entropy

MDPI

Article

Graph Regression Model for Spatial and Temporal Environmental Data—Case of Carbon Dioxide Emissions in the United States

Roméo Tayewo [1,*,†], François Septier [1,†], Ido Nevat [2,†] and Gareth W. Peters [3,†]

[1] Univ Bretagne Sud, CNRS UMR 6205, LMBA, F-56000 Vannes, France; francois.septier@univ-ubs.fr
[2] TUMCREATE, 1 Create Way, #10-02 CREATE Tower, Singapore 138602, Singapore; ido.nevat@tum-create.edu.sg
[3] Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA 93106, USA; garethpeters@ucsb.edu
* Correspondence: romeo.tayewo@univ-ubs.fr
† These authors contributed equally to this work.

Abstract: We develop a new model for spatio-temporal data. More specifically, a graph penalty function is incorporated in the cost function in order to estimate the unknown parameters of a spatio-temporal mixed-effect model based on a generalized linear model. This model allows for more flexible and general regression relationships than classical linear ones through the use of generalized linear models (GLMs) and also captures the inherent structural dependencies or relationships of the data through this regularization based on the graph Laplacian. We use a publicly available dataset from the National Centers for Environmental Information (NCEI) in the United States of America and perform statistical inferences of future CO_2 emissions in 59 counties. We empirically show how the proposed method outperforms widely used methods, such as the ordinary least squares (OLS) and ridge regression for this challenging problem.

Keywords: graph regression model; spatio-temporal data; CO_2 emission

check for updates

Citation: Tayewo, R.; Septier, F.; Nevat, I.; Peters, G.W. Graph Regression Model for Spatial and Temporal Environmental Data—Case of Carbon Dioxide Emissions in the United States. *Entropy* **2022**, *25*, 1272. https://doi.org/10.3390/e25091272

Academic Editor: Donald J. Jacobs

Received: 10 July 2023
Revised: 18 August 2023
Accepted: 24 August 2023
Published: 29 August 2023

1. Introduction

Statistical models for spatio-temporal data are invaluable tools in environmental applications, providing insights, predictions, and actionable information for understanding and managing complex environmental phenomena [1]. Such models help uncover complex patterns and trends, providing insights into how environmental variables change geographically and temporally. Many environmental datasets are collected at specific locations and times, leaving gaps in information. Statistical models help interpolate and map values between observation points, providing a complete spatial and temporal picture of the phenomenon being studied. Moreover, environmental applications frequently require predicting future values or conditions. Statistical models allow for accurate predictions by capturing the spatial and temporal dependencies present in the data. Such predictions provided by these models provide valuable information for decision makers by quantifying the effects of various factors on the environment and projecting the consequences of different actions.

Let $\{y_{t,s} : s \in \Omega_s, t \in \Omega_t\}$ denote the spatio-temporal random process for a phenomenon of interest evolving through space and time. As an example, $y_{t,s}$ might be the CO_2 emission level at a geographical coordinate $s = (\text{latitude, longitude})$ on the sphere at a given time t. Traditionally, one considers models for such a process from a *descriptive* context, primarily in terms of the first few moments of a probability distribution (i.e., mean and covariance functions in the case of a Gaussian process). Descriptive models are generally based on the *spatio-temporal mixed-effect model* [1,2], in which the spatio-temporal

process is described with a deterministic mean function and some random effects capturing the the spatio-temporal variability and interaction:

$$y_{t,s} = \mu_{t,s} + \epsilon_{t,s} \tag{1}$$

where $\mu_{t,s}$ is a deterministic (spatio-temporal) mean function or trend, and $\epsilon_{t,s}$ a zero-mean random effect, which generally depends on some finite number of unknown parameters. A common choice for the trend is to consider the following linear form $\mu_{t,s} = \boldsymbol{\phi}_{t,s}\boldsymbol{\beta}$, where $\boldsymbol{\phi}_{t,s}$ represents a vector of known covariates and $\boldsymbol{\beta}$ a set of unknown coefficients. Generally, with such a model, it is generally assumed that the process of interest is Gaussian. However, in real-world scenarios, data can exhibit heavy tails or outliers, which can significantly affect the distribution's shape and parameters. If these extreme values are not accounted for, it can lead to biased estimates and incorrect inferences. As a consequence, a more advanced model based on a generalized linear model (GLM) has been proposed [3]. The systematic component of the GLM specifies a relationship between the mean response and the covariates through a possibly nonlinear but known link function. Note that some additional random effects can be added in the transformed mean function, leading to the so-called *generalized linear mixed model* (GLMM) [4].

The main challenge of such models lies in estimating the unknown parameters. Once this important step is done, the different tasks of interest (prediction, decision, etc.) can be performed. Unfortunately, the inference of these parameters can lead to overfitting, multicollinearity-related instability, and lack of variable selection, resulting in complex models with high variance. As a consequence, regularization methods using the ℓ_1 and/or ℓ_2 norm as penalty function are generally used in practice to mitigate these issues by controlling the model complexity, improving generalization, and enhancing the stability of coefficient estimates [5,6].

Contributions

Graph signal processing is a rapidly developing field that lies at the intersection between signal processing, machine learning and graph theory. In recent years, graph-based approaches to machine learning problems have proven effective at exploiting the intrinsic structure of complex datasets [7]. Recently, graph penalties were applied successfully to the reconstruction of a time-varying graph signal [8,9] or to the regression with a simple linear model [10,11]. In these works, the results highlight that regularization based on the graph structure could have an advantage over more traditional norm-based ones in situations where the data or variables have inherent structural dependencies or relationships. The main advantage of graph penalties is that they take into account the underlying graph structure of the variables, capturing dependencies and correlations that might not be adequately addressed by norm-based penalties.

In this work, we propose a novel and general spatio-temporal model that incorporates a graph penalty function in order to estimate the unknown parameters of a spatio-temporal mixed-effect model based on a generalized linear model. In addition, different structures of graph dependencies are discussed. Finally, the proposed model is applied to a real and important environmental problem: the prediction of CO_2 emissions in the the United States. As recently discussed in [12], regression analysis is one of the most widely used statistical method to characterize the influence of selected independent variables on a dependent variable and thus has been widely used to forecast CO_2 emissions. To the best of our knowledge, this is the first time that a more advanced model, i.e., a GLM-based spatio-temporal mixed effect model with graph penalties, is proposed to predict CO_2 emissions.

2. Problem Statement—The Classical Approach

In this section, we first provide a background of graphs and their properties, then we introduce the system model of our problem followed by the classical approach which uses a linear regression structure.

