Enhancing Predictive Modeling of Nested Spatial Data through Group-Level Feature Disaggregation

Boyang Liu
Michigan State University
East Lansing, Michigan
liuboya2@msu.edu

Pang-Ning Tan
Michigan State University
East Lansing, Michigan
ptan@cse.msu.edu

Jiayu Zhou
Michigan State University
East Lansing, Michigan
jiayuz@msu.edu

ABSTRACT

Multilevel modeling and multi-task learning are two widely used approaches for modeling nested (multi-level) data, which contain observations that can be clustered into groups, characterized by their group-level features. Despite the similarity of the problems they address, the explicit relationship between multilevel modeling and multi-task learning has not been carefully examined. In this paper, we present a comparative analysis between the two methods to illustrate their strengths and limitations when applied to two-level nested data. We provide a detailed analysis demonstrating the equivalence of their formulations under a mild condition from an optimization perspective. We also demonstrate their limitations in terms of their predictive performance and especially, their difficulty in identifying potential cross-scale interactions between the local and group-level features when applied to datasets with either a small number of groups or limited training examples per group. To overcome these limitations, we propose a novel method for disaggregating the coarse-scale values of the group-level features in the nested data. Experimental results on both synthetic and real-world data show that the disaggregated group-level features can help enhance the prediction accuracy of the models significantly and identify the cross-scale interactions more effectively.

KEYWORDS

Nested data, Multi-task learning, Multilevel modeling

ACM Reference Format:
Boyang Liu, Pang-Ning Tan, and Jiayu Zhou. 2018. Enhancing Predictive Modeling of Nested Spatial Data through Group-Level Feature Disaggregation. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220091

1 INTRODUCTION

Nested data are prevalent across many application domains, from ecology and environmental sciences to education and bioinformatics [4, 23, 26, 7]. Such data contain observations sampled from individuals that are clustered into groups, with coarse-level features available to characterize properties of individuals belonging to the same group. Table 1 shows a toy example of a two-level ecological dataset, where each observation corresponds to a set of measurements pertaining to a lake. The lakes are grouped into coarser scale regions known as hydrological units, identified by their unique code called HUC ID. Each lake is also characterized by a fine-level ("local") feature, maximum lake depth, and a group-level ("regional") feature, average HUC temperature. These local and regional features can be used to predict a lake nutrient concentration variable such as total phosphorous.

Table 1: A toy example of a two-level lake ecology data

Lake ID	HUC ID	Maximum Lake Depth	Average HUC Temperature	Total Phosphorous
1	1	8.2	73	3
2	1	15.2	73	2.4
3	1	14.6	73	4.3
4	2	10.6	67	6
5	2	14.8	67	9
6	3	11.2	74	3.6

Traditional regression methods are not well-suited for modeling such nested data as the observations are not independent of each other. Alternative statistical methods, such as those based on multilevel models (MLM), also known as mixed models or hierarchical linear models [11], have thus been proposed to handle the nested data. MLM provides a principled way to integrate the local features with group-level features and can be used to infer the presence of cross-scale interactions in the data [8, 23]. A cross-scale interaction (CSI) refers to the joint effect between the local and group-level features on the response variable of interest [23]. For example, consider the following linear model for predicting the total phosphorous in lake i located in region j:

$$\mathrm{TP}_{ij} = \beta_L \mathrm{Depth}_{ij} + \beta_R \mathrm{Temp}_j + \beta_{\mathrm{CSI}} \Big(\mathrm{Depth}_{ij} \times \mathrm{Temp}_j \Big) + \beta_0. \quad (1)$$

The β_{CSI} coefficient in the preceding equation provides an estimate of the magnitude of cross-scale interaction between the local feature, i.e., maximum lake depth, and the regional feature, i.e., average HUC temperature, on the response variable, total phosphorous (TP). Detection of CSIs is an important research problem as it can help reveal the nonlinear relationships that exist in a complex system. For example, previous studies have shown the existence of CSI pattern between broad-scale hurricane-induced disturbance and fine-scale historical land use, which influences the biodiversity of land snails [28]. Previous studies have also found a strong evidence

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220091

Table 2: A two-level lake ecology data with disaggregated local temperature. The shaded cells illustrate the collinearity problem, as the values in the fifth column is simply a scalar multiplication of the values in the third column.

