selection of the nodes in the graph, such as subgraph detection [21], structured feature selection [17], and subgraph clustering [29]. The other type learns the selection and weights of the edges in the graph in order to learn the dependency among the variables [2, 44]. A standard approach for this is based on the classic result that the zeros in the correlation matrix correspond to zero partial correlations among variables. Numerous works have focused on learning the precision matrix for graphical models; this problem is particularly relevant to graphical LASSO, proposed by Banerjee et al. [2] and Friedman et al. [16], and a large variety of alternative penalties have been suggested to extend the priors of graphical LASSO [13, 20, 45]. Loh et al. [28] focused on inferring the graph structure using discrete Markov Random Fields. Instead of learning the precision matrix, other works have directly estimated the Laplacian matrix or the adjacency matrix. For example, [12] developed a fitness metric as the surrogate for learning the graph topology, rather than directly optimizing it. To avoid the bias introduced by the surrogate metric and directly learn the graph topology, Lake et al. [23] proposed learning an adjacency matrix with a regularized Laplacian matrix. Dong et al. [14] proposed learning a smoothing regularizer for a given signal using a nonconvex problem formulation. Unlike the above existing work, this paper focuses on the optimization of the contiguity matrix for spatial locations, where the final output for each location is determined by both the input and the output of other locations. Moreover, additional constraints required for this spatial regression problem includes symmetric, non-negative, zero-diagonal, and spatial topological constraints such as a tree-structure (e.g., a watershed network), a grid-structure (e.g., Manhattan Neighborhood Network), or a connected subgraph (e.g., an electrical grid network).

Traffic flow prediction. Traffic flow is traditionally measured using stationary sensors such as induction loops. Their measurements can be used to train models for the short-term prediction at those locations [26, 32, 40]. This work uses traffic data from various locations to study the spatial correlation of flow and to train auto-regressive dependency learning models based on spatial topological constraints of the road network. The fundamental traffic flow diagram [19] could be used to infer traffic volume from speed and density. However, learning these relationships requires a large amount of training data, which is not always available for all parts of the road network, an important limitation this work is able to overcome.

3 PROBLEM FORMULATION

Denote $X = \{X_t\}_t^{\mathcal{T}}$ as the time-ordered collection of **input data**. Each $X_t \in X$ represents the input data, which are typically observations from sensors, at time t such that $X_t \in \mathbb{R}^{|S|\times|K|}$ is a matrix for all the spatial locations S and all the input features K. Naturally, denote $X_{t,s}$ as the data vector at time t in the location s, and denote $X_{t,s,k}$ as the data point for the feature k in the location s at time t. Similarly, denote $Y = \{Y_t\}_t^T$ as the time-ordered collection of **output data** to be predicted. Each $Y_t \in Y$ denotes the output data at time t such that $Y_t \in \mathbb{R}^{1\times|S|}$. Therefore, $Y_{t,s} \in \mathbb{R}$ denotes the prediction output for location s at time t. In conventional prediction problems, the independent and identity distribution (i.i.d) condition is typically assumed. However, for spatial prediction, i.i.d condition cannot be held according to the first law of geography where "everything is related to everything else, but near things are more related than distant things" [42]. For example, if we want to predict the traffic congestion for a road segment, then the traffic condition of its adjacent road segments is highly important to be considered since they are correlated instead of being independent from each other.

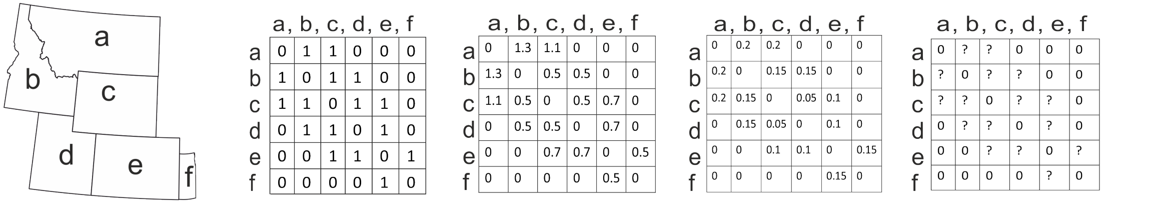
Therefore, the problem of spatial prediction can be formulated as follows:

Problem Formulation: At time τ, given the input signal $X_\tau \in \mathbb{R}^{|S|\times|K|}$ on a set of different locations S with a set of input features K, we aims at predicting the output Y_t at a future time t by learning a predictive model: $f : X_\tau | D, W \longrightarrow Y_t$, where $W \in \mathbb{R}^{1\times|K|}$ encodes the weights of the features to the prediction output, and $D \in \mathbb{R}^{|S|\times|S|}$ encodes the spatial dependencies among all the locations S. For example, $D_{i,s}$ denotes the spatial dependency between locations i and s. Typically D needs to be non-negative and must follow some reasonable spatial topological constraints (such as road network and river network). In addition, the lead time is $p = t - \tau$, which means the time span between the current time τ and future time t.

In essence, the prediction of a spatial location is not only determined by the input of this location, but also the output in other locations: $Y_{t,s} = \sum_{i\in S, i\neq s} \rho \cdot D_{i,s} Y_{t,i} + W \cdot X_{\tau,s} + \varepsilon$, where $D_{i,s}$ denotes the spatial correlation from location i to location s, ρ is the trade-off parameter balancing the contribution of the input and the location correlation to the

Table 1. Important Notations

Notations	Explanations
S	All the spatial locations
T	All the time intervals
$X_{s,t} \in \mathbb{R}^{1\times\lvert K\rvert}$	Input feature vector in location s at time t
$Y_{t,s} \in \mathbb{R}$	Response variable in location s at time t
$D \in \mathbb{R}^{\lvert S\rvert\times\lvert S\rvert}$	Spatial dependency matrix
$W \in \mathbb{R}^{1\times\lvert K\rvert}$	feature weights
G	The sets of all the connected components on spatial dependency
$D_g \in \mathbb{R}^{\lvert g\rvert\times\lvert g\rvert}$	The spatial dependency for a connected component g



(a) spatial locations

	a	b	c	d	e	f
a	0	1	1	0	0	0
b	1	0	1	1	0	0
c	1	1	0	1	1	0
d	0	1	1	0	1	0
e	0	0	1	1	0	1
f	0	0	0	0	1	0

(b) contiguity relation

	a	b	c	d	e	f
a	0	1.3	1.1	0	0	0
b	1.3	0	0.5	0.5	0	0
c	1.1	0.5	0	0.5	0.7	0
d	0	0.5	0.5	0	0.7	0
e	0	0	0.7	0.7	0	0.5
f	0	0	0	0	0.5	0

(c) border length

	a	b	c	d	e	f
a	0	0.2	0.2	0	0	0
b	0.2	0	0.15	0.15	0	0
c	0.2	0.15	0	0.05	0.1	0
d	0	0.15	0.05	0	0.1	0
e	0	0	0.1	0.1	0	0.15
f	0	0	0	0	0.15	0

(d) border ratio

	a	b	c	d	e	f
a	0	?	?	0	0	0
b	?	0	?	?	0	0
c	?	?	0	?	?	0
d	0	?	?	0	?	0
e	0	0	?	?	0	?
f	0	0	0	0	?	0

(e) topological constraint

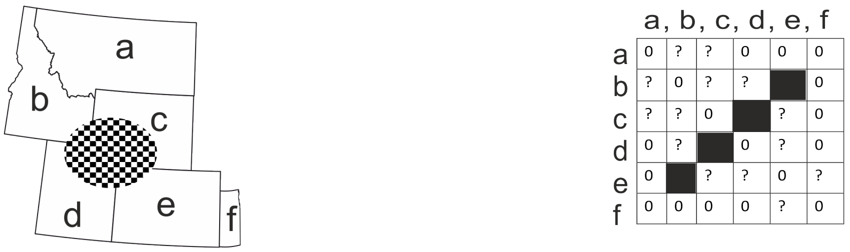
Fig. 3. The correlation matrix of spatial locations, different conventional heuristics to determine the correlation matrix, and an example of topological constraints leveraged by our method.

prediction. Using matrix notation, this notion can be concisely rewritten in the following form:

$$Y_t = \rho \cdot D \cdot Y_t + X_\tau \cdot W^T + \varepsilon, \tag{1}$$
$$\text{where } D \geq \mathbf{0}, \operatorname{diag}(D) = 0$$

which is called *spatial autoregressive model* [23], where $D \geq \mathbf{0}$ denotes that all its elements are nonnegative. $\operatorname{diag}(D) = 0$ means all of its diagonal elements are equal to 0 because a location is not meaningful to be dependent on itself. ε denotes the noise following a Gaussian distribution $\mathcal{N}(0, \sigma)$. The matrix D in the above equation could be further normalized (e.g., row-normalized) to avoid potential the singularity of $I - D$, where I is an identity matrix. However, the framework proposed in this paper can avoid singularity even without row-normalization, which will be introduced in the next section.