2.1. Preliminaries

Let us consider a weighted, undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$ composed of $|\mathcal{V}| = N$ vertices. $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix, where $A_{ij} \geq 0$ represents the strength of the interaction between nodes i and j. An example of such a graph is depicted in Figure 1. $\mathcal{E}$ is the set of edges, and therefore $(i, j) \in \mathcal{E}$ implies $A_{ij} > 0$ and $(i, j) \notin \mathcal{E}$ implies $A_{ij} = 0$. The graph can be defined through the (unnormalized) Laplacian matrix $\boldsymbol{L} \in \mathbb{R}^{N \times N}$:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \tag{2}$$

where $\boldsymbol{D}$ corresponds to the degree matrix of the graph as $\boldsymbol{D} = \text{diag}(D_{11}, D_{22}, \dots D_{NN})$, where D_{ii} is the i-th column sum (or row sum) of the adjacency matrix A.

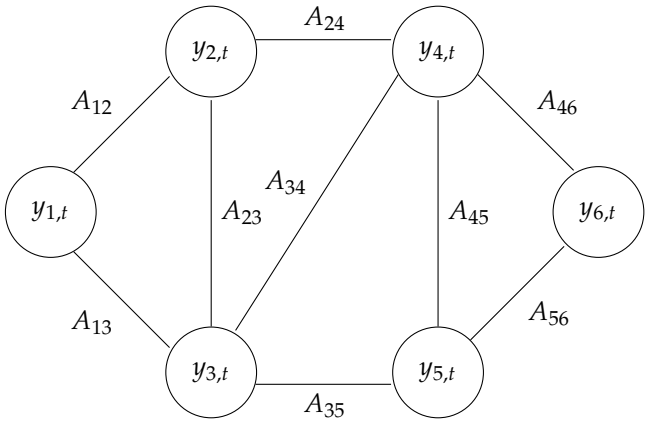


Figure 1. Example of a graph with $|\mathcal{V}| = 6$ vertices at time t.

The graph Laplacian, closely related to the continuous domain Laplace operator, has many interesting properties. One of them is the ability to inform about the connectedness of the graph. By combining this property with any graph signal at time t, $\boldsymbol{y}_t \in \mathbb{R}^N$, in the following quadratic sum,

$$\boldsymbol{y}_t^\top \boldsymbol{L} \boldsymbol{y}_T = \sum_{(i,j)} A_{ij} \left(y_{t,i} - y_{t,j}\right)^2 \tag{3}$$

can be considered a measure of the cross-sectional similarity of the signal, with smaller values indicating a smoother signal reaching a minimum of zero for a function that is constant on all connected sub-components [13].

2.2. System Model

The main objective of this paper is to design a statistical regression model in order to characterize and predict CO_2 emissions across time and space. More precisely, the paper is concerned with the situation where a signal $\boldsymbol{y}_t = (y_{t,1}, y_{t,2}, \dots, y_{t,N})^\top \in \mathbb{R}^N$ is measured on the vertices of a fixed graph at a set of discrete times $t \in [1, 2, \dots, T]$. This vector corresponds to the CO_2 emission measured at N different spatial locations at time t. At each of these time instants, a vector of K covariates $\boldsymbol{x}_t \in \mathbb{R}^K$ is also measured, which is not necessarily linked to any node or set of nodes.

Objectives:

1. Determine, for each of the N different locations, the specific relationship between the response variables $\{y_{t,i}\}_{t=1}^{T}$ and the set of covariates $\{\boldsymbol{x}_t\}_{t=1}^{T}$.
2. Based on this relationship, make a prediction of the CO_2 levels in different locations in space and time.

2.3. Problem Formulation with a Classical Linear Regression Model

The most common form of structural assumption is that the responses are assumed to be related to predictors through some deterministic function f and some additive random error component ϵ_i so that for the i-th location and $\forall t = 1, \ldots, T$ we have that

$$y_{t,i} = f_i(\boldsymbol{x}_t) + \epsilon_i, \tag{4}$$

where ϵ_i is a zero-mean error random variable. Therefore, a classical procedure consists of approximating the true function f_i by a linear combination of basis functions:

$$f_i(\boldsymbol{x}_t) \approx \sum_{p=1}^{P} \beta_{i,p}\phi_{i,p}(\boldsymbol{x}_t) = \boldsymbol{\phi}_i(\boldsymbol{x}_t)^T\boldsymbol{\beta}_i, \tag{5}$$

where $\boldsymbol{\beta}_i = \left[\beta_{i,1} \ldots \beta_{i,P}\right]^T$ is the set of coefficients corresponding to basis functions $\boldsymbol{\phi}_i(\boldsymbol{x}_t) = \left[\phi_{i,1}(\boldsymbol{x}_t) \ldots \phi_{i,P}(\boldsymbol{x}_t)\right]^T$ in order to approximate the function $f_i(\cdot)$ associated to the signal over time at i-th location, i.e., $\{y_{t,i}\}_{t=1}^{T}$.

The linear regression model over all the N different locations could be formulated in a matrix form as follows $\forall t \in [1, 2, \ldots, T]$:

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \tag{6}$$

where

$$\boldsymbol{\Phi}_t = \begin{bmatrix} \boldsymbol{\phi}_1(\boldsymbol{x}_t) & \mathbf{0}_{1\times P} & \ldots & \mathbf{0}_{1\times P} \\ \mathbf{0}_{1\times P} & \boldsymbol{\phi}_2(\boldsymbol{x}_t) & & \vdots \\ \vdots & & \ddots & \mathbf{0}_{1\times P} \\ \mathbf{0}_{1\times P} & \ldots & \mathbf{0}_{1\times P} & \boldsymbol{\phi}_N(\boldsymbol{x}_t) \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix} \text{ and } \boldsymbol{\epsilon}_t = \begin{bmatrix} \epsilon_{t,1} \\ \vdots \\ \epsilon_{t,N} \end{bmatrix} \tag{7}$$

As a consequence, this linear regression can be fully summarized as

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{y} = \left(\boldsymbol{y}_1^T\boldsymbol{y}_2^T \ldots \boldsymbol{y}_T^T\right)^\top \in \mathbb{R}^{NT\times 1}$ and

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 \\ \vdots \\ \boldsymbol{\Phi}_T \end{bmatrix} \in \mathbb{R}^{NT\times NP} \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_T \end{bmatrix} \in \mathbb{R}^{NT\times 1},$$

where $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}_{NT\times 1}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$.

In such a model, the most common approach to estimate the regression coefficients is the generalized least square (GLS) method, which aims at minimizing the squared Mahalanobis distance of the residual vector:

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}). \tag{9}$$

Theorem 1 (Aitken [14]). *Consider that the following conditions are satisfied:*

(A1) The matrix $\boldsymbol{\Phi}$ *is nonrandom and has full rank, i.e., its columns are linearly independent,*

(A2) The vector $\boldsymbol{y}$ *is a random vector such that the following hold:*

(i) $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\Phi}\boldsymbol{\beta}_0$ *for some* $\boldsymbol{\beta}_0$*;*

(ii) $Var(\boldsymbol{y}) = \boldsymbol{\Sigma}$ *is a known positive definite matrix.*

Then, the generalized least square estimator from (9) *is given by*

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

Moreover, $\hat{\boldsymbol{\beta}}_{GLS}$ *corresponds to the best linear unbiased estimator for* $\boldsymbol{\beta}_0$ *and its covariance matrix is* $Var\left(\hat{\boldsymbol{\beta}}_{GLS}\right) = \left(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\right)^{-1}$.