Lake ID	HUC ID	Max Depth	Avg HUC Temp	Depth × Avg HUC Temp	Total Phosphorous
1	1	8.2	73	8.2 × 73	3
2	1	15.2	73	15.2 × 73	4.3
3	1	14.6	73	14.6 × 73	2.4

(a) Difficulty of using regional features for finding CSI using MTL.

Lake ID	HUC ID	Max Depth	Recovered Local Temp	Depth × Local Temp	Total Phosphorous
1	1	8.2	72.5	8.2 × 72.5	3
2	1	15.2	73.5	15.2 × 73.5	4.3
3	1	14.6	73.0	14.6 × 73.0	2.4

(b) Using "localized" regional feature for finding CSI using MTL.

of CSI between local wetland cover and regional agriculture land use on nutrient concentration in freshwater lakes [8, 23].

From a machine learning perspective, the modeling of nested data can be naturally formulated as a multi-task learning (MTL) problem [5, 32, 6], where each task corresponds to learning the relationship between predictor and response variables for the observations in a group. Instead of training the model for each group independently, MTL learns the models jointly by leveraging the common structure among the tasks. This strategy is particularly useful when there is limited training data available among some of the groups [34].

Despite the similarity of the problem addressed by MLM and MTL, the explicit relationship between the two methods has not been carefully investigated. In this paper, we present a comparative analysis between MLM and MTL when applied to a two-level nested dataset. Our goal is to shed light on their potential strengths and pitfalls, especially in terms of model accuracy and their ability to detect CSI patterns in the data. Specifically, we show that the inherent assumption in the model specification of MLM may lead to its suboptimal predictive performance and misinterpretation of CSI patterns when applied to datasets with limited number of groups or training instances. While MTL is generally helpful to improve prediction accuracy on datasets with imbalanced distribution of training data, it cannot capture the CSI patterns in nested data due to the rank deficiency problem. Table 2(a) provides a simple illustration of MTL's limitations. Since MTL fits a local model to each region, the local model cannot effectively utilize the regional feature (average HUC temperature) as its value is identical for all the lakes in the same region. Furthermore, adding a nonlinear CSI feature (Depth × Temperature) explicitly into the design matrix introduces collinearity [9] in the data (see the shaded columns of the table), making it impossible to separate the β_L coefficient of the local feature in Equation (1) from the β_{CSI} coefficient using existing MTL methods. Nevertheless, we show that it is possible to learn a unique solution for β_{CSI} through a subsequent post-processing step after learning the local model for each region using MTL. Similarly, a subsequent post-processing step can also be performed to improve the prediction accuracy of MLM. More importantly, our analysis suggests the equivalence between the two-stage MTL (i.e., applying MTL followed by a post-processing step) and the two-stage MLM (i.e., applying MLM followed by a post-processing step) as they both optimize the same objective function in different order of variables.

Finally, although the two-stage MLM and MTL methods can improve prediction accuracy and identify the CSI patterns within a nested dataset, these methods tend to perform poorly when the number of groups (regions) or the number of samples per group are small [27, 17]. To overcome this problem, we propose a novel framework to disaggregate the group-level feature values to their finer scales. For example, instead of using the average temperature for the whole region, it would be better to estimate the local temperature of each lake (see Table 2(b)) and use this finer grain information to fit the MTL model. Unfortunately, it would be impossible to recover the local values of the group-level features without any prior assumptions as there are infinitely many ways to disaggregate the values. To overcome this problem, we present a novel feature disaggregation framework that is suitable for nested data with spatial contiguity properties. We show that the disaggregated feature values can enhance accuracy of the prediction models and identify the CSI patterns more effectively.

2 RELATED WORK

The modeling of nested data has been widely studied by statisticians and computer scientists. Multilevel modeling (MLM) [11], also known as linear mixed model, hierarchical linear model, and random effect model, is a mature statistical method designed to learn a model that not only explains the differences between individual samples in a group, but also the differences between groups. Examples of MLM methods include the random intercept and random slope models [11]. These are special cases of the more general, cross level interaction (CLI) model [11], which considers the random effect on both the slope and intercept of the model.