Challenges: In the existing works, the spatial dependency matrix D is typically known beforehand. Conventionally, it is determined by external prior knowledge such as the continuity correlation among the locations [29]. These works assume the continuous locations have spatial correlation of "1" if they are contiguous while "0" if they have no contiguity. However, it is prohibitively difficult to accurately determine the spatial dependency only by human prior knowledge and heurstics without bias. Specifically, as shown in Figure 3, there are six spatial locations shown in Figure 3(a) whose spatial contiguity relation is shown by the weight matrix in Figure 3(b). However, it is not reasonable to assume that all the locations share the same strength of spatial correlation. For example, the weight matrix in Figure 3(b) assumes the spatial correlation strengths between locations c and d are equal to that between locations a and b. But obviously the latter has much longer border shared between locations c and d, which cannot be differentiated in Figure 3(b). Figure 3(c), though differentiates the length of the shared border among locations, cannot consider their percentage out of the total lengths. For example, the border between locations c and d, and that between locations e and f are the same in length, but obviously the latter are potentially higher correlated because the percentage of border shared between locations e and f among the total length of all of their borders is much larger than the percentage of border that c and d share. Figure 3(d), though considers the percentage of border, still cannot consider the shapes and other factors that could be critical in determining the spatial correlation strength. **Also, the spatial correlation should differ for**



$$\begin{array}{c|cccccc} & a & b & c & d & e & f \\ \hline a & 0 & ? & ? & 0 & 0 & 0 \\ b & ? & 0 & ? & ? & ■ & 0 \\ c & ? & ? & 0 & ■ & ? & 0 \\ d & 0 & ? & ■ & 0 & ? & 0 \\ e & 0 & ■ & ? & ? & 0 & ? \\ f & 0 & 0 & 0 & 0 & ? & 0 \end{array}$$

(a) spatial locations with incomplete information (b) weight matrix with unknown connectivity pattern

Fig. 4. The information about the spatial contiguity among locations b, c, d, and e is incomplete (marked by the oval grid mask). Correspondingly, the absence of spatial dependency between locations c and d, as well as that between locations b and e is unknown, which is marked as black in the weight matrix D in subfigure (b).

different prediction tasks. For example, when predicting the air pollution, the contiguous locations are related with each other. But for predicting the water pollution, the same spatial locations need to correlate under the constraint of stream networks. **Furthermore, usually the prior knowledge on spatial dependency D is incomplete or only in high-level such as the type of its geographical topology.** For example, in Figure 4, the information about the spatial contiguity among locations b, c, d, and e is incomplete, which is marked as the oval grid mask in Figure 4(a). Correspondingly, the existence or absence of spatial connectivity between locations c and d, as well as that between locations b and e is unknown, which is marked as black in the spatial correlation weight matrix D in Figure 4(b). Therefore, we need to infer the unknown connectivity among the locations. To achieve this, it is crucial to infer whether the unknown connectivity exists (nonzero value in the weight matrix) or not (zero value in the weight matrix), namely to predict whether the black-marked entries in Figure 4(b) is zero or not, and the specific value if nonzero. It is very challenging to instruct the optimization of D under such partial constraints due to its inherent nature related to combinatorial and non-convex optimization which will be addressed by our new algorithms elaborated in Section 4.3.

4 MODELS

In order to address the above challenges, we first propose a generic framework SADL to jointly optimize the spatial dependency as well as the prediction mapping from the input features. By inserting different prior knowledge on the spatial topological constraints in SADL, it leads to two different optimization problems. First, when the existence and absence of connection between any pair of all the spatial locations are known *a priori*, the first model named Spatial Autoregressive Dependency Learning-I (SADL-I) is proposed to learn the numerical strength of these connections. Second, when the connectivity among the locations is unknown, the second model named Spatial Autoregressive Dependency Learning-II (SADL-II) is proposed to automatically learn both the connections and their numerical strength, following the spatial topological constraint.

4.1 New generic framework for spatial dependency learning

The connectivity among different locations commonly exists in reality, such as the connections among the road segments through a road network, the connections among the segments of rivers among a water network, and the contiguity relationship among different provinces in a country. Although the binary value on the existence (or not) of connection is easy to obtain, the strengths of them which are numerical values typically are difficult to be pre-defined accurately. Specifically, we propose a new framework that is able to learn such numerical values in terms of spatial dependency during spatial prediction, given the binary-valued existence of connectivity. To avoid the overspecification of spatial correlation, we propose to leverage topological constraint. Figure 3(e) shows one example of topological constraint, which only specifies which two locations are conditionally independent (shown as "0" in Figure 3(e)) with each other, but empower the model to automatically and adaptively optimizes the values of the strengths of the spatial dependency (namely, the unknown entries with "?" in Figure 3(e)), which is impossible to hand-craft. Our problem is formulated as follows:

This new generic framework can be formulated as the following objective function:

$$\min_{D,W} -\ell\ell(D,W,\sigma^2|Y)+\lambda\mathcal{R}(W), \qquad s.t., D\in\mathcal{G} \tag{2}$$

where $-\ell\ell(D,W,\sigma^2|Y)$ is the negative log-likelihood which minimizes the spatial prediction error. $\mathcal{R}(W)$ denotes the regularization term on feature weight W which enforces feature sparsity and ensures model generalizability. And $D\in\mathcal{G}$ is the spatial topological constraints on D based on external prior knowledge. The detailed explanations about all these terms are as follows:

1) The log-likelihood $\ell\ell(D,W,\sigma^2|Y)$.

$\ell\ell(D,W,\sigma^2|Y)$ is the log-likelihood conditioning on the variable Y. The following introduces the deduction of it. According to Equation (1), we have $\varepsilon=(I-\rho\cdot D)\cdot Y_t-X_\tau\cdot W^\intercal$ which follows the Gaussian distribution $\varepsilon\sim\mathcal{N}(0,\sigma)$, where Y_t denotes the output data at time t. Because we are optimizing D, the scaling factor ρ can be absorbed into D and hence we have $\varepsilon=(I-D)\cdot Y_t-X_\tau\cdot W^\intercal$. Since it is easy to obtain the likelihood $\ell(\sigma|\varepsilon)$ in terms of the variable ε, we can utilize $\ell(\sigma|\varepsilon)$ to calculate $\ell(D,W,\sigma^2|Y)$ using the well-known Jacobian factor [30]. And finally we can obtain the log-likelihood as shown in the below lemma whose detailed deduction is elaborated in the proof in the appendix section.

Lemma 1. *The log-likelihood $\ell\ell(D,W,\sigma^2|Y)$ in Equation (2) is calculated as follows:*

$$\ell\ell(D,W,\sigma^2|Y)=\sum\nolimits_t^\intercal\frac{1}{2\sigma^2}\|(I-D)Y_t-X_\tau W^\intercal\|_2^2) \tag{3}$$
$$+\frac{1}{2}\cdot|S||T|\ln(2\pi\sigma^2)-|T|\ln(\det(I-D))$$

where $\det(x)$ *denotes the determinant of the matrix x. It can be seen that the ρ in Eq. (1) has been absorbed into D. Y is the output data to be predicted and Y_t is the output data for time t.*

Proof. The detailed proof is included in appendix section. □

2) The regularization on feature weights W.

$\mathcal{R}(W)$ is the regularization term, such as the ℓ_1-norm for the feature weights W to enforce reasonable sparsity patterns when W has high dimension. The trade-off parameter λ can be selected by cross-validation as detailed in the experimental section.

3) The spatial topological constraint $D\in\mathcal{G}$.

$D\in\mathcal{G}$ is the spatial topological constraint, which needs to be specified according to the prior knowledge that we have on $\mathcal{G}$. Therefore, in the following two subsections we focus on the two different formulations of $\mathcal{G}$: **1) SADL-I Model**: Section 4.2 describes the situation when the complete information of the connectivity among spatial locations is available; and **2) SADL-II Model**: Section 4.3 elaborates the situation when the connectivity is incomplete or unavailable but the spatial topological constraint is available.

[1] For the basics about this equivalence, please refer to [30] and Section 1.2.1 of [5]

In addition, it can be seen from Equation (27) that, the aforementioned singularity issue of $I - D$ can be avoided in our framework, even without doing row-normalization, thanks to the optimization of D. Specifically, as shown in Equations (27), it minimizes the term of $-|T|\ln(\det(I-D))$, which will avoid $I - D$ to be singular because if it is singular, then the optimization objective will be positive infinity, which is inherently avoided by our minimization process. Moreover, our framework is also generic to, and can easily adopt any normalization of matrix D.

Remarks for the novelty of the proposed SADL framework: The general framework proposed in Equation 2 is new and has never been proposed before, to the best of our knowledge, which is significant and novel in three specific aspects: *1) Automatic inference of the weight matrix D.* Existing work typically predefined the matrix D purely based on human hand-crafting and heuristics. However, as mentioned above, such predefined matrix involves bias and cannot accurately reflect the true spatial correlation specific to the corresponding prediction task. To address this, our framework innovatively enables to optimize the matrix D adaptively to minimize the prediction error, based on our proposed techniques on sparse learning and optimization. *2) Inclusion of the spatial topological constraints.* Currently, there is no existing work that has leveraged the spatial topological constraint for the optimization of weight matrix D, which is important to many applications as mentioned above but is very challenging to be formulated. This is because the graph topological related problems are naturally discrete optimization problem typically with discrete constraints, while the spatial autoregressive is typically continuous optimization problem, and hence it is extremely challenging for current techniques to jointly address these two types of problem. To address this, we innovatively formulate the original discrete formulation of spatial topological constraint into its continuous equivalence by our Theorem 1 in Section 5.2. *3) Optimization of the new objective with spatial topological constraints.* Specifically, existing spatial autoregressive methods typically only optimize the scaling factor ρ (see in Equation (1)) and input feature weight W, which can be solved directly using traditional algorithms such as Expectation-Maximization algorithms. However, the new involvement of the optimization of the additional matrix D with numerous parameters and nonlinear and nonsmooth constraints on spatial topologies requires new efficient and effective algorithm to handle this problem. Accordingly, this paper proposes new optimization methods based on ADMM and proximal operators, which decomposes the original big problem into equivalent subproblems that are then respectively solved by the new algorithms we designed and presented in Section 5.

4.2 SADL-I Model

This section present SADL-I Model which can optimize the (numerical) strength of spatial dependency when information on the existence of connection between any pair of locations is known *a priori*.

In many applications, the existence or not of connection among different locations can be prepared accurately without too much effort. In this situation, it means the sparsity pattern (namely binary information on whether each element is zero or not) of D can thus be provided *a priori*. And thus the optimization task is to learn the specific weights for those existing edges (namely those nonzero entries) which represent the corresponding spatial dependency among different locations. Therefore, we have a constraint $supp(D) = \mathbb{M}$ where $\mathbb{M}$ denotes the indices of all the existing connections among the locations based on prior knowledge. Here the function $supp(x)$ denotes the *support* of a matrix x which is the set of the indices of all the nonzero elements of x.

In addition, spatial dependency is commonly assumed non-negative, namely $D \geq \mathbf{0}$. Also, the concept of spatial dependency exists only between different locations, which indicates the zero values for the diagonal elements of the matrix D. Moreover, spatial dependency is typically undirected, which means that if a location A is correlated with another location B, then B is equally correlated to location A. Therefore, we have $D = D^{\mathsf{T}}$. Thus, the spatial topological constraint can be denoted as:

$$D \in \mathcal{G}, \quad \text{where } \mathcal{G} = \{d | d \geq \mathbf{0}, \text{diag}(d) = \mathbf{0}, d \in \mathbb{R}^{|S| \times |S|}, \\ d = d^{\mathsf{T}}, supp(d) = \mathbb{M}\} \tag{4}$$

And the overall objective function for spatial dependency learning can be written as:

$$\min_{D,W} \sum\nolimits_t^T \frac{1}{2\sigma^2} \|(I - D)Y_t - X_\tau W^{\mathsf{T}}\|_2^2) \\ -|T| \ln(\det(I - D)) + \lambda \mathcal{R}(W), \\ s.t., D \geq \mathbf{0}, \text{diag}(D) = 0, D = D^{\mathsf{T}}, supp(D) = \mathbb{M} \tag{5}$$

where the hyper-parameters λ and σ can balance the importance of different terms. These hyper-parameters are determined based on cross-validation in the experimental section.