Let us remark that the ordinary least square (OLS) estimator is nothing but a special case of the GLS estimator. They are indeed equivalent for any diagonal covariance matrix $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$.

2.4. Generalized Linear Models

In this paper, we propose to use the generalized linear model (GLM) structure [15], which is a flexible generalization of linear regression model discussed previously. In this model, the additivity assumption of the random component is removed and more importantly, the response variables can be distributed from more general distributions in the standard linear model for which one generally assumes normally distributed responses, see discussions in [16,17]. The likelihood distribution of the response variables $f_Y(\boldsymbol{y}|\boldsymbol{\beta})$ is a member of the *exponential family*, which includes the normal, binomial, Poisson and gamma distributions, among others.

Moreover, in a GLM, a smooth and invertible function $g(\cdot)$, called *link function*, is introduced in order to transform the expectation of the response variable, $\mu_{t,i} \equiv \mathbb{E}[y_{t,i}]$

$$g(\mu_{t,i}) = \eta_{t,i} = \boldsymbol{\phi}_i(\boldsymbol{x}_t)^T\boldsymbol{\beta}_i. \tag{10}$$

Because the link function is invertible, we can also write

$$\mu_{t,i} = g^{-1}(\eta_{t,i}) = g^{-1}\left(\boldsymbol{\phi}_i(\boldsymbol{x}_t)^T\boldsymbol{\beta}_i\right), \tag{11}$$

and, thus, the GLM may be thought of as a linear model for a transformation of the expected response or as a nonlinear regression model for the response. In theory, the link function can be any monotonic and invertible function. The inverse link g^{-1} is also called the *mean function*. Commonly employed link functions and their inverses can be found in [15]. Note that the *identity link* simply returns its argument unaltered $\mu_{t,i} = g^{-1}(\eta_{t,i}) = \eta_{t,i} = \boldsymbol{\phi}_i(\boldsymbol{x}_t)^T\boldsymbol{\beta}_i$ and therefore is equivalent to the assumption (A2)-(i) of Theorem 1 used in the classical linear model.

In GLM, due to the nonlinearity induced by the link function, the regression coefficients are generally obtained with the maximum likelihood technique, which is equivalent to minimizing a cost function defined as the negative log-likelihood function $f_Y(\boldsymbol{y}|\boldsymbol{\beta})$ as [16]

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} V(\boldsymbol{y};\boldsymbol{\beta}), \tag{12}$$

with $V(\boldsymbol{y};\boldsymbol{\beta}) = -\ln f_Y(\boldsymbol{y}|\boldsymbol{\beta})$.

3. Proposed Graph Regression Model

In this section, we develop our *penalized regression model over graph*. We first show how to overcome some of the deficiencies in traditional regression models by introducing

new penalty terms which regulate the solution. Finally we provide details regarding the estimation procedure and the algorithm we develop.

3.1. Penalized Regression Model over Graph

In the previous section, we introduced a flexible generalization in order to model our spatial and temporal response variables of interest. Unfortunately, two main issues could arise. On the one hand, the solution of the optimization problem defined in (12) may not be unique if $\boldsymbol{\Phi}$ has full rank deficiency or when the number of regression coefficients exceeds the number of observations (i.e., $NP > NT$). On the other hand, the learned model could suffer from poor generalization due to, for example, the choice of an overcomplicated model. To avoid such problems, the most commonly used approach is to introduce a penalty function in the optimization problem to further constrain the resulting solution as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (V(\boldsymbol{y}; \boldsymbol{\beta}) + h(\boldsymbol{\beta}; \boldsymbol{\gamma})). \tag{13}$$

The penalty term $h(\boldsymbol{\beta}; \boldsymbol{\gamma})$ can be decomposed as the sum of p penalty functions and therefore depends on some positive tuning parameters $\{\gamma_i\}_{i=1}^{p}$ (regularization coefficients), which controls the importance of each elementary penalty function in the resulting solution. When every parameter is null, i.e., $\{\gamma_i\}_{i=1}^{p} = 0$, we obtain the classical GLM solution in (12). On the contrary, for large values of $\boldsymbol{\gamma}$, the influence of the penalty term on the coefficient estimate increases. The most commonly used penalty functions are the ℓ_2 norm (ridge), ℓ_1 norm (LASSO) or a combination of both (Elastic-net)—see [18] for details.

In this paper, we propose to use an elementary penalty function, which takes into account the specific graph structure of the observations. As in [10,11], a penalty function can be introduced in order to enforce some smoothness of the predicted mean of the signal $\mathbb{E}[\boldsymbol{y}_t]$ over the underlying graph at each time instant. More specifically, we propose to use the following estimator:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \left(V(\boldsymbol{y}; \boldsymbol{\beta}) + \gamma_1 \boldsymbol{\beta}^\top \boldsymbol{\beta} + \gamma_2 \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{y}_t]^\top \boldsymbol{L} \mathbb{E}[\boldsymbol{y}_t] \right) \\
&= \arg\min_{\boldsymbol{\beta}} \left(V(\boldsymbol{y}; \boldsymbol{\beta}) + \gamma_1 \boldsymbol{\beta}^\top \boldsymbol{\beta} + \gamma_2 \sum_{t=1}^{T} \boldsymbol{g}^{-1}(\boldsymbol{\phi}(\boldsymbol{x}_t)\boldsymbol{\beta})^\top \boldsymbol{L} \boldsymbol{g}^{-1}(\boldsymbol{\phi}(\boldsymbol{x}_t))\boldsymbol{\beta} \right) \\
&= \arg\min_{\boldsymbol{\beta}} \left(V(\boldsymbol{y}; \boldsymbol{\beta}) + \gamma_1 \boldsymbol{\beta}^\top \boldsymbol{\beta} + \gamma_2 \boldsymbol{g}^{-1}\left(\boldsymbol{\beta}^\top \boldsymbol{\Phi}^\top\right)(\boldsymbol{I}_T \otimes \boldsymbol{L}) \boldsymbol{g}^{-1}(\boldsymbol{\Phi}\boldsymbol{\beta}) \right),
\end{aligned} \tag{14}$$

where the function $\boldsymbol{g}^{-1}(\cdot) : \mathbb{R}^{NT} \mapsto \mathbb{R}^{NT}$ corresponds to the element-wise application of the inverse link function introduced in (11) on the input argument. $\boldsymbol{I}_T \otimes \boldsymbol{L}$ stands for the tensor product between the identity matrix of size T ($\boldsymbol{I}_T$) and the Laplacian matrix of the underlying graph ($\boldsymbol{L}$). The penalty function is therefore the sum of two elementary ones with $\gamma_1, \gamma_2 \geq 0$, their regularization coefficients. The regularization $\boldsymbol{\beta}^\top \boldsymbol{\beta} = \|\boldsymbol{\beta}\|^2$ imposes some smoothness conditions on possible solutions, which also remain bounded. Finally, the regularization based on the graph Laplacian L enforces the expectation of the response variable through the GLM model to be smooth over the considered graph $\mathcal{G}$ at each time t. It comes from the property of the Laplacian matrix discussed in Section 2.1.