Cross-scale interactions play an important role for understanding the complex interactions between processes operating at different scales in macrosystem ecology and other disciplines [23]. Such interactions may lead to surprising outcomes and can have significant impact on the ecosystem and society [21]. Since the influence of CSIs may not be as pronounced compared to the effects of the local and regional features, there have been several studies focusing on estimating the statistical power of CSIs [1, 17, 18, 27]. Various methods have also been developed to measure CSIs in nested data. For example, Soranno et al. [23] employed a Bayesian hierarchical model to estimate the CSI between local wetland and regional agriculture. However, such methods suffer from the insufficient sample size problem, which may lead to contradicting cross-scale interactions [10].

Multi-task learning (MTL) is another commonly used approach to deal with grouped data, whereby the modeling of observations in each group is considered a separate learning task. A survey on MTL approaches can be found in [33]. There have been several recent studies focusing on the application of MTL to spatio-temporal data. For example, Xu et al. [29] proposed a multi-task framework that assumes a low rank clustering structure among the different regions. Lin et.al. [14] developed a MTL approach with feature interaction whereas Lozano et al. [16] presented a multi-level lasso approach. However, these approaches considered only the group

membership of the observations and did not utilize the group-level feature information. In contrast, Yuan et al. [31] recently proposed a multi-task multi-level modeling approach that utilizes both the local and group-level features of the nested data. The approach assumes that the local and group-level features share some common latent factors, but the CSI patterns found using the approach vary by regions, which makes it harder to be interpreted.

3 BACKGROUND

This section formalizes the learning problem and reviews the application of MLM and MTL to nested data. We consider linear models in this study as they have been widely used in the modeling of nested data in ecology, climate science, and other application domains [23, 29]. Not only do linear models have a lower bias, which makes them more robust to overfitting for small sample size problems, they can also identify salient features in the data.

3.1 Problem Statement

Consider a two-level nested dataset $\mathcal{D} = \{\mathbf{X}_j, \mathbf{z}_j, \mathbf{y}_j\}_{j=1}^{M}$, where $\mathbf{X}_j \in \mathbb{R}^{n_j \times d_L}$ is a matrix containing the fine-level (local) features of observations in group (region) j, $\mathbf{z}_j \in \mathbb{R}^{d_R \times 1}$ is the corresponding vector of group-level (regional) features, and $\mathbf{y}_j \in \mathbb{R}^{n_j \times 1}$ is a vector containing the values of response variables for the observations in group j. Note that the terms *group* and *region* will be used interchangeably throughout this paper. The subscript j refers to the index of a group, M denotes the number of groups, d_L is the number of local features, d_R is the number of group-level features, and n_j is the number of observations that belong to the j^{th} group.

One approach is to fit a linear model globally to the entire nested data. The same *global model* will then be applied to make predictions in all the regions. Since the approach assumes that the regression coefficients are invariant across different regions, its predictive performance is likely to be poor as it ignores the inherent spatial heterogeneity of the data [10]. Alternatively, a local model can be trained for each region using only the training instances available for the region. This *independent modeling* approach is also likely to perform poorly for two reasons. First, as shown in Table 1, the approach cannot effectively utilize the group-level features as their values are identical for all observations from the same region. Second, the local models are susceptible to overfitting especially for regions with limited labeled data. Due to these limitations, alternative methods have been developed for modeling nested data, including the MLM and MTL approaches to be described next.

3.2 MLM for Nested Data

A two-level linear model for the i-th observation in region j is given by the following equations [11]:

$$y_{i,j} = \mathbf{x}_{i,j}^T \boldsymbol{\beta}_j + \beta_{0j} + \epsilon_{i,j}, \tag{2}$$

$$\boldsymbol{\beta}_j = \mathbf{G}z_j + \boldsymbol{\gamma} + \mathbf{u}_j, \tag{3}$$