Lemma 2. *The optimization problem in Equation (5) is convex.*

Proof. The sufficient condition of the convexity of the first term is that the term $\|DY_t + X_\tau W^{\mathsf{T}}\|_2^2$ is convex, which is equal to prove the equivalent term $\|mat(Y_t) \cdot vec(D) + X_\tau W^{\mathsf{T}}\|_2^2$, where $vec(D) \in \mathbb{R}^{|S| \cdot |S| \times 1}$ is the flattened vector form of the matrix D such that $[vec(D)]_{(i-1) \cdot |S| + j, 1} = D_{i,j}$, and $mat(Y_t) \in \mathbb{R}^{|S| \times |S| \cdot |S|}$ is the matrix form of the vector $Y_t \in \mathbb{R}^{|S| \times 1}$ such that $[mat(Y_t)]_{i,(i-1) \cdot |S| + j} = [Y_t]_{j,1}$ and all the other elements of $mat(Y_t)$ are zeros. And it is easy to see that $mat(Y_t) \cdot vec(D) = D \cdot Y_t$ and that $\|mat(Y_t) \cdot vec(D) + X_\tau W^{\mathsf{T}}\|_2^2$ is convex due to the fact that the Hessian matrix with respect jointly to $vec(D)$ and W is positive semi-definite. Therefore, the first term in Equation (5) is convex. Moreover, the logarithm determinant term $|T| \ln(\det(I - D))$ is well-known to be concave. Finally, because in Equation (5), all the equality constraints are affine while all the inequality constraints are convex, the proof is completed. □

In Section 5, our newly proposed algorithm is able to address this problem with a global optimal solution.

4.3 SADL-II Model

For the situations when the information on the conditional independence among spatial locations is unknown or incomplete where SADL-I cannot apply. To address it, in this subsection, we first present the motivation of the SADL-II model, then propose our new equivalent formulation of spatial topological constraints using Theorem 1, and finally present the concrete mathematical formulation of our SADL-II model in Equation (6).

In general, a wide variety of applications suffer from the incomplete knowledge on the connectivity among spatial regions, such as in the domains of environment science, epidemiology, sociology, criminology, neuroscience, and chemistry, especially for those open and crucial domains where the spatial patterns have not been completely figured out. For these domains, it is usually technically infeasible or too complex and expensive to investigate the prior knowledge on the fine-grained connectivity among different locations. For example, the connectivity of the underground water network could be expensive and typically prohibitive to be investigated. In most situations, although we do not know the detailed connection structure, it is easy to obtain knowledge on which subset of locations form connected components and what is the topological type of each connected component. In epidemiology domain, the spatial disease-transmissibility correlation among different locations are crucial in determining the spread of epidemics, which, however, is typically too complex to be 100% completely observed and modeled. For example, to comprehensively consider the spatial transmissibility correlation among locations, one needs to consider transportations, flight connections, climates, and even the migration of animal species for some pest-borne diseases. In neuroscience, brain connectome is a famous spatial network which can be considered as a network of voxels (i.e., spatial regions), and different voxels can have connections if there are fibers between them. However, currently the state-of-the-arts are still far from being able to figure out all the connectivity between voxels and among all the neurons. However, a spatial topological constraint is typically known: “the brain network (e.g., structural connectome) should be a connected graph)”. In biochemistry, being determined by protein folding, the spatial structure of protein molecular is crucial for determining the protein function and hence important for crucial domains such as drug design. This directly motivates the research on protein structure prediction whose goal is just to predict the unknown spatial proximity among atoms, which is still an open question. Here a spatial topological constraint is: “a protein must be a connected graph with backbone-chain structure”. Spatial topological constraints are also available for many other domains. For example, given a set of watersheds where we want to predict the water quality, we know each watershed it is typically a connected component and tree-structured. Given a set of streets in areas of different cities, we know each of them is a connected graph under a “street grid” [18]. This type of knowledge is extremely instructive for optimizing the spatial dependency under a correct spatial structure.

Therefore, we are given prior knowledge on G which is the list of the location sets in all the different connected components, where each $g \in G$ is a subset of locations such that $g \subseteq S$. Concretely, this prior knowledge is equivalent to the joint satisfaction of both the following two properties: ***1) Property 1:*** *Locations in each connected components have path(s) to each other.* For each pair of locations i and j in a connected component $g \in G$, they must have at least one path to each other. ***2) Property 2****: No connections among locations from different connected components.* For location $i \in g$ and location $j \in h$ where $g \in G$ and $h \in G$ are two different connected components, there is no path between locations i and j.

So the crux now is how to embed such prior knowledge into our framework in Equation (2), namely how to mathematically encode such information of connected components in terms of the constraints upon $D \in \mathcal{G}$. This is a new and challenging problem. Specifically, the constraints to enforce connected components is inherently discrete. And enforcing the connectivity for each component is naturally a discrete optimization problem with discrete constraints which typically requires graph theory discrete algorithm. On the other hand, the spatial autoregressive is typically continuous optimization problem which relies on maximal likelihood and gradient descent. However, our model requires the simultaneous optimization of both of them together, and hence it is extremely challenging for current techniques to jointly address these two types of problem in a unified framework. To address this issue, we innovatively propose a convex formulation for $D \in \mathcal{G}$ which exactly encodes the prior knowledge on the connected components, as shown in Theorem 1.

Theorem 1 (connected-component constraints). *A graph has a set of connected components G if and only if the following two conditions are satisfied:*

- *$diagm(D_g \cdot \mathbf{1}^\top) - D_g + \mathbf{1}^\top \mathbf{1}/|g| \succ 0, D_g \geq \mathbf{0},\ g \in G$, which formulates the above* ***Property 1****.*
- *$D[i,j] = 0, \forall\ i \in g, j \in h, g \neq h, g, h \in G$, which formulates the above* ***Property 2****.*

where $D[i,j]$ denotes the connectivity of locations i and j. D_g is the adjacency matrix only for those locations in g. Thus D_g is a block matrix of D. $\mathbf{1} \in \mathbb{R}^{1 \times |g|}$ denotes an all-one vector. $diagm(x)$ denotes the diagonal matrix with elements as the vector x. The symbol $\succ 0$ denotes the positive definite property.

Proof. First, if there are $|G|$ connected components, then there must be no connections across different components such that $D[i,j] = 0,\ i \in g, j \in h, g \neq h, g, h \in G$. Then in the following, we prove the connectedness of each components.

The proof is based on the property of graph connectedness constraint [42]. Define $Z_g \equiv \mathrm{diagm}(D_g \cdot \mathbf{1}^\top) - D_g + \mathbf{1}^\top \mathbf{1}$. We first prove for any $x \neq 0$, we have $x^\top Z_g x > 0$. This is equivalent to proving $x^\top \cdot Z_g x = \sum_{i \neq j} D_{g,i,j}(x_i - x_j)^2/2 + 1/|g| \cdot (\sum_i^{|g|} x_i)^2 \neq 0$. Assuming $x^\top \cdot Z_g \cdot x = 0$, this means that $\sum_{i=1} x_i = 0$ and $x_i = x_j$ for all i and j such that $D_{g,i,j} > 0$. Then x must be 0 if there is only one component in the components consisting of g. Next, assume the components has at least two components, then we have two sets c and e such that $c \cup e = g, c \cap e = \emptyset$ and $\forall (i,j) \in c \times e : D_{g,i,j} = 0$. By setting $x_i = 1/|c|$ and $x_j = -1/|e|$ we have $x^\top \cdot Z_g x = (|c|/|c| - |e|/|e|)^2/|g| = 0$. The proof is completed. □

When there is only one connected-component, namely all the locations have path(s) two each other, it is easily seen that our Theorem 1 is still valid.

Applying the formulated spatial topological constraint in Theorem 1, we obtain the objective function for SADL-II:

$$\begin{gathered}\min\nolimits_{D,W} - \ell\ell(D, W, \sigma^2|Y_g) + \lambda\mathcal{R}(D), \\ s.t., \mathrm{diagm}(D_g \cdot \mathbf{1}^\top)/|g| - D_g + \mathbf{1}^\top \mathbf{1} > 0, D_g \geq \mathbf{0}, \mathrm{supp} \subseteq \mathbb{M}' \\ D[i,j] = 0,\ i \in g, j \in h, g \neq h, g, h \in G, \\ \mathrm{diagm}(D_g) = 0,\ D_g = D_g^\top \end{gathered} \tag{6}$$

$\mathbb{M}'$ denotes the partial prior knowledge on the connectivity among locations, which can be incomplete. When $\mathbb{M}'$ is all-one matrix, there is not any prior knowledge on connectivity available. To solve Equation (6), a new algorithm is proposed to obtain the global optimal solution for in Section 5.

5 OPTIMIZATION ALGORITHMS

Both of the proposed models SADL-I and SADL-II are not easy to solve efficiently based on existing convex optimization algorithms, due to the existence of the determinant term and several constraints on D. In the following, the two algorithms for SADL-I and SADL-II are proposed and described.

5.1 Algorithm for SADL-I

In the objective function of SADL-I, namely Equation (5), there are two variables with several constraints. In order to optimize it efficiently, Alternating Direction Method of Multipliers (ADMM) [6] has been utilized. First, we transfer the original problem into an equivalent formulation, as follows:

$$\begin{gathered}\min_{D,W} \sum\nolimits_t^\top \frac{1}{2\sigma^2} \|(I-D)Y_t - X_\tau W^\top\|_2^2) \\ - |T| \ln(\det(U)) + \lambda\mathcal{R}(V) \\ s.t., D \geq \mathbf{0}, \mathrm{diag}(D) = 0, U = I - D, \\ V = W, D = D^\top, supp(D) = \mathbb{M}\end{gathered} \tag{7}$$

In ADMM, the constrained optimization problem needs to be formulated in forms of augmented Lagrangian, as follows:

$$\begin{aligned} L_p = &\sum\nolimits_t^\top \frac{1}{2\sigma^2} \|(I-D)Y_t - X_\tau W^\top\|_2^2) - |T| \ln(\det(U)) + \\ &\lambda\mathcal{R}(V) + \frac{\gamma}{2}\|U - I + D + \Gamma_1\|_F^2 + \frac{\gamma}{2}\|V - W + \Gamma_2\|_F^2 \\ &- (\gamma/2)(\|\Gamma_1\|_F^2 + \|\Gamma_2\|_F^2)\end{aligned} \tag{8}$$

where γ is the penalty term which is typically initialized as 1 [6]. And Γ_1 and Γ_2 are dual variables.