As recently discussed in both [8,9], in some practical applications, the reconstruction of a time-varying graph signal can be significantly improved by adequately exploiting the correlations of the signal in both space and time. The authors show from several real-world datasets that the time difference signal (i.e., $\mathbb{E}[\boldsymbol{y}_t] - \mathbb{E}[\boldsymbol{y}_{t-1}]$ in our case) exhibits smoothness on the graph, even if signals $\mathbb{E}[\boldsymbol{y}_t]$ are not smooth on the graph. The proposed model can be simply rewritten as follows in order to take into account this property:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(V(\boldsymbol{y}; \boldsymbol{\beta}) + \gamma_1 \boldsymbol{\beta}^\top \boldsymbol{\beta} + \gamma_2 \boldsymbol{g}^{-1}\left(\boldsymbol{\beta}^\top \boldsymbol{\Phi}^\top\right) \widetilde{\boldsymbol{L}} \boldsymbol{g}^{-1}(\boldsymbol{\Phi}\boldsymbol{\beta}) \right), \tag{15}$$

With this general formulation, several cases can be considered:

- *Case 1*—$\widetilde{\boldsymbol{L}} = \boldsymbol{I}_T \otimes \boldsymbol{L}$: the penalization induces the smoothness of the successive mean vectors $\mathbb{E}[\boldsymbol{y}_1], \ldots, \mathbb{E}[\boldsymbol{y}_T]$ over a static graph structure $\boldsymbol{L}$.
- *Case 2*—$\widetilde{\boldsymbol{L}} = \text{diag}(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_T)$: the penalization induces the smoothness of the successive mean vectors $\mathbb{E}[\boldsymbol{y}_1], \ldots, \mathbb{E}[\boldsymbol{y}_T]$ over a time-varying graph structure, $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_T$.
- *Case 3*—$\widetilde{\boldsymbol{L}} = \boldsymbol{D}_h^\top (\boldsymbol{I}_{T-1} \otimes \boldsymbol{L}) \boldsymbol{D}_h$ or $\widetilde{\boldsymbol{L}} = \boldsymbol{D}_h^\top \text{diag}(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_{T-1}) \boldsymbol{D}_h$: The penalization induces the smoothness of the time difference mean vectors $\mathbb{E}[\boldsymbol{y}_2] - \mathbb{E}[\boldsymbol{y}_1], \ldots, \mathbb{E}[\boldsymbol{y}_T] - \mathbb{E}[\boldsymbol{y}_{T-1}]$ over a graph structure which could be either static or time varying, respectively. The matrix $\boldsymbol{D}_h^\top$ of dimension $NT \times N(T-1)$ defined as

$$\boldsymbol{D}_h^\top = \begin{bmatrix} -\boldsymbol{I}_N & \boldsymbol{0}_N & \ldots & \ldots & \ldots & \boldsymbol{0}_N \\ \boldsymbol{I}_N & -\boldsymbol{I}_N & \boldsymbol{0}_N & \ldots & \ldots & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{I}_N & -\boldsymbol{I}_N & \boldsymbol{0}_N & \ldots & \boldsymbol{0}_N \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{0}_N & \ldots & \boldsymbol{0}_N & \boldsymbol{I}_N & -\boldsymbol{I}_N & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \ldots & \ldots & \boldsymbol{0}_N & \boldsymbol{I}_N & -\boldsymbol{I}_N \\ \boldsymbol{0}_N & \ldots & \ldots & \ldots & \boldsymbol{0}_N & \boldsymbol{I}_N \end{bmatrix},$$

allows to transform the mean vector into the time difference mean vector.

Proposition 1. *When the response variables are considered to be normally distributed, i.e., $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, then the solution that minimizes the cost function defined in Equation (15) is given by*

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} + \gamma_1 \boldsymbol{I}_{NP} + \gamma_2 \boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} \tag{16}$$

Proof. See Appendix A. □

3.2. Learning and Prediction Procedure

As discussed in the previous section, our proposed estimator in (15) results from a regression model with a penalization function over the graph, which depends on some hyperparameters, i.e., $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2\}$. Cross-validation techniques are the most commonly used strategies for the calibration of such hyperparameters, as they allow us to obtain an estimator of the generalization error of a model [19]. In this paper, a cross-validation technique is used by partitioning the dataset into train, validation and test sets. Only the train and validation sets are used to obtain the selected parameters/hyperparameters set. Finally, the model with the selected set is evaluated using the test set.

Cross validation (CV) is a resampling method that uses different portions of the data to test and train a model through different iterations. Resampling may be useful while working with *iid* data. However, as opposed to the latter, time-series data usually posses temporal dependence, and therefore, one should respect the temporal structure while performing CV in that context. To that end, we follow the procedure of forward validation (we refer to it as time series CV) originally due to [20]. More specifically, the dataset is partitioned as follows $\mathcal{D}_{train} = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=1}^{\rho_{train} T}$, $\mathcal{D}_{val} = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=\rho_{train} T+1}^{(\rho_{train}+\rho_{val}) T}$ and $\mathcal{D}_{test} = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=(\rho_{train}+\rho_{val})T+1}^{T}$, where ρ_{train} and ρ_{val} correspond to the percentage of the dataset used for training and validation, respectively. In this paper, we set $\rho_{val} = \frac{1-\rho_{train}}{2}$ to have the same number of data in both the validation and test sets. The set of hyperparameters and parameters are obtained by minimizing the generalization error approximated using the validation set. In practice, the hyperparameters are optimized using either numerical optimization methods that do not require a gradient (e.g., Nelder–Mead optimizer) or a grid of discrete values. The proposed learning procedure used in this work is summarized in Algorithm 1.

Algorithm 1 Learning procedure of the proposed penalized regression model over graph

Input: $\mathcal{D}_{train} = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=1}^{\rho_{train}T}$,
$\mathcal{D}_{val} = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=\rho_{train}T+1}^{(\rho_{train}+\rho_{val})T}$
$\mathcal{D}_{test} = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=(\rho_{train}+\rho_{val})T+1}^{T}$

1: Iterations of a numerical optimization method

2: **while** $E^*_{\mathcal{D}_{val}} \neq E^{min}_{\mathcal{D}_{val}}$ **do**

3: Let $\boldsymbol{\gamma}^*$ denote the candidate for the values of hyperparameters for this iteration of the chosen derivative-free optimization technique.