$$\beta_{0j} = \mathbf{z}_j^T \mathbf{w}^R + \gamma_0 + u_{0j}, \tag{4}$$

$$\mathbf{u}_j \sim N(\mathbf{0}, \Sigma_u^2),\ u_{0_j} \sim N(0, \sigma_{u0}^2),\ \epsilon_{i,j} \sim N(0, \sigma_\epsilon^2), \tag{5}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{d_L \times 1}$ is a vector of coefficients representing the slope of the local linear model with respect to each local feature and β_{0j} corresponds to the model intercept. These local regression coefficients are related to the group-level features $\mathbf{z}_j$ according to Eqs. (3) and (4), respectively, where $\mathbf{w}_j^R \in \mathbb{R}^{d_R \times 1}$ is a vector of coefficients for the group-level features and $\mathbf{G} \in \mathbb{R}^{d_L \times d_R}$ is a matrix of coefficients for the cross-scale interaction between the local and group-level features.

The preceding model can be further simplified by substituting Eqs. (3) and (4) into Eq. (2), which leads to the following formulation, also known as the cross-level interaction (CLI) model [11]:

$$y_{i,j} = \mathbf{x}_{i,j}^T(\gamma + u_j) + \mathbf{z}_j^T \mathbf{w}^R + \mathbf{x}_{i,j}^T \mathbf{G}\mathbf{z}_j + \gamma_0 + (u_{0j} + \epsilon_{i,j}) \tag{6}$$

$$= \mathbf{x}_{i,j}^T \gamma + \mathbf{z}_j^T \mathbf{w}^R + \mathbf{x}_{i,j}^T \mathbf{G}\mathbf{z}_j + \gamma_0 + \mathbf{x}_{i,j}^T u_j + \eta_{ij} \tag{7}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2),\ \mathbf{u}_j \sim N(0, \Sigma_u^2),\ \epsilon_{i,j} \sim N(0, \sigma_\epsilon^2).$$

The parameters of the model, i.e., γ_0, $\boldsymbol{\gamma}$, $\mathbf{w}^R$, and $\mathbf{G}$, along with the variance components σ_{u0}^2, Σ_u, and σ_ϵ^2, are often estimated using the restricted maximum likelihood approach [12]. CLI is a popular two-level MLM approach for inferential data analysis [13] as it allows us to quantify the different types of relationships present in the data, such as cross-scale interactions ($\mathbf{G}$), along with their standard errors. However, it has an inherent limitation when applied to predictive modeling problems since it provides only the global estimate for $\boldsymbol{\gamma}$, γ_0, Σ_u^2, and σ_{u0}^2 instead of explicitly calculating the region-specific values $\mathbf{u}_j$ and u_{0j}. During the prediction step, we may compute the expected value of the response variable for any given test instance $(\mathbf{x}^*, \mathbf{z}_j)$ as follows:

$$E[y|\mathbf{x}^*, \mathbf{z}_j] = \mathbf{x}^{*T}\boldsymbol{\gamma} + \mathbf{z}_j^T \mathbf{w}^R + \mathbf{x}^{*T}\mathbf{G}\mathbf{z}_j + \gamma_0. \tag{8}$$

Note the difference between the expected value given in Eq. (2), i.e., $E[y|\mathbf{x}^*, \mathbf{z}_j] = \mathbf{x}^{*T}\boldsymbol{\beta}_j + \beta_{0j}$, from the one given in Eq. (8). The latter applies the same prediction function to all regions since the coefficients $\boldsymbol{\gamma}$, $\mathbf{w}^R$, $\mathbf{G}$, and γ_0 are independent of j. More importantly, Eq. (8) excludes the random effects u_j and u_{0j} from the model prediction, which explains its poor predictive performance when there are significant spatial heterogeneity in the data [10].

3.3 MTL for Nested Data

In MTL, the prediction problem for each region is treated as a separate learning task. However, unlike the independent modeling approach, the local models are simultaneously trained to optimize a joint objective function for all the regions:

$$\underset{\{\boldsymbol{\beta}_j\}, \boldsymbol{\beta}_{0j}}{\arg\min} \sum_{j=1}^{M} \|\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j - \beta_{0j}\mathbf{1}\|_2^2 + \Omega(\mathbf{B}, \mathbf{B}_0), \tag{9}$$

where $\mathbf{1} = [1, 1, \cdots, 1]^T$, $\mathbf{B}_0 = [\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \cdots, \boldsymbol{\beta}_{0M}]$, $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_M]$, and $\Omega(\mathbf{B}, \mathbf{B}_0)$ is a regularization term that relates the model parameters from different regions.