Algorithm 1 Parameter Optimization for SADL-I

```
Require: $X$, $Y$, $\sigma$, $\lambda$, and $\mathbb{M}$.
Ensure: solution $W$, $V$, $U$, and $D$.
 1: Initialize $\gamma = 1$, $W, U, V, D = \mathbf{0}$.
 2: Choose $\varepsilon_p > 0$ and $\varepsilon_d > 0$.
 3: repeat
 4:     $W \leftarrow$Equation (10).
 5:     $V \leftarrow$Equation (11).
 6:     $U \leftarrow$Equation (12).
 7:     repeat
 8:         $\Delta D = D - \eta \nabla H$
 9:         $D \leftarrow ((\Delta D/2 + \Delta D^\top/2)_+ \odot D_M)_+$
10:     until Convergence
11:     Calculate the primal residual $p$ and dual residual $d$ according to [6].
12:     if $r > 10d$ then
13:         $\gamma \leftarrow 2\gamma$                                # Update penalty parameter
14:     else if $10r < d$ then
15:         $\gamma \leftarrow \gamma/2$                               # Update penalty parameter
16:     else
17:         $\gamma \leftarrow \gamma$                                 # Update penalty parameter
18:     end if
19: until $p < \varepsilon^p$ and $d < \varepsilon^d$
```

Then the variables U, V, W, and D are optimized iteratively until convergence by fixing other variables. In our algorithm, we provide the situation when the regularization term is an ℓ_1 norm such that $\mathcal{R}(V) = \|V\|_1$, but the algorithm is generic for other popular norms. The procedure of the algorithm for SADL-I is illustrated in Algorithm 1. The algorithm is initialized in Lines 1-2. And then the parameters W, V, U, and D are optimized iteratively in Lines 3-10. D is optimized by projected gradient descent where the proximal gradient is calculated in Lines 8-9. Then the primal and dual residuals are calculated [6], which are used to update the γ in Lines 11-18 and determine the termination of the iterations in Line 19. The solutions for the subproblems are described in the following.

1. Update W. The update of W amounts to the following subproblem:

$$\min_W \sum\nolimits_t^\top \frac{1}{2\sigma^2}\|(I-D)Y_t - X_\tau W^\top\|_F^2 + \\ \|W - V + \Gamma_2\|_F^2 \tag{9}$$

The optimization of W has an analytical solution as follows:

$$W = [\frac{1}{\sigma^2}\sum\nolimits_t^\top Y_t^\top (I-D)^\top X_t + \gamma(V - \Gamma_2)] \\ \cdot (\frac{1}{\sigma^2}\sum\nolimits_t^\top X_t^\top X_t + \gamma I)^{-1} \tag{10}$$

2. Update V. Update V amounts to the following subproblem.

$$\min_V \|W - V + \Gamma_2\|_F^2 + \|V\|_1 \tag{11}$$

which has analytical solution based on soft-thresholding [6].

3. Update U**.** The optimization problem for updating U is:

$$\min_U \frac{\gamma}{2}\|I - D - U + \Gamma_1\|_F^2 - \ln|U| \tag{12}$$

Let $PQP^\top$ denote the eigendecomposition of $\gamma(I - D - \Lambda_1)$, therefore:

$$U = \gamma/2 \cdot P(Q + (QQ + 4\gamma I)^{1/2})P^\top \tag{13}$$

4. Update D**.** The optimization problem for updating U is:

$$\begin{aligned}\min_D H(D) = &\sum\nolimits_t^\top \frac{1}{2\sigma^2} \|(I - D)Y_t - X_\tau W^\top\|_2^2) \\ &+ \frac{\gamma}{2} \|U - I + D + \Gamma_1\|_F^2 \\ s.t.,\ &D \geq \mathbf{0}, \operatorname{diag}(D) = 0, D = D^\top, supp(D) = \mathbb{M}\end{aligned} \tag{14}$$

The above problem can be easily solved by projected gradient descent [10], where for each gradient step we calculate the following projected gradient:

$$\begin{gathered}\operatorname{proj}_0(D - \eta\nabla H), \\ \text{where } \operatorname{prox}_0(x) = ((x/2 + x^\top/2)_+ \odot D^{(M)})_+\end{gathered} \tag{15}$$

where η is the step size for each iteration of the projected gradient descent. $(x)_+$ denotes an operation on a matrix x which maps those negative elements to 0's while it retains the non-negative elements' values. $\odot$ denotes the element-wise multiplication between two matrices. $D^{(M)}$ is the binary matrix that satisfies both $\operatorname{diag}(D) = 0$ and $supp(D) = \mathbb{M}$.

Since the problem in Equation (5) is convex according to Lemma 2 and the optimal solution for each subproblem can be obtained, the algorithm for SADL-I will converge to global optimal solution based on the ADMM convergence properties for convex problems with simple equality constraints [6].

5.2 Algorithm for SADL-II

The equivalence form of the original objective function for SADL-II model in Equation (6) is as follows:

$$\begin{aligned}\min_{D,W,V,E,U} &\sum\nolimits_t^\top \frac{1}{2\sigma^2} \|(I - D)Y_t - X_\tau W^\top\|_F^2 + \frac{|S|}{2}\ln(2\pi\sigma^2) \\ &- \ln|U| + \lambda\|V\|_1 + \|D\|_1 \\ s.t.,\ &I - D = U, \operatorname{diagm}(D_g \cdot \mathbf{1}) - D_g + \mathbf{1}^\top\mathbf{1}/|g| - \epsilon I = E_g \\ W = V, &\operatorname{supp} \in \mathbb{M}', E_g \geq 0, D \geq \mathbf{0}, \operatorname{diag}(D_g) = 0, g \in G\end{aligned} \tag{16}$$

The augmented Lagrangian of the Equation (16) is as follows.

$$\begin{aligned}L_\gamma = &\frac{1}{T}\sum\nolimits_t^\top \frac{1}{2\sigma^2} \|(I - D)Y_t - X_\tau W^\top\|_F^2 + \frac{|S|}{2}\ln(2\pi\sigma^2) \\ &- \ln|U| + \lambda\|V\|_1 + \frac{\gamma}{2}\|I - D - U + \Lambda_1\|_F^2 - (\gamma/2)\|\Lambda_1\|_F^2 \\ &+ \frac{\gamma}{2}\sum\nolimits_g^G \|\operatorname{diagm}(D_g \cdot \mathbf{1}) - D_g + \mathbf{1}^\top\mathbf{1} - \epsilon I - E_g + \Lambda_2\|_F^2 \\ &+ \|W - V + \Lambda_3\|_F^2 + \|D\|_1 - (\gamma/2)(\|\Lambda_2\|_F^2 + \|\Lambda_3\|_F^2)\end{aligned}$$

where Λ_1, Λ_2, and Λ_3 are the dual variables. Similar to Section 5.1, the augmented Lagrangian will be solved by alternately solving five subproblems of U, V, W, E, and D until convergence, as summerized in the pseudo-code in Algorithm 2. The subproblems of U, V, and W in Lines 5-7 are the same as those described in Section 5.1 and thus they are omitted in this section. However, the solving of the subproblems of D and E is different from that of SADL-I. Similar

Algorithm 2 Parameter Optimization for SADL-II

```
Require: $X$, $Y$, $\sigma$, $\lambda$, $G$, and $\mathbb{M}'$.
Ensure: solution $W$, $V$, $U$, $E$, and $D$.
 1: Initialize $\gamma = 1$, $W$, $U$, $V$, $D_g = \mathbf{0}$, $g \in G$.
 2: Initialized $E = \text{diagm}(D_g \cdot \mathbf{1}) - D_g + \mathbf{1}^\top\mathbf{1} - \epsilon I$.
 3: Choose $\varepsilon_p > 0$ and $\varepsilon_d > 0$.
 4: repeat
 5:     $W \leftarrow$Equation (10).
 6:     $V \leftarrow$Equation (11).
 7:     $U \leftarrow$Equation (12).
 8:     for $g \in G$ do
 9:         repeat
10:             $\Delta D_g = D_g - \eta \nabla H(D_g)$
11:             $D_g \leftarrow ((\Delta D_g/2 + \Delta D_g^\top/2)_+ \odot D^{(M)})_+$
12:         until Convergence
13:         repeat
14:             $\Delta E = \text{diagm}(D_g) \cdot \mathbf{1}^\top - D_g + \mathbf{1}^\top\mathbf{1} - \epsilon \cdot I_g + \Lambda_{2,g}$
15:             $D_g \leftarrow \text{prox}_{\geq}(\Delta E)$
16:         until Convergence
17:     end for
18:     Calculate the primal residual $p$ and dual residual $d$ according to [6].
19:     if $r > 10d$ then
20:         $\gamma \leftarrow 2\gamma$                                  # Update penalty parameter
21:     else if $10r < d$ then
22:         $\gamma \leftarrow \gamma/2$                                 # Update penalty parameter
23:     else
24:         $\gamma \leftarrow \gamma$                                   # Update penalty parameter
25:     end if
26: until $p < \varepsilon^p$ and $d < \varepsilon^d$
```

to Algorithm 1, the update of γ and stop criteria are illustrated in Lines 18-26. Both subproblems of D and E are solved based on projected gradient descent in Lines 9-12 and Lines 13-16, respectively, and are elaborated in the following.