4: Given $\boldsymbol{\gamma}^*$, obtain the optimal regression coefficient $\hat{\boldsymbol{\beta}}^*$ in (15) using only the data from the training set $\mathcal{D}_{train}$:

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}} \left(V(\boldsymbol{y} \in \mathcal{D}_{train}; \boldsymbol{\beta}) + \gamma_1^* \boldsymbol{\beta}^\top \boldsymbol{\beta} + \gamma_2^* \sum_{t \in \mathcal{D}_{train}} \boldsymbol{g}^{-1}(\boldsymbol{\phi}(\boldsymbol{x}_t)\boldsymbol{\beta})^\top \widetilde{\boldsymbol{L}} \boldsymbol{g}^{-1}(\boldsymbol{\phi}(\boldsymbol{x}_t))\boldsymbol{\beta} \right).$$

either by a numerical optimization technique or Equation (16) in case of Gaussian likelihood.

5: Compute the estimator of the generalization error using the validation set:

$$E^*_{\mathcal{D}_{val}} = \frac{1}{\rho_{val} T} \sum_{t \in \mathcal{D}_{val}} ||\boldsymbol{y}_t - g^{-1}(\boldsymbol{\phi}(\boldsymbol{x}_t))\hat{\boldsymbol{\beta}}^*||^2$$

6: **end while**

Output: Optimal hyperparameters $\hat{\boldsymbol{\gamma}}$ and regression coefficients $\hat{\boldsymbol{\beta}}$

4. Numerical Study—CO_2 Prediction in the United States

In this section, we empirically assess the benefit of using our proposed penalized regression model over graph for the prediction of CO_2 in the United States. For this purpose, the CO_2 emission levels were obtained from the Vulcan project (https://vulcan.rc.nau.edu/ (accessed on 1 August 2023)) [21] and more especially the dataset (https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1810 (accessed on 1 August 2023)), which provides emissions on a 1 km by 1 km regular grid with an hourly time resolution for the 2010–2015 time period. More specifically, the response variable vector $\boldsymbol{y}_t$ corresponds to the CO_2 emissions for the t-th day after 1 January 2011 at $N = 59$ different counties on the east coast of the United States of America (see Appendix B for the full list of selected counties).

On the other hand, among the explanatory variables presented in detail below, there are weather data from weather daily information available on the platform https://www.ncdc.noaa.gov/ghcnd-data-access (accessed on 1 August 2023) of National Centers for Environmental Information (NCEI) in the United States of America. NCEI manages one of the largest archives of atmospheric, coastal, geophysical, and oceanic research in the world.

4.1. Choice of Covariates and Data Pre-Processing

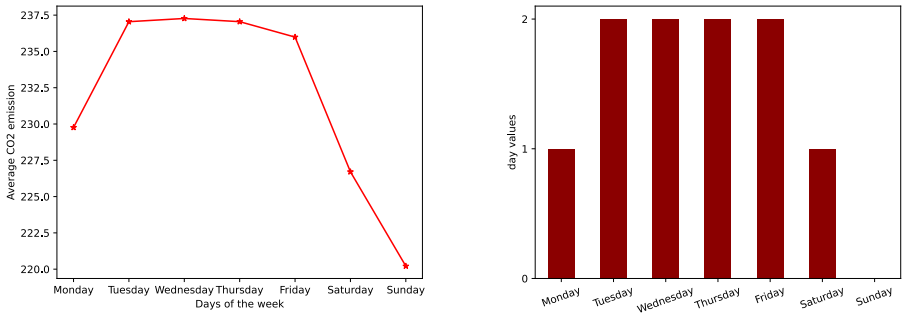
The covariates we propose to use to model the daily CO_2 emissions at the US counties level are composed of three types of data:

- Daily weather data (available on the platform of National Centers for Environmental Information (NCEI) https://www.ncdc.noaa.gov/ghcnd-data-access (accessed on 1 August 2023)) in the United States of America including maximal temperature (*TMAX*), minimal temperature (*TMIN*) and precipitation (*PREC*);
- Temporal information to capture the time patterns of the data;
- Lagged CO_2 emission variables to take into account the time correlation of the response.

All the variables related to the first two points are commonly used as covariates for each county, whereas lagged variables are county-specific.

Firstly, for the weather data, a number of steps are taken to pre-process them before feeding into the learning procedure described in Algorithm 1. Firstly any weather stations from the 59 US counties with a large proportion of missing values over the period of time are discarded. Missing values in the retained weather stations are interpolated linearly between the available readings. Then, the weather data are summarized at the state level—the 59 counties are part of 19 different states. As a consequence, for each state, the 3 weather variables (*TMAX*, *TMIN* and *PREC*) are averaged over the retained weather stations of that state. Whatever the county considered, weather variables from all 19 states are utilized as covariates in $\{\boldsymbol{\phi}_i\}_{i=1}^{N}$ of Equation (7). The final step before estimation is to transform all variables so that they are scaled and translated to achieve a unit marginal variance and zero mean.

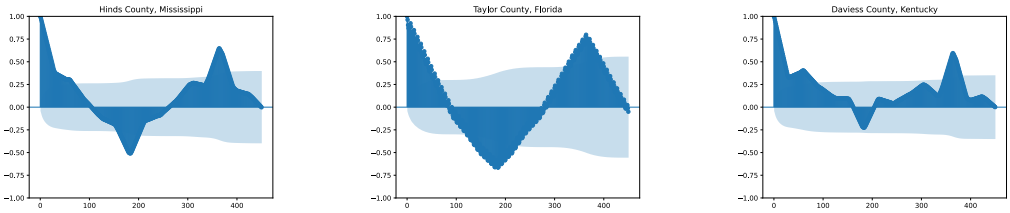
Secondly, for the temporal patterns in the data, we consider three types: a week identifier (*WD*), a weight associated to each day of a week (*WD*) and a trend variable (*TREND*). The variable *WI* simply corresponds to a one-hot encoding of the week number of the year. The variable *WD* is added after observing that a regular pattern can be observed concerning the evolution of the CO_2 emission with the day of the week—as shown in Figure 2, less emissions typically are observed during the weekend. The trend variable (*TREND*) is simply a linear and regularly increasing function at the daily rate from 0 (1 January 2010) to 1 (31 December 2015).



(**a**) (**b**)

Figure 2. Choice of the covariate *WD* to encapsulate information about the weekday for the CO_2 emission. (**a**) Spatial and temporal average of the CO_2 emission per weekday. (**b**) Values assigned to the covariates *WD* depending on the current weekday.