MTL is an effective approach for predictive modeling of nested data since it promotes information sharing between different regions, which is particularly useful for regions with limited training data [29, 31]. Unlike the MLM formulation shown in Eq. (7), MTL attempts to directly solve the slope and intercept terms of the regression function for each region, similar to the Eq. (2) for MLM. To incorporate the regional features and cross-scale interaction terms into the MTL formulation, a naïve approach would be to modify

the objective function given in Eq. (9) as follows:

$$\underset{\{\mathbf{w}_j^L\},\mathbf{w}^R,\mathbf{G}}{\arg\min} \sum_{j=1}^{M} \|\mathbf{y}_j - \mathbf{X}_j\mathbf{w}_j^L - \mathbf{Z}_j^T\mathbf{w}^R - \mathbf{X}_j\mathbf{G}\mathbf{z}_j^T - \boldsymbol{\gamma}_{0j}\|_2^2 + \Omega\big[\{\mathbf{w}_j^L\}, w^R, G\big], \tag{10}$$

where we have replaced $\boldsymbol{\beta}_j = \mathbf{G}\mathbf{z}_j + (\gamma + u_j) = \mathbf{G}\mathbf{z}_j + \mathbf{w}_j^L$ using Eq. (3) and $\beta_{0j} = \mathbf{z}_j^T\mathbf{w}^R + \gamma_{0j}$ using Eq. (4). Furthermore, $\mathbf{Z}_j \equiv [\mathbf{z}_j, \mathbf{z}_j, \cdots, \mathbf{z}_j]$. As will be shown in the next section, this approach may not be able to learn the regression coefficients and cross-scale interactions correctly due to the ill-posed nature of its formulation.

4 RELATIONSHIP BETWEEN MLM AND MTL

In this section, we perform a comparative analysis between MLM and MTL, and show that the performance of both approaches can be improved by applying a subsequent postprocessing step to refine the models. We termed these approaches as two-stage MLM and two-stage MTL, respectively. We then illustrate the equivalence between these two approaches for two-level data in Section 4.3.

4.1 Two-Stage MLM

As noted in Section 3.2, most MLM implementations would provide only an estimate of $E[\mathbf{w}_j^L] = \boldsymbol{\gamma}$ and variance components such as Σ_u instead of the region specific values for $\mathbf{w}_j^L$ or $\boldsymbol{\beta}_j$. While this may be sufficient for inferential data analysis [13] to determine the existence of certain relationships in data, it is not optimal for prediction purposes. It is possible to refine the MLM predictions by re-fitting the model to the residuals computed from the regional features and the cross-scale interaction terms given in Eq. (8). Specifically, let $\hat{\mathbf{y}}_j = \mathbf{y}_j - \mathbf{z}_j^T\mathbf{w}^R - \mathbf{X}_j\mathbf{G}\mathbf{z}_j$ be the residual errors for all the instances in region j. By fitting a regression model between $\mathbf{X}_j$ and the residual error $\hat{\mathbf{y}}_j$, it is possible to recover $\mathbf{w}_j^L$ for each region using Eq. (7). The two-stage MLM approach can be summarized as follows:

Stage 1: Apply cross-level interaction model to the nested data:

$$(\mathbf{w}^R, \mathbf{G}, \gamma, \gamma_0) = \underset{\hat{\mathbf{w}},\hat{\boldsymbol{\gamma}},\hat{\gamma}_0,\hat{\mathbf{G}}}{\arg\min} \sum_j \|\mathbf{y_j} - \mathbf{X_j}\hat{\boldsymbol{\gamma}} - \mathbf{Z}_j^T\hat{\mathbf{w}} - \mathbf{X}_j\hat{\mathbf{G}}\mathbf{z}_j^T - \hat{\gamma}_0\mathbf{1}\|^2.$$