1. Update D**.** According to Theorem 1, D can be partitioned into several block matrices corresponding to all the different connected components. Thus the optimization of D is broken into the following subproblems on each different block matrix D_g. Similarly, in the following, a symbol with a subscript "g" also denotes its block matrix corresponding to that of D_g. For example, U_g denotes the block matrix the original matrix U which corresponds to those locations $g \in G$ which forms a connected component. The objective funtion for D_g is as follows.

$$\min_{D_g \geq \mathbf{0}, D_g = D_g^\top, D_{g,M}=0} H(D_g) \tag{17}$$

where

$$\begin{aligned} H(D_g) = \frac{\gamma}{2} \|U_g - I_g + D_g + \Lambda_{1,g}\|_F^2 \\ + \frac{\gamma}{2} \|\text{diagm}(D_g \mathbf{1}^\top) - D_g + \mathbf{1}^\top \mathbf{1}/|g| - \epsilon I_g - E_g + \Lambda_{2,g}\|_F^2 \\ + \sum\nolimits_t^\top \frac{1}{2\sigma^2} \|(I_g - D_g)Y_g - X_g W^\top\|_F^2 \end{aligned} \tag{18}$$

The above is a convex objective with convex constraint, which can be efficiently solved by projected gradient descent, where the gradient for each iteration is calculated as follows:

$$\begin{aligned}\nabla_{D_g} H(D_g) = \gamma \cdot D_g \cdot \mathbf{1}^\top \mathbf{1} + \gamma \cdot \mathrm{diagm}(\mathrm{diag}(C_g)) \cdot \mathbf{1}^\top \mathbf{1}\\ +\gamma(U_g - I_g + \Lambda_{1,g}) - \gamma \cdot (\mathrm{diag}(D_g)\mathbf{1}^\top \mathbf{1} + \mathrm{diagm}(D_g \cdot \mathbf{1}^\top))\\ +\frac{1}{\sigma^2}\sum\nolimits_t^\top (D_g Y_{g,t}^\top Y_{g,t} Y_{g,t}^\top Y_{g,t} - X_{g,t} W^\top Y_{g,t}) + 2\gamma D_g - \gamma C_g\end{aligned}$$

where $C_g = \mathbf{1}^\top \mathbf{1}/|g| - \epsilon I_g - E_g + \Lambda_2$. Therefore the update of D_g is denoted as follows:

$$D_g \leftarrow \mathrm{prox}_0(D_g - \eta \cdot \nabla_{D_g} H(D_g)) \tag{19}$$

where the projection prox_0 is defined in Equation 15.

2. Update E. The update of E amounts to the following optimization problem:

$$\begin{aligned}\min_{E_g \geq \mathbf{0}} P(E_g) = \frac{\gamma}{2}\|\mathrm{diagm}(D_g) \cdot \mathbf{1}^\top - D_g\\ + \mathbf{1}^\top \mathbf{1}/|g| - \epsilon I_g - E_g + \Lambda_{2,g}\|_F^2\end{aligned} \tag{20}$$

which can be written as proximal operator:

$$\begin{aligned}E_g \leftarrow \mathrm{prox}_{\geq}(\mathrm{diagm}(D_g) \cdot \mathbf{1}^\top - D_g\\ + \mathbf{1}^\top \mathbf{1}/|g| - \epsilon \cdot I_g + \Lambda_{2,g})\end{aligned} \tag{21}$$

where $\mathrm{prox}_{\geq}(\cdot)$ is the projection of update of E_g denoted:

$$\mathrm{prox}_{\geq}(A) = \sum\nolimits_i^n (\lambda_i)_+ u_i u_i^\top \tag{22}$$

where the vector λ_j and u_j are the jth ($j = 1, 2, \cdots, n$) eigenvalue and eigenvector such that $A \cdot u_j = \lambda_j \cdot u_j$. And $A = \sum_j^n \lambda_j u_j \cdot u_j^\top$ is called the eigenvalue decomposition.

The optimization problem in Equation 16 is convex because: 1) the optimization objective is convex similar to the proof of Lemma 2, 2) all the equality constraints are affine, and 3) the inequality constraint $E_g \succeq 0$ is well-known to be convex. Therefore, the proposed algorithm for SADL-II converges to global optimal solution due to ADMM's convergence properties for convex problems with simple equality constraints [6].

6 EXPERIMENTS

In this section, the performance of the proposed models SADL-I and SADL-II is evaluated using several real-world datasets from different domains. In the following, we first introduce the experimental setup. The effectiveness of the proposed models is then evaluated against several existing methods. Finally, an analysis of the discovered spatial dependency patterns is presented. All experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU@3.40GHz) and 16.0GB memory.

6.1 Experimental Settings

6.1.1 Datasets and Metrics. The experimentation uses three different datasets related to three different domains: influenza outbreaks, traffic volume, and water quality.

1) **Influenza outbreak dataset.** The task for this dataset is to forecast the spatio-temporal influenza outbreaks based on social media data. The input data consists of a set of tweets from Jan 1, 2011 to May, 2015 for the United States.

Table 2. Datasets Descriptions

Dataset	Time Period	Sample Rate	Locations	Features
Influenza dataset	2011-01-01 to 2015-05-01	weekly	major states in the United States	525 keywords
Traffic flow dataset	2017-01-02 00:00 to 2017-02-01 00:15	quarter-hourly	37 road segments	47 traffic features
Water quality dataset	2016-06-28 to 2018-01-01	daily	36 sites in Georgia	12 water indices

And each location is a state. Each of these tweets must contain at least one of 525 predefined flu-related keywords (e.g., “cold", “fever", “cough") provided by [35]. The data is partitioned into a sequence of week-interval bins for week-wise forecasting. The predictions were validated against the flu statistics reported by the Centers for Disease Control and Prevention (CDC). CDC publishes weekly influenza-like illness (ILI) population size within each state in the United States using the proportion of outpatient visits to healthcare providers for ILI. The task is to predict for each state the size of ILI Population in the next week. An example of a ground truth ILI population size is a tuple: {’State’: ’New York’, ’Week’: ’01-09-2013 to 01-15-2013’,’ILI Population Size’: 657}. The contiguity relationship among the US is utilized to form the default contiguity matrix for the proposed and the comparison methods.

2) **Traffic volume dataset.** The task for this dataset is to forecast the spatio-temporal traffic volume based on the historical traffic volume and other features in neighboring locations. Specifically, the traffic volume is measured every 15 minutes at 36 sensor locations along two major highways in Northern Virginia/Washington D.C. capital region. The 47 features include: 1) the historical sequence of traffic volume sensed during the 10 most recent sample points (10 features), 2) week day (7 features), 3) hour of day (24 features), 4) road direction (4 features), 5) number of lanes (1 feature), and 6) name of the road (1 feature). The goal is to predict the traffic volume 15 minutes into the future for all sensor locations. With a given road network, we know the spatial connectivity between sensor locations. While traditional approaches train regression or ARIMA models for short-term traffic volume prediction of each road segment separately [26, 32, 40], we study the spatial correlation of flow and we train auto-regressive dependency learning models based on the topological constraints of the road network.

3) **Water quality dataset.** Here we want to forecast the spatio-temporal water quality in terms of the “power of hydrogen (pH)” value for the next day based on the input data, which is the historical data of other water measurement indices. The input data consists of daily samples for 36 sites, providing measurements related to pH values in Georgia, USA. The input features consist of 12 common indices including volume of dissolved oxygen, temperature, and specific conductance [2]. This dataset is published by the United States Geological Survey[3]. Due to the complexity of the water system, there is no prior knowledge on the specific connections among all the sites through water streams, i.e., spatial connectivity. High-level prior spatial knowledge is provided based on the water system they belong to, including the water system of Atlanta, the watershed of the Savannah River, and the watershed of the Timmons River [2].

In the experiments, all the input data in all the datasets has been normalized by zero mean and one standard deviation. All the performance of all the methods are compared under the metrics of Root-Mean-Square Error (RMSE $= \sqrt{\frac{1}{n} \cdot \sum_{t=1}^{T} (Y_t - \tilde{Y}_t)^2}$), Normalized Root-Mean-Square Error (NRMSE=RMSE/$average(Y)$), Mean Absolute Error (MAE=$\frac{1}{n} \sum_{t=1}^{T} |Y_t - \tilde{Y}_t|$), and Normalized Mean Absolute Error (NMAE=RMAE/$average(Y)$), where $\tilde{Y}_t$ is the set of predictions for all the locations for a future time point t, and Y_t is the corresponding labeled ground truth, $average(Y)$ returns the average value of all the labeled ground truth. For each of all the datasets, the data is evenly split in two time spans, where the data of the first time span is training set while that of the second time span is for testing.

[2] Refer to: https://waterdata.usgs.gov/nwis/dv/?referred_module=qw

[3] USGS: https://www.usgs.gov/. Accessed Feb, 2018.

Table 3. Spatiotemporal prediction performance for all the methods in all the datasets

	Flu outbreak				Traffic volume				Water quality			
	RMSE	NRMSE	MAE	NMAE	RMSE	NRMSE	MAE	NMAE	RMSE	NRMSE	MAE	NMAE
MLR	0.0659	0.5050	0.0415	0.3180	0.0454	0.1739	0.0310	0.1188	0.0120	0.0460	0.0072	0.0276
LASSO	0.0540	0.4138	0.0341	0.2614	0.0461	0.1766	0.0312	0.1195	0.0240	0.0092	0.0218	0.0835
SAR	0.0587	0.4498	0.0331	0.2536	0.0492	0.1885	0.0336	0.1287	N/A	N/A	N/A	N/A
SAR L1	0.0560	0.4492	0.0341	0.2614	0.0454	0.1739	0.0309	0.1184	0.0279	0.1069	0.0264	0.1011
GL	0.0608	0.4660	0.0372	0.2850	0.0464	0.1778	0.0317	0.1215	N/A	N/A	N/A	N/A
SADL-I	**0.0485**	**0.3716**	0.0305	0.2338	0.0455	0.1743	0.0309	0.1184	N/A	N/A	N/A	N/A
SADL-II	0.0505	0.3870	**0.0266**	**0.2038**	**0.0448**	**0.1716**	**0.0308**	**0.1180**	**0.0115**	**0.0441**	**0.0068**	**0.0261**

6.1.2 Comparison Methods. In the experiment, the performance of the two proposed methods is compared to well-established state-of-the-art methods for spatiotemporal forecasting. For all the methods that have tunable parameter(s), 5-fold cross-validation is performed on the training dataset. The parameter combinations with the best performance was adopted in subsequent experimentation.

Multivariate Linear Regression (MLR) [30] learns a linear mapping from each multivariate input to each prediction.

LASSO [6]. LASSO is an MLR with an ℓ_1 regularization over the weights of the input features. The trade-off between the empirical loss and the regularization term is a key parameter. 5-fold cross validation was performed to select the parameter in a large range of 22 candidate values among $2^{-10,-9,\cdots,9,10}$ and 0.

Spatial Autoregressive Model (SAR) [8, 38]. SAR predicts by jointly considering the input data and the spatial dependency among all the locations. The contiguity matrices of the first and second datasets were utilized as one input of SAR.