Finally, to take into account the time correlation of the CO_2 emissions, we decided to use some lagged response variables as covariates. More precisely, after analyzing the autocorrelation function (ACF) of the time series of CO_2 for each county (see Figure 3 for the ACF of three different counties), we proposed to use as covariates three lagged versions of the response variable. More precisely, for the i-th county at time t, $y_{t,i}$, the following lagged variables are used as predictors: the 365-day lagged variable $y_{t-365,i}$ (one year), the 182-day lagged variable $y_{t-182,i}$ (about six months) and the 14-day lagged variable $y_{t-14,i}$ (about 2 weeks).



(**a**) ACF for Hinds County (**b**) ACF for Taylor County (**c**) ACF for Daviess County

Figure 3. Illustration of the time correlation of the daily CO_2 emissions per county with the autocorrelation function (ACF) of three different counties.

4.2. Graph Construction of the Spatial Component

In this work, the 59 counties are considered the nodes of a common graph. The locations of the chosen counties are depicted in Figure 4. As a consequence, case 1 of the graph penalty function of Section 3.1 is considered, i.e., $\widetilde{\boldsymbol{L}} = \boldsymbol{I}_T \otimes \boldsymbol{L}$. The single Laplacian matrix $\boldsymbol{L}$ is defined through the adjacency matrix.

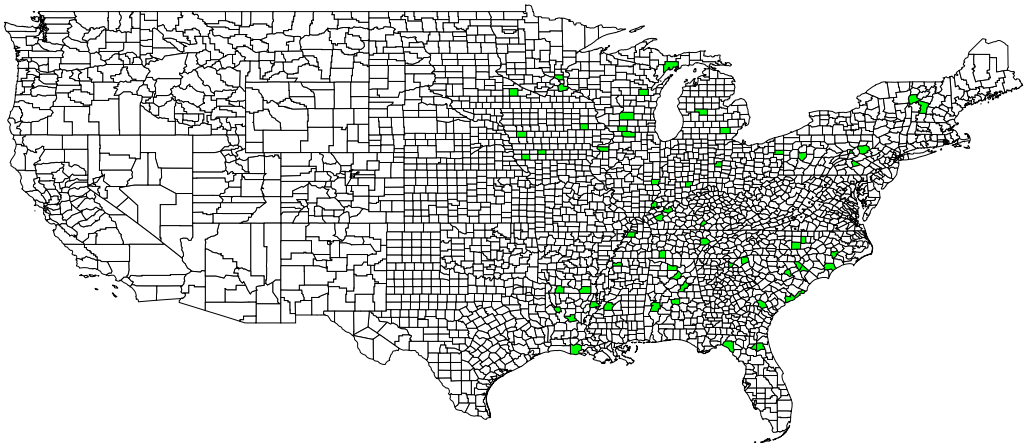


Figure 4. US counties selected as nodes of the graph depicted in green.

A graph adjacency matrix should reflect the tendency for measurements made at node pairs to have similar values in mean. There are many possible choices for the design of this adjacency matrix. In this work, two different choices of matrix are compared. As in [11], we firstly construct the adjacency matrix based on distances by setting

$$A_{i,j}^{dist} = e^{-l\frac{d_{i,j}^2}{\sum_{i,j} d_{i,j}^2}}, \tag{17}$$

where $d_{i,j}$ denotes the geodesic distance between the i-th and j-th counties in kilometers and l is a scaling hyperparameter to be optimized using Algorithm 1. A heat map of the geodesic distances in kilometers between counties is represented in Figure 5.

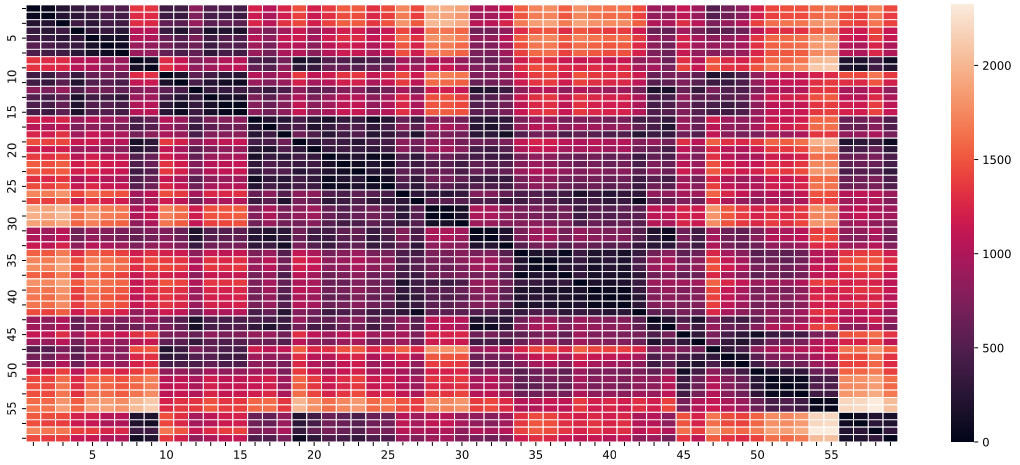


Figure 5. Geodesic distances in kilometers between counties.

The second proposition for the adjacency matrix is to utilize the empirical correlations between counties CO_2 emissions. For two counties i and j, the adjacency coefficient is defined as follows:

$$A_{i,j}^{corr} = e^{-l \max\left(0,\, \rho_{i,j}^2\right)} \tag{18}$$

where $\rho_{i,j}$ is the empirical correlation between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, the CO_2 emissions of the i-th and j-th counties, respectively.

4.3. Numerical Experiments

In the following numerical experiments, the proposed penalized regression model over graph is compared to two other classical models, namely, the ridge and the ordinary least square (OLS) solution. In fact, these two models are nothing but special cases of the proposed model by setting in Equation (16) either $\gamma_2 = 0$ or $(\gamma_1 = 0, \gamma_2 = 0)$, respectively.

Firstly, we empirically study the performance of the penalized regression model over graph with the two possible choices for the Laplacian matrix. As shown in Table 1, using the adjacency matrix based on geodesic distances rather than on empirical correlations improves the RMSE on both the validation and the test sets. A smaller RMSE on the training set using the correlation-based adjacency matrix shows that this choice could lead to overfitting.

Table 1. RMSE of the penalized regression model over graph with the Laplacian defined using an adjacency matrix based either on geodesic distances or on empirical correlations.

Root Mean Square Error (RMSE): Distances Versus Empirical Correlations						
	Testing Set		**Validation Set**		**Training Set**	
Perc. Train	**Graph (Distance)**	**Graph (Correlation)**	**Graph (Distance)**	**Graph (Correlation)**	**Graph (Distance)**	**Graph (Correlation)**
70%	**16.42**	27.04	**13.67**	14.92	13.40	**7.96**

Table 2 shows the root mean squared error (RMSE) over the different sets (training, validation and test) with a varying number of training data. Let us remark as described more precisely in Section 3.2 that since we use the same number of data, increasing the size of training set reduces the size of both the validation and test sets. As expected, since the proposed model is a generalization of both the ridge and OLS solution, smaller RMSE is obtained on all configurations. More importantly, the proposed model allows us to obtain a quite significant improvement on the test set compared to both the ridge and the OLS

solutions, which clearly demonstrates the superiority in terms of the generalization of the proposed model.