Stage 2: Learn $\mathbf{w}_j^L$ by regressing the residual error on $\mathbf{X}_j$:

$$\forall j:\ \hat{\mathbf{y}}_j = \mathbf{y}_j - \mathbf{z}_j^T\mathbf{w}^R - \mathbf{X}_j\mathbf{G}\mathbf{z}_j,\ \ \mathbf{w}_j^L = \arg\min_{\hat{\mathbf{w}}_j} \|\hat{\mathbf{y}}_j - \mathbf{X}_j\hat{\mathbf{w}}_j\|_2^2.$$

An unlabeled instance $(\mathbf{x}^*, \mathbf{z}_j)$ can then be predicted as follows:

$$E[y|\mathbf{x}^*, \mathbf{z}_j] = \mathbf{x}^{*T}\mathbf{w}_j^L + \mathbf{z}_j^T\mathbf{w}^R + \mathbf{x}^{*T}\mathbf{G}\mathbf{z}_j + \gamma_0. \tag{11}$$

Unlike Eq. (8), the prediction function here accounts for differences in $\mathbf{w}_j^L$ between the different regions.

4.2 Two-Stage MTL

For MTL, one way to estimate the regression coefficients $\mathbf{w}^L$ and $\mathbf{w}^R$ as well as the cross-scale interaction term $\mathbf{G}$ is by solving Eq. (10). However, since the value of the group-level feature is the same for all observations in the same region, this may lead to severe multicollinearity problem, as shown in Table 2. Worse still, comparing Eq. (9) to Eq. (10), solving $\mathbf{w}_j^L$ and $\mathbf{G}$ from $\boldsymbol{\beta}_j$ essentially requires optimizing the following objective functions:

$$\underset{\{\mathbf{w}_j^L\},\mathbf{G}}{\arg\min} \sum_{j=1}^{M} \|\boldsymbol{\beta}_j - \mathbf{G}\mathbf{z}_j - \mathbf{w}_j^L\|_2^2 \iff \underset{\mathbf{W}^L,\mathbf{G}}{\arg\min} \|\mathbf{B} - \mathbf{G}\mathbf{Z} - \mathbf{W}^L\|_F^2.$$

Unfortunately, this is an ill-posed problem since there are infinitely many ways to decompose $\mathbf{B}$ into the sum of $\mathbf{W}^L$ and $\mathbf{G}\mathbf{Z}$ given $\mathbf{Z}$. Thus, the naïve approach of optimizing Eq. (10) will not be able to recover the coefficients $\mathbf{w}_j^L$ and $\mathbf{G}$ accurately. The same argument also applies to the decomposition of β_{0j} into $\mathbf{w}_R$ and γ_0.

To overcome this limitation, a two-stage approach can be used to recover the regression coefficients and cross-scale interaction terms. First, we apply standard MTL methods to learn the $\boldsymbol{\beta}_j$'s and $\boldsymbol{\beta}_{0j}$'s for all the regions (see Eq. (9)). During the second stage, the $\boldsymbol{\beta}_j$'s and $\boldsymbol{\beta}_{0j}$'s will be fitted against their corresponding group-level features $\mathbf{z}_j$ to learn the $\mathbf{G}$, $\mathbf{w}^R$, and $\boldsymbol{\gamma}$. The regression coefficients for the local features $\mathbf{w}_j^L$ can then be obtained using Eq. (3). The two-stage MTL approach can be summarized as follows:

Stage 1: Apply MTL to learn $\boldsymbol{\beta}_j$ and β_{0j} by solving Eq. (9).

Stage 2: $(\boldsymbol{\gamma}, \mathbf{G}) = \arg\min_{\hat{\boldsymbol{\gamma}},\hat{\mathbf{G}}} \sum_{j=1}^{M} \|\boldsymbol{\beta}_j - \hat{\mathbf{G}}\mathbf{z}_j - \hat{\boldsymbol{\gamma}}\|_2^2$

$(\mathbf{w}^R, \gamma_0) = \arg\min_{\hat{\mathbf{w}},\hat{\gamma}} \sum_{j=1}^{M} (\beta_{0j} - \mathbf{z}_j^T\hat{\mathbf{w}} - \hat{\gamma_0})^2$

$\mathbf{w}_j^L = \boldsymbol{\beta}_j - \mathbf{G}\mathbf{z}_j.$

Note that the key difference between the two-stage MTL approach and the ill-posed matrix decomposition problem stated earlier is that the former imposes an additional constraint on $\mathbf{w}_j^L$, namely that it can be decomposed into $\boldsymbol{\gamma}$ and a Gaussian noise term with mean zero, thereby reducing the number of possible solutions.