Graph Laplacian regularized Linear Regression
(GL) [14]. GL utilizes the spatial dependency to enforce a smooth pattern of the spatial prediction. The adjacency matrix of GL is designated as the contiguity matrix of each of the first and second datasets. The trade-off parameter between the empirical loss and the regularization term is tuned based on 5-fold cross validation over the range of 22 candidate values among $2^{-10,-9,\cdots,9,10}$ and 0.

ℓ_1-regularized Spatial Autoregressive Model
(LSAR) [37] LSAR is the baseline of this paper. Different from the SAR model, LSAR does not utilize a predefined contiguity matrix, but optimizes the contiguity matrix based on the sparsity assumption. The trade-off parameter between the empirical loss and the sparsity regularization term is tuned again based on the 5-fold cross validation over a range of 22 candidate values among $2^{-10,-9,\cdots,9,10}$ and 0.

SADL-I & SADL-II - the proposed methods. SADL-I utilizes the predefined contiguity matrix as a constraint, while SADL-II does not, but instead only utilizes the grouping information of the different locations. There are two parameters to tune, σ^2 and λ. They were tuned jointly based on 5-fold cross validation over a range of 12 candidate values among $2^{-10,-8,\cdots,8,10}$ and 0 for each. The parameter sensitivity analyses are provided in Section 6.2.2.

6.2 Performance and Discussions

The following sections discuss the performance comparison, convergence and sensitivity analyses, and the discovered spatial dependency patterns of our methods.

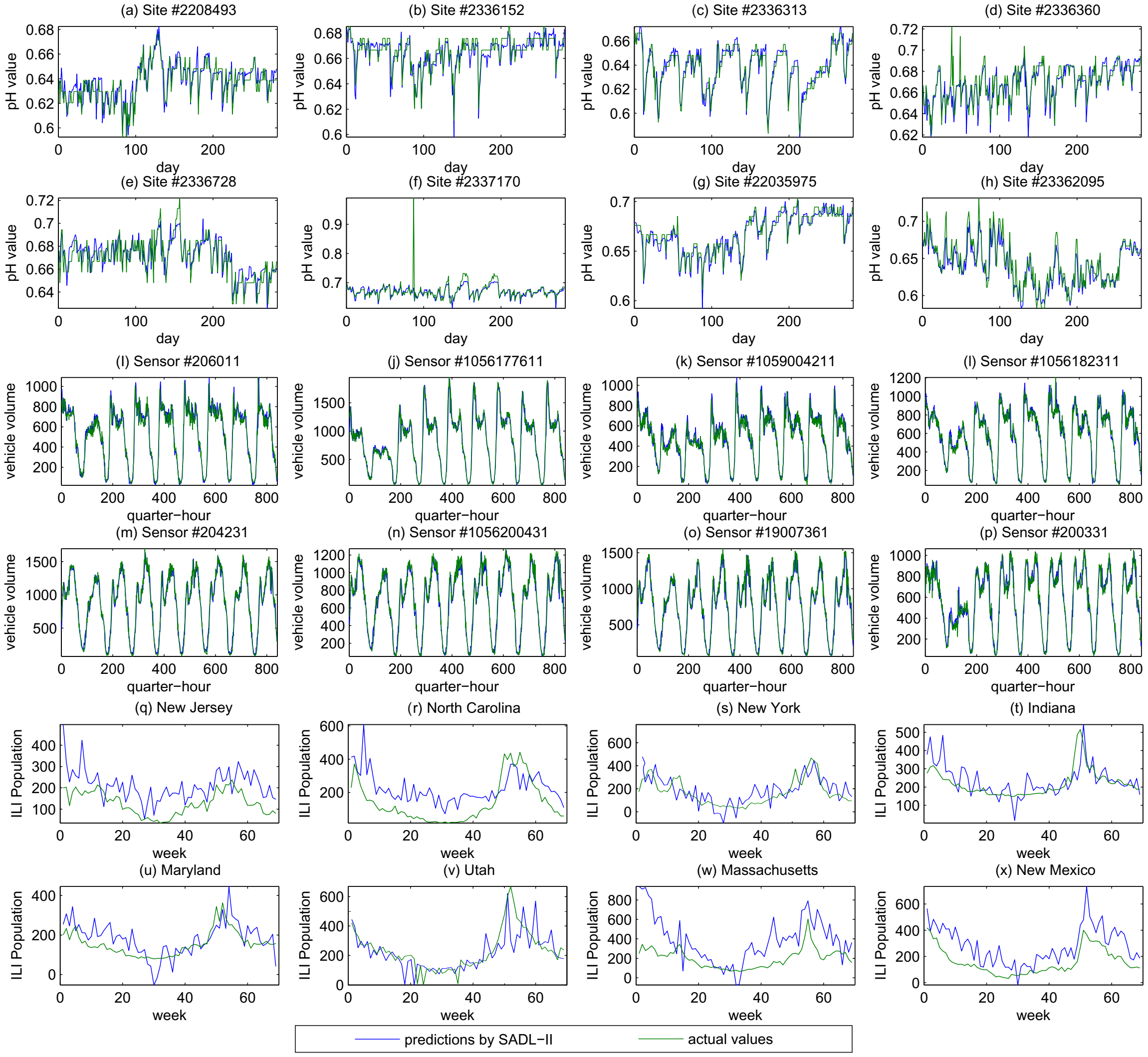


Fig. 5. SADL-II's predictions vs. ground truth for parts of the three datasets. (a)-(h) water quality data; (i)-(p) traffic data; and (q)-(x) flu outbreak data. (Better see in color.)

6.2.1 Prediction performance. As shown in Table 3, four different metrics have been utilized to evaluate the performance on three datasets, RMSE tends to highlight the large errors among the predictions while MAE directly shows the average errors of all the predictions. Additionally, as their normalized versions, NRMSE and NMAE enables the comparison of the performance across different datasets with different scales. The performance of both proposed methods SADL-I and SADL-II outperform the comparison methods by an obvious margin, which highlights the effectiveness of spatial dependency inference for event forecasting. Specifically, SADL-I achieves the lowest RMSE and NRMSE in the flu outbreak dataset, about 10% less than the best competing performer, LASSO. However, SADL-II generally achieved the best overall performance for all the datasets, with the best performance for traffic and water datasets as well as second lowest RMSE values for flu dataset. Since no contiguity matrix was available, we do not have results for SADL-I

Table 4. The training runtime of different methods on different datasets.

Methods	Flu outbreak	Traffic volume	Water quality
MLR	8.2577	**0.3781**	**0.0068**
LASSO	19.0304	18.6455	0.3454
SAR	35.2545	8.8084	N/A
SAR L1	42.0920	15.0020	0.9460
GL	11.1521	0.7683	N/A
SADL-I	10.9684	5.3734	N/A
SADL-II	**4.7294**	4.4816	1.2285

as well as SAR and GL for the water quality data. Based on NRMSE and NMAE, all the methods tend to have better performance on traffic and water datasets than that of flu datasets because the latter one is based on social media data which is very noisy and thus is a more challenging forecasting task.

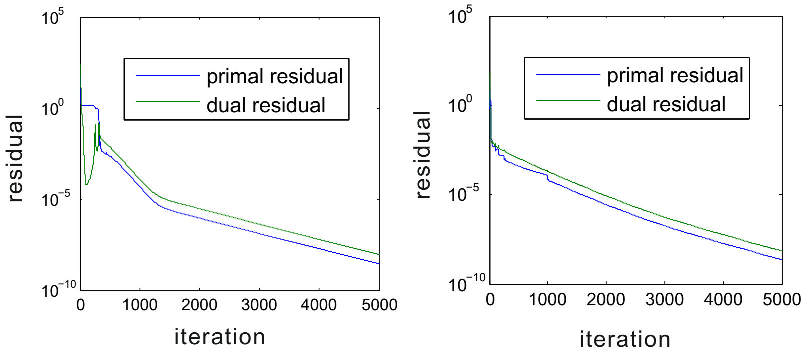
In particular, for the flu dataset, SADL-I and SADL-II achieved the best performance in RMSE and MAE, respectively, indicating that SADL-I's prediction errors were more consistent while there might be some larger errors in the predictions of SADL-II. Comparing the performance on flu and traffic datasets, it can be seen that the proposed SADL-I and SADL-II outperformed other methods by roughly 10% and 1%, respectively and thus the advantage looks much larger on flu datasets. This is because it is relatively more difficult to define a reasonable spatial dependency for flu outbreaks only based on prior knowledge, compared to the traffic dataset where the road network could be used to provide a reasonably good spatial dependency among traffic volume. This made the comparison methods also perform reasonably well on traffic data. For the flu dataset, MLR performed poorly, much worse than LASSO, while it outperformed LASSO for the other two datasets. This is because only the flu dataset, which has over 500 features, has a serious feature sparsity issue. This is also the reason for LASSO, SADL-I, and SADL-II, which enforce feature sparsity, to performed well for the flu data. Since the water quality data do not have prior location connectivity knowledge, the methods requiring this information, AR, GL, and SADL-I could not be used for it.

Furthermore, for the water quality data, which cannot provide prior knowledge on the spatial dependency, methods such as SAR, GL, and SADL-I cannot be directly utilized. The proposed SADL-II clearly outperformed the remaining methods by around 5% in all the metrics. This is because SADL-II sufficiently utilized the spatial topological constraints to infer the spatial dependency, which further enforces the dependency among the predictions in different locations, and thus boosted up the model generalizability. More importantly, the proposed SADL-II is able to infer the spatial dependency automatically, which is very valuable in analyzing the correlations among different locations toward future event occurrence. This will be further discussed in Section 6.2.3.

Figure 5 shows the temporal predictions of SADL-II for eight locations of each dataset. The other locations also follow the same patterns. Specifically, Figures 5(a)-5(h) demonstrate that the water quality data in terms of the pH value for the eight sites (e.g., Sites # 2198840 and # 2336152) was accurately predicted. This case shows a situation without strong periodicity and with large and random fluctuation of the respective water indices due to various reasons including precipitation, seasonal change, and natural disasters. SADL-II can still achieve high-quality results, and precisely predict all the large fluctuations of indices very close to the actual values.