Table 2. RMSE of the different regression models for different sizes of the training set.

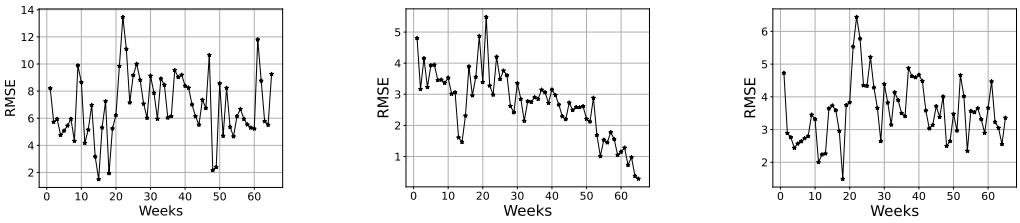
Root Mean Square Error (RMSE)									
	Testing Set			Validation Set			Training Set		
Perc. Train	**Graph Reg.**	**Ridge**	**OLS**	**Graph Reg.**	**Ridge**	**OLS**	**Graph Reg.**	**Ridge**	**OLS**
50%	**35.65**	41.43	42.10	**16.80**	17.86	17.65	9.13	6.74	**6.55**
60%	**30.02**	36.77	41.41	**15.02**	19.60	19.73	21.73	**6.52**	**6.52**
70%	**16.42**	22.65	49.52	**13.67**	17.13	16.44	13.40	7.94	**7.02**

Next, in Table 3 we present the RMSE obtained when the models are applied without any lagged variables as covariates. By comparing the values obtained with these variables in Table 2, we can clearly see the benefit of using an auto-regressive structure in the regression model by the introduction of such lagged response variables.

Table 3. RMSE of the different regression models without the use of the lagged response variables as covariates.

Root Mean Square Error (RMSE) without Lagged Variables									
	Testing Set			Validation Set			Training Set		
Perc. Train	**Graph Reg.**	**Ridge**	**OLS**	**Graph Reg.**	**Ridge**	**OLS**	**Graph Reg.**	**Ridge**	**OLS**
70%	**38.54**	**38.54**	41.76	**20.28**	**20.28**	20.34	9.65	9.65	**9.64**

In Figure 6, the weekly RMSE is depicted as a function of time for three different counties. These weekly RMSEs are obtained by aggregating the daily forecasted values from the proposed regression model which is trained on 50% of the dataset. It is interesting to observe that the weekly RMSE does not explode with time but rather stays quite stable with respect to time.

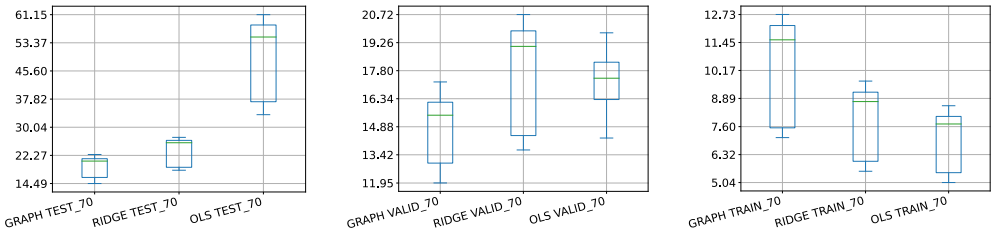


(**a**) Union County (**b**) Red River Parish (**c**) Lebanon County

Figure 6. RMSE as a function of time for three different counties.

In order to ensure that the previously observed conclusions are not too sensitive to the specific 59 chosen counties, we compute the RMSE on the three different sets for the different regression models by randomly selecting 2 counties for each of the 19 states. Let remark that we use transfer learning for the hyperpamemeters of the models (i.e., γ_1 and γ_2). They are not optimized on each random choice of data but are set to their optimized values in the previous scenario in which all 59 counties are used. From the results depicted in Figure 7, the same conclusions as before can be drawn. It is worth noting that, even if the hyperparameters are not optimized for each random choice, the RMSE on the validation set is still smaller using the proposed model. Finally, the boxplots obtained on the test sets

empirically show better predictive power for the proposed penalized regression model over graph prediction.



(**a**) RMSE on testing set (**b**) RMSE on validation set (**c**) RMSE on training set

Figure 7. Boxplots of the RMSE obtained after 50 random choices of two counties per state for the different regression models (70% of the dataset is used for training).

5. Conclusions

In this paper, we propose a novel GLM-based spatio-temporal mixed-effect model with graph penalties. This graph penalization allows us to take into account the inherent structural dependencies or relationships of the data. Another advantage of this model is its ability to model more complicated and realistic phenomena through the use of generalized linear models (GLMs). To illustrate the performance of our model, a publicly available dataset from the National Centers for Environmental Information (NCEI) in the United States of America is used, where we perform statistical inference of future CO_2 emissions over 59 counties. We show that the proposed method outperforms widely used methods, such as the ordinary least squares (OLS) and ridge regression models. In the future, we will further study how to improve this model to this specific CO_2 prediction. In particular, the use of different likelihood and link functions will be studied along with other adjacency matrices. We will also study whether considering, for the graph penalties, time differences instead of the direct mean values as discussed in Section 3.1 could improve the prediction accuracy. Finally, it will be interesting to connect this prediction model to some decision-making problems as in [22].

Author Contributions: Conceptualization, R.T., F.S., I.N. and G.W.P.; methodology, R.T., F.S., I.N. and G.W.P.; software, R.T. and F.S.; writing—original draft preparation, R.T., F.S., I.N. and G.W.P.; writing—review and editing, R.T., F.S., I.N. and G.W.P.; visualization, R.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available on request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Proposition 1

With the normal assumption of the response variables, the resulting estimator $\hat{\boldsymbol{\beta}}$ defined in (15) is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) + \gamma_1 \boldsymbol{\beta}^\top \boldsymbol{\beta} + \gamma_2 \boldsymbol{\beta}^\top \boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}} \boldsymbol{\Phi} \boldsymbol{\beta}.$$

The partial derivative with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned}\frac{\partial C(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) + 2\gamma_1\boldsymbol{\beta} + 2\gamma_2\boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}}\boldsymbol{\Phi}\boldsymbol{\beta} \\ &= -2\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{y} + 2\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\boldsymbol{\beta} + 2\gamma_1\boldsymbol{\beta} + 2\gamma_2\boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}}\boldsymbol{\Phi}\boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\frac{\partial C(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 &\Leftrightarrow \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\hat{\boldsymbol{\beta}} + \gamma_1\hat{\boldsymbol{\beta}} + \gamma_2\boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}}\boldsymbol{\Phi}\hat{\boldsymbol{\beta}} = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ &\Leftrightarrow \left(\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \gamma_1\boldsymbol{I}_{NP} + \gamma_2\boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}}\boldsymbol{\Phi}\right)\hat{\boldsymbol{\beta}} = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}\end{aligned}$$

We finally obtain that

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \gamma_1\boldsymbol{I}_{NP} + \gamma_2\boldsymbol{\Phi}^\top \widetilde{\boldsymbol{L}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}$$

Appendix B. List of Counties Used in the Numerical Study

Table A1. List of counties.

List of Counties					
Number	**Counties**	**States**	**Number**	**Counties**	**States**
1	Anoka County	Minnesota	31	Daviess County	Kentucky
2	Dakota County	Minnesota	32	Hopkins County	Kentucky
3	Lyon County	Minnesota	33	Russel County	Kentucky
4	Buchanan County	Iowa	34	Alamance County	North Carolina
5	Crawford County	Iowa	35	Lenoir County	North Carolina
6	Page County	Iowa	36	Pender County	North Carolina
7	Union County	Iowa	37	Randolph County	North Carolina
8	Ashley County	Arkansas	38	Charleston County	South Carolina
9	Columbia County	Arkansas	39	Dillon County	South Carolina
10	Outagamie County	Wisconsin	40	Lee County	South Carolina
11	Dane County	Wisconsin	41	Marlboro County	South Carolina
12	Clark County	Illinois	42	Pickens County	South Carolina
13	Mercer County	Illinois	43	Bartholomew County	Indiana
14	Ogle County	Illinois	44	Posey County	Indiana
15	Stephenson County	Illinois	45	Mahoning County	Ohio
16	Lawrence County	Tennessee	46	Shelby County	Ohio
17	Obion County	Tennessee	47	Delta County	Michigan
18	Cumberland County	Tennessee	48	Montcalm County	Michigan

Table A1. *Cont.*

List of Counties					
Number	**Counties**	**States**	**Number**	**Counties**	**States**
19	Hinds County	Mississipi	49	Washtenaw County	Michigan
20	Tate County	Mississipi	50	Armstrong County	Pennsylvania
21	Blount County	Alabama	51	Montour County	Pennsylvania
22	Autauga County	Alabama	52	Lebanon County	Pennsylvania
23	Marengo County	Alabama	53	Luzerne County	Pennsylvania
24	Morgan County	Alabama	54	Addison County	Vermont
25	Talladega County	Alabama	55	Windsor County	Vermont
26	Bulloch County	Georgia	56	Grant Parish	Louisiana
27	Habersham County	Georgia	57	Red River Parish	Louisiana
28	Bradford County	Florida	58	Vermilion Parish	Louisiana
29	Clay County	Florida	59	Madison Parish	Louisiana
30	Taylor County	Florida			

References

1. Cressie, N.; Wikle, C. *Statistics for Spatio-Temporal Data*; Wiley: Hoboken, NJ, USA , 2011.
2. Wikle, C. Modern Perspectives on Statistics for Spatio-Temporal Data. *Wires Comput. Stat.* **2014**, *7*, 86–98. [CrossRef]
3. Wikle, C.K.; Zammit-Mangion, A.; Cressie, N. *Spatio-Temporal Statistics with R*; Chapman & Hall/CRC: Boca Raton, FL, USA , 2019.
4. Stroup, W. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*; Chapman & Hall/CRC Texts in Statistical Science; Chapman & Hall/CRC: Boca Raton, FL, USA , 2012.
5. St-Pierre, J.; Oualkacha, K.; Bhatnagar, S.R. Efficient penalized generalized linear mixed models for variable selection and genetic risk prediction in high-dimensional data. *Bioinformatics* **2023**, *39*, btad063.
6. Schelldorfer, J.; Meier, L.; Bühlmann, P. GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ_1-Penalization. *J. Comput. Graph. Stat.* **2014**, *23*, 460–477. [CrossRef]
7. Shuman, D.I.; Narang, S.K.; Frossard, P.; Ortega, A.; Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **2013**, *30*, 83–98. [CrossRef]
8. Qiu, K.; Mao, X.; Shen, X.; Wang, X.; Li, T.; Gu, Y. Time-Varying Graph Signal Reconstruction. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 870–883. [CrossRef]
9. Giraldo, J.H.; Mahmood, A.; Garcia-Garcia, B.; Thanou, D.; Bouwmans, T. Reconstruction of Time-Varying Graph Signals via Sobolev Smoothness. *IEEE Trans. Signal Inf. Process. Over Netw.* **2022**, *8*, 201–214. [CrossRef]
10. Belkin, M.; Niyogi, P.; Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *J. Mach. Learn. Res.* **2006**, *7*, 2399–2434.
11. Venkitaraman, A.; Chatterjee, S.; Händel, P. Predicting Graph Signals Using Kernel Regression Where the Input Signal is Agnostic to a Graph. *IEEE Trans. Signal Inf. Process. Over Netw.* **2019**, *5*, 698–710. [CrossRef]
12. Karakurt, I.; Aydin, G. Development of regression models to forecast the CO_2 emissions from fossil fuels in the BRICS and MINT countries. *Energy* **2023**, *263*, 125650. [CrossRef]
13. Fouss, F.; Saerens, M.; Shimbo, M. *Algorithms and Models for Network Data and Link Analysis*; Cambridge University Press: Cambridge, UK , 2016.
14. Aitken, A.C. On Least-squares and Linear Combinations of Observations. *Proc. R. Soc. Edinb.* **1936**, *55*, 42–48. [CrossRef]
15. Nelder, J.A.; Baker, R. *Generalized Linear Models*; Wiley Online Library: Hoboken, NJ, USA , 1972.
16. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989; p. 500.
17. Denison, D.G. *Bayesian Methods for Nonlinear Classification and Regression*; John Wiley & Sons: Hoboken, NJ, USA , 2002; Volume 386.
18. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
19. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
20. Hjorth, U.; Hjort, U. Model Selection and Forward Validation. *Scand. J. Stat.* **1982**, *9*, 95–105.

21. Gurney, K.R.; Liang, J.; Patarasuk, R.; Song, Y.; Huang, J.; Roest, G. The Vulcan Version 3.0 High-Resolution Fossil Fuel CO_2 Emissions for the United States. *J. Geophys. Res. Atmos.* **2020**, *125*, e2020JD032974. [CrossRef] [PubMed]
22. Nevat, I.; Mughal, M.O. Urban Climate Risk Mitigation via Optimal Spatial Resource Allocation. *Atmosphere* **2022**, *13*, 439. [CrossRef]