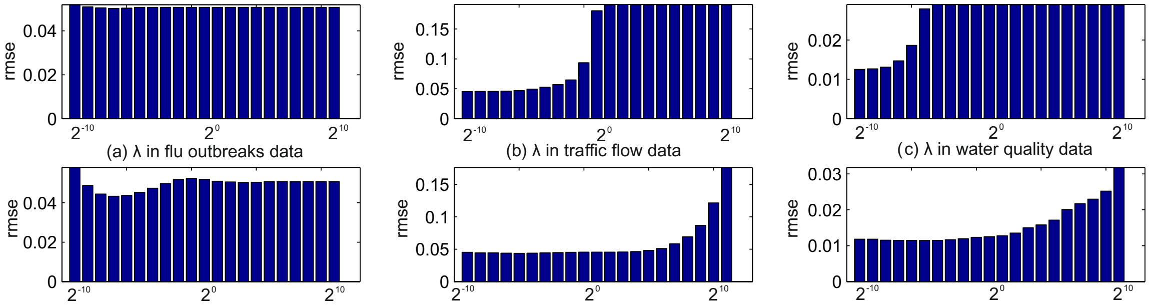
As shown in Figures 5(i)-5(p), the *traffic data* exhibits a more obvious periodicity, which is successfully predicted by SADL-II. These commute patterns capture regular traffic between downtown and the suburbs in the Washington D.C. area. Moreover, all the peaks and valleys were accurately predicted for all eight locations and for the entire two-month prediction periods. Finally, for the flu data, the predictions and the actual values for eight states such as New York and Utah are shown. First, we see a predicted seasonal periodicity of flu outbreaks, which follows the actual values well. Second, SADL-II can effectively detect the peaks and valleys of the predictions as shown in Figures 5(q)-5(x).

Table 4 shows the runtime of all the methods on all the datasets during training phase, each of which is the average runtime over 100 runs for each method on each dataset. As can be seen from the table, the runtime of SADL-I and SADL-II was relatively fast on larger dataset: SADL-II and SADL-I ranked first and third on flu outbreak dataset which is the largest dataset among all the three. For traffic volume dataset the runtime of them is still competitive. On the water quality dataset which is the smallest one, the proposed SADL-II requires 1.2 second which is the largest because it needs to optimize the spatial connectivity among all the locations. But having a training runtime of around only one second still ensures that it is highly efficient and practical in real-world applications. After model training, all the models run instantly to get the predictions in test phase.



(a) SADL-I (b) SADL-II

Fig. 6. The convergence process of SADL-I and SADL-II



(a) λ in flu outbreaks data (b) λ in traffic flow data (c) λ in water quality data

(d) σ in flu outbreaks data (e) σ in traffic flow data (f) σ in water quality data

Fig. 7. Sensitivity Analysis. The proposed models are not sensitive in σ^2. And λ is influenced by feature sparsity.

6.2.2 *Evaluation of the Method Properties.*

In this section, both the convergence and the parameter sensitivity are analyzed. Figure 6 shows the convergence of both SADL-I and SADL-II for the traffic dataset, which is the largest and second largest dataset in terms of data points and number of features, respectively. The convergence for the other datasets follows a similar pattern and was not included due to space limitations. As shown in Figure 6, both SADL-I and SADL-II converge sublinearly and continuously down to the residual errors of around 10^{-8} after 5000 iterations.

This verifies the claim on the methods' convergence in Section 5. A residual error of 10^{-8} is more than sufficient for practical applications. Typically, a value below 0.005 would be sufficient for the water quality prediction, which can be achieved with less than 200 iterations. Overall, each iteration is computed efficiently due to the effective utilization of closed form solutions. Figure 7 illustrates the influence of the tunable parameters of SADL-I and SADL-II on the performance. Figures 7(a), (b), and (c) show the performance changes when λ changes and σ^2 was set to 1. Figures 7(d), (e), and (f) show the performance variations when σ^2 changes and λ was set to 1. It is easy to see that the patterns of the flu dataset differ from those of the other two datasets. This is due to the feature sparsity in our flu data. Also, for the flu data, the best choice of λ, which controls the strength of the regularization on the feature sparsity, is around 2^{-7}. Which is much larger than the choice of λ for the other two datasets. This again shows that the models tend to enforce more feature sparsity to match the actual situation in the dataset. For the traffic and water quality datasets, the model performed be for small λ values. This implies that for datasets without feature sparsity problems (e.g., with few dense features), a relatively small value of λ is preferred. Finally, the performance is not very sensitive to σ^2 over such a large range of candidate values. This applies to all datasets. This shows the stability of the performance in terms of different values of σ^2.

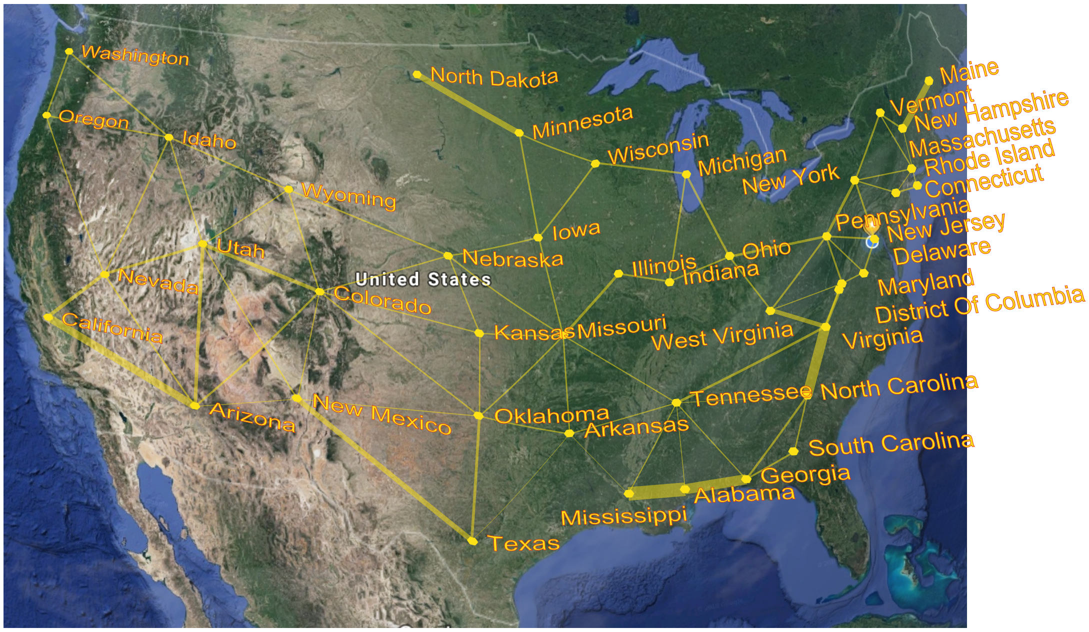
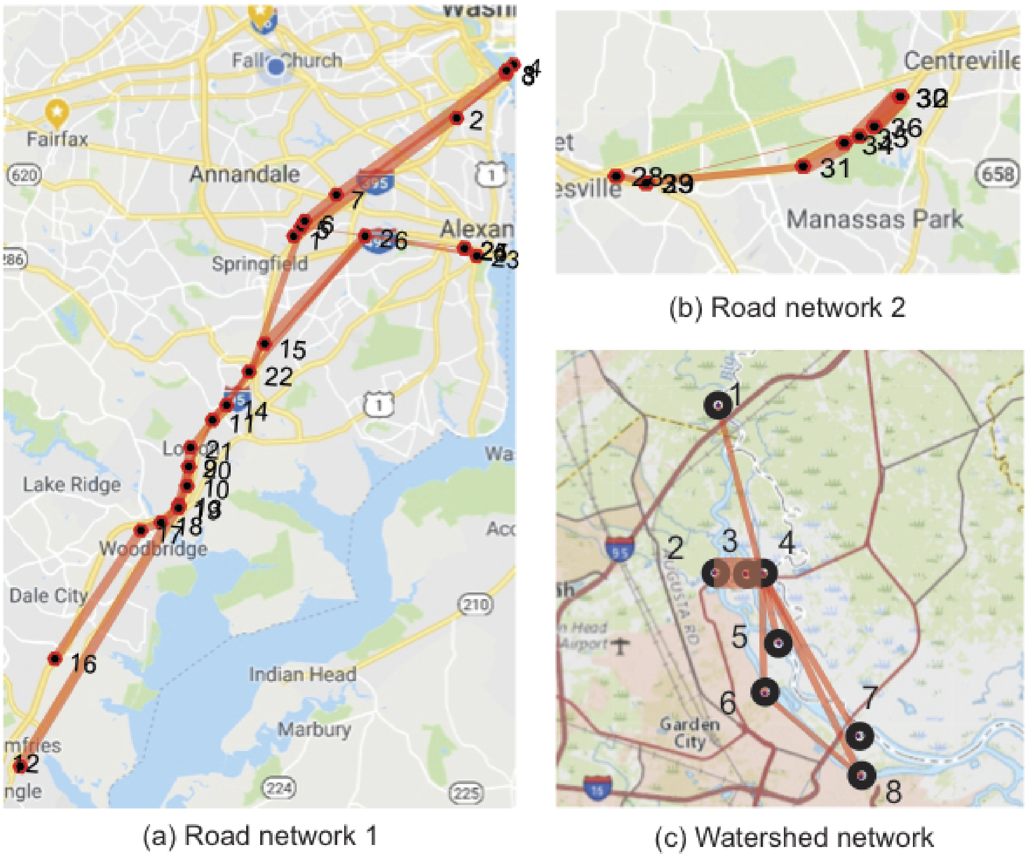


Fig. 8. The discovered spatial dependency towards influenza outbreaks for different states on flu outbreak dataset. (Better see in color)

6.2.3 Spatial Dependency Analysis. This section illustrates the spatial dependency discovered for all the three datasets based on the proposed SADL-I and SADL-II methods.

In contrast to existing work, the proposed methods SADL-I and SADL-II are able to discover the corresponding spatial dependency for the specific prediction tasks by achieving an optimal trade-off between exogenous and endogenous knowledge. Figure 8 and Figures 9(a), (b), (c), all show spatial networks in which a node represents a sensor location,



(a) Road network 1

(b) Road network 2

(c) Watershed network

Fig. 9. The discovered spatial dependency for traffic flow prediction and water quality prediction. (a) and (b) are for traffic dataset and (c) is for water quality dataset.

while an edge captures the spatial dependency. The thicker and edge, the higher its weight, which indicates a stronger dependency between the two locations. The weight reflects how much the measurements of a location contribute to a prediction and how much it is affected by its neighbors. Figure 8 illustrates the spatial dependency of the influenza outbreaks. In contrast to the conventional contiguity matrix, which assumes all the spatial locations have identical dependency strength with each other, the proposed methods found that the states of Mississippi, Alabama, and Georgia have relatively strong spatial dependency in influenza outbreaks. This corresponds to the fact that these states are the closest states within Mississippi region and share large, accessible boundaries. Similarly, the state pairs of Virginia and North Carolina, Maine and New Hampshire have strong connections. In contrast, the state of Iowa has a weak connection to Illinois, which is due to the fact that their common boundary is small.

Figures 9(a) and 9(b) shows the spatial dependency learned by SADL-I for the road network. Figure 9(a) shows I-95 and connecting highways, while Figure 9(b) shows I-66. Both roads are outside of Washington, D.C. It is easy to see that the dependency for nearby sensors along the same highway tends to be similar and relatively strong, while the dependency of sensors across different highways is weak. Consider here the example of the existing, but weak connection between Sensors 1 and 26 in Figure 9(a). Moreover, on the same highway, a sensor's spatial dependency to another distant sensor is much smaller than to a neighboring sensor. For example, the dependency between Sensors 29 and 34 is much smaller than the dependency between neighboring sensors. Finally, Figure 9(c) shows the spatial dependency automatically learned by SADL-II for the unknown watershed network. By verifying the learned spatial dependency using the real-world watershed map as shown in Figure 9(c), the plausibility of the learned dependency is shown. Specifically, the model successfully learned a connected graph of all the nodes in Figure 9(c), which matches an actual watershed network in which all the sites are connected. Second, the identified spatial dependency tends to be stronger for nearby sites, such as 2, 3, 4, but not for distant ones. This is much more reasonable to contiguity correlation used in conventional methods which, for instance, will assume Sites 1 and 4 have the same dependency strength as that between Sites 3 and 4. The inferred spatial dependency is also more reasonable than those based on inverse distance. For example, SADL-II inferred that there is no direct dependency between Sites 5 and 6, which are on the two side of an eyot, while the prior knowledge based only on inverse distance weighting [31] assumes a strong dependency between those in near proximity.

7 CONCLUSIONS

Being widely utilized, spatial regression models typically rely on spatial dependency, which is either manually defined or heuristically estimated. Without tuning these models to the context of specific applications, prediction results would be sub-optimal. In addressing this drawback, this paper develops a novel framework that jointly learns the predictive mapping and the spatial dependency. If the connectivity between locations is known, the SADL-I model learns the strength of the connectivity between locations. If the connectivity is incomplete or unknown, the SADL-II model can automatically learn the spatial connectivity and dependency based on spatial topological constraints. The proposed models are convex and iteratively optimized by our new ADMM-based algorithms, converging to a global optimal solution. Extensive experimentation on three real-world datasets demonstrates that the proposed models significantly outperform existing methods. Moreover, the spatial dependency results show the effectiveness of the proposed methods to automatically discover interpretable correlation patterns between different spatial locations.

Appendices

Appendix A: Proof of Lemma 1

Proof. Recall that the noise term $\varepsilon = (I - D) \cdot Y_t - X_\tau \cdot W^\mathsf{T}$ follows the Gaussian distribution $\varepsilon \sim \mathcal{N}(0, \sigma)$. Hence the likelihood in terms of the variable ε is:

$$\ell(\sigma|\varepsilon) = (1/(2\pi\sigma^2))^{|S||T|/2} \exp(-(\varepsilon^\mathsf{T}\varepsilon/(2\sigma^2))) \tag{23}$$

Recall the correlation between ε and Y:

$$\varepsilon = (I - D) \cdot Y_t - X_\tau \cdot W^\mathsf{T} \tag{24}$$

and leveraging the Jocabian factor theory [5], the likelihood in terms of the variable Y is:

$$\ell(D, W, \sigma^2|Y) = (1/(2\pi\sigma^2))^{|S||T|/2} \cdot J \cdot \exp(-(\varepsilon^\mathsf{T}\varepsilon/(2\sigma^2))) \tag{25}$$

$$\text{where the Jacobian factor is: } J = |\frac{\partial \varepsilon}{\partial Y}| = |I - D| \tag{26}$$

.

Combining Equations (24), (25), and (26) and perform logarithm transformation, the log-likelihood $\ell\ell(D, W, \sigma^2|Y)$ is calculated as below:

$$\ell\ell(D, W, \sigma^2|Y) = \sum\nolimits_t^{\mathsf{T}} \frac{1}{2\sigma^2} \|(I - D)Y_t - X_\tau W^{\mathsf{T}}\|_2^2) \\ + \frac{1}{2} \cdot |S||T| \ln(2\pi\sigma^2) - |T| \ln(\det(I - D)) \tag{27}$$

And hence the lemma is proved. □

REFERENCES

[1] Achim Ahrens and Arnab Bhattacharjee. Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128–155, 2015.

[2] Onureena Banerjee, Laurent El Ghaoui, and Alexandre dâĂŹAspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[3] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

[4] Arnab Bhattacharjee and Chris Jensen-Butler. Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617–634, 2013.

[5] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[7] Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.

[8] Mete Celik, Baris M Kazar, Shashi Shekhar, and Daniel Boley. Parameter estimation for the spatial autoregression model: a rigorous approach. In *Second NASA Data Mining Workshop: Issues and Applications in Earth Science with the 38th Symposium on the Interface of Computing Science, Statistics and Applications*, 2006.

[9] Turgay Celik. Spatial entropy-based global and local image contrast enhancement. *IEEE Transactions on Image Processing*, 23(12):5298–5308, 2014.

[10] Feng Chen, Baojian Zhou, Adil Alim, and Liang Zhao. A generic framework for interesting subspace cluster detection in multi-aributed networks. In *in international conference on Data Mining*, pages 41–50. IEEE, 2017.

[11] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

[12] Samuel I Daitch, Jonathan A Kelner, and Daniel A Spielman. Fitting a graph to vector data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 201–208. ACM, 2009.

[13] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[14] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.

[15] Steven Farber, Antonio Páez, and Erik Volz. Topology and dependency tests in spatial and network autoregressive models. *Geographical Analysis*, 41(2):158–180, 2009.

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[17] Tian Gao, Ziheng Wang, and Qiang Ji. Structured feature selection. In *ICCV*, pages 4256–4264, 2015.

[18] Virgilio Gomez-Rubio, Roger S Bivand, and Håvard Rue. Estimating spatial econometrics models with integrated nested laplace approximation. *arXiv preprint arXiv:1703.01273*, 2017.

[19] BD Greenshields, Ws Channing, Hh Miller, et al. A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board, 1935.

[20] David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 205–213. ACM, 2017.

[21] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In *International Conference on Machine Learning*, pages 928–937, 2015.

[22] Harry H Kelejian and Ingmar R Prucha. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of econometrics*, 157(1):53–67, 2010.

[23] Brenden Lake and Joshua Tenenbaum. Discovering structure by learning sparse graphs. *CogSci 2010*, 2010.

[24] Clifford Lam and Pedro CL Souza. Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, 35(8-10):1347–1376, 2016.

[25] Clifford Lam, Pedro CL Souza, et al. *Regularization for spatial panel time series using the adaptive lasso*. LSE, STICERD, 2014.

[26] Sangsoo Lee and Daniel Fambro. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Transportation Research Record: Journal of the Transportation Research Board*, (1678):179–188, 1999.

[27] Hongfei Li, Catherine A Calder, and Noel Cressie. One-step estimation of spatial dependence parameters: Properties and extensions of the aple statistic. *Journal of Multivariate Analysis*, 105(1):68–84, 2012.

[28] Po-Ling Loh and Martin J Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In *Advances in Neural Information Processing Systems*, pages 2087–2095, 2012.

[29] Jesús Mur and Ana Angulo. The spatial durbin model and the common factor tests. *Spatial Economic Analysis*, 1(2):207–226, 2006.

[30] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.

[31] E. Oktavia, Widyawan, and I. W. Mustika. Inverse distance weighting and kriging spatial interpolation for data center thermal monitoring. In *1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 69–74, Aug 2016.

[32] Iwao Okutani and Yorgos J Stephanedes. Dynamic prediction of traffic volume through kalman filtering theory. *Transportation Research Part B: Methodological*, 18(1):1–11, 1984.

[33] Philipp Otto and Rick Steinert. Estimation of the spatial weighting matrix for spatiotemporal data under the presence of structural breaks. *arXiv preprint arXiv:1810.06940*, 2018.

[34] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[35] Michael J Paul and Mark Dredze. A model for mining public health topics from twitter. *Health*, 11:16–6, 2012.

[36] Xi Qu and Lung-fei Lee. Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*, 184(2):209–232, 2015.

[37] Somwrita Sarkar and Sanjay Chawla. Inferring the contiguity matrix for spatial autoregressive analysis with applications to house price prediction. *arXiv preprint arXiv:1607.01999*, 2016.

[38] Shashi Shekhar, Zhe Jiang, Reem Y Ali, Emre Eftelioglu, Xun Tang, Venkata Gunturi, and Xun Zhou. Spatiotemporal data mining: a computational perspective. *ISPRS International Journal of Geo-Information*, 4(4):2306–2338, 2015.

[39] Wei Shi and Lung-fei Lee. Spatial dynamic panel data models with interactive fixed effects. *Journal of Econometrics*, 197(2):323–347, 2017.

[40] Brian L Smith, Billy M Williams, and R Keith Oswald. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4):303 – 321, 2002.

[41] David S Stoffer. Estimation and identification of space-time armax models in the presence of missing data. *Journal of the American Statistical Association*, 81(395):762–772, 1986.

[42] Martin Sundin, Arun Venkitaraman, Magnus Jansson, and Saikat Chatterjee. A connectedness constraint for learning sparse graphs. In *Signal Processing Conference (EUSIPCO)*, pages 151–155. IEEE, 2017.

[43] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[44] Sen Yang, Qian Sun, Shuiwang Ji, Peter Wonka, Ian Davidson, and Jieping Ye. Structural graphical lasso for learning mouse brain connectivity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1385–1394. ACM, 2015.

[45] Yao Zhang, Yun Xiong, Xinyue Liu, Xiangnan Kong, and Yangyong Zhu. Meta-path graphical lasso for learning heterogeneous connectivities. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 642–650. SIAM, 2017.

[46] Ali Ziat, Edouard Delasalles, Ludovic Denoyer, and Patrick Gallinari. Spatio-temporal neural networks for space-time series forecasting and relations discovery. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 705–714. IEEE, 2017.