4.3 On the Equivalence between MLM and MTL

It is not difficult to see the connection between the MTL and MLM formulations. First, consider the cross-level interaction model given in Eq. (6). The formulation is obtained by substituting the regression coefficients shown in Eq. (3) and Eq. (4) into Eq. (2). The objective function for MTL, which is given by Eq. (9), is equivalent to learning the $\boldsymbol{\beta}_j$'s and β_{0j}'s in Eq. (2) directly without considering the constraints imposed by Eqs. (3) and (4). Without such constraints, MTL will not be able to obtain unique solutions for $\mathbf{w}_j^L$, $\mathbf{w}_R$, and $\mathbf{G}$ due to the ill-posed nature of the problem, as given in Eq. (10). Estimating these coefficients are useful for many applications as the coefficients convey important information about the important relationships that exist in the data. By employing a two-stage MTL approach, we can restrict the solution space to satisfy a feasible set defined by the constraints given in Eqs. (3) and (4) by regressing the $\boldsymbol{\beta}_j$'s and β_{0j}'s against the group-level features $\mathbf{z}_j$'s. The slopes of these linear functions would determine $\mathbf{G}$ and $\mathbf{w}^R$, while their intercepts can be used to identify $\mathbf{w}_j^L$ and γ_{0j}. The two-stage MTL approach can thus be stated as the following constrained optimization problem:

$$\begin{aligned} &\underset{\{\boldsymbol{\beta}_j,\beta_{0j}\}}{\arg\min} \sum_{j=1}^{M} \|\mathbf{X}_j\boldsymbol{\beta}_j + \beta_{0j}\mathbf{1} - \mathbf{y_j}\|_2^2, \\ \text{s.t.}\quad &\boldsymbol{\beta}_j = \mathbf{w}_j^L + \mathbf{G}\mathbf{z}_j,\ \beta_{0j} = \mathbf{z}_j^T\mathbf{w}^R + \gamma_{0j}, \\ &\mathbf{w}_j^L = \boldsymbol{\gamma} + \mathbf{u}_j,\ \gamma_{0j} = \gamma_0 + u_{0j}. \end{aligned} \tag{12}$$

The first stage involves solving the unconstrained optimization problem using MTL[1], whereas the second stage projects the initial solution to the feasible set.

The equivalent formulation for two-stage MLM can be found by substituting the equality constraints in Eq. (12) directly into the objective function to remove $\boldsymbol{\beta}_j$ and β_{0j} from the formulation. This leads to the following constrained optimization problem for two-stage MLM:

$$\underset{\mathbf{w}_R, \boldsymbol{\gamma}, \gamma_0, \mathbf{G}}{\arg\min} \sum_{j=1}^{M} \|\mathbf{X}_j\boldsymbol{\gamma} + \mathbf{Z}_j^T\mathbf{w}^R + \mathbf{X}_j\mathbf{G}\mathbf{z}_j + \gamma_0\mathbf{1} - \mathbf{y}_j\|_2^2$$
$$\text{s.t. } \mathbf{w}_j^L = \boldsymbol{\gamma} + \mathbf{u}_j,\ \gamma_{0j} = \gamma_0 + u_{0j}. \quad (13)$$

The unconstrained optimization problem solves the cross-level interaction model in the first stage, and the second stage learns the varying slopes $\mathbf{w}_j^L$ and intercepts γ_{0j} of the local models.

Note that the following two conditions are needed to ensure good performance by the two-stage MTL and two-stage MLM approaches:

(1) There must be sufficient number of samples available in each region to correctly estimate the $\boldsymbol{\beta}$'s in order to improve prediction accuracy.
(2) There must be sufficient number of regions available in the nested data in order to correctly estimate the cross-scale interaction matrix $\mathbf{G}$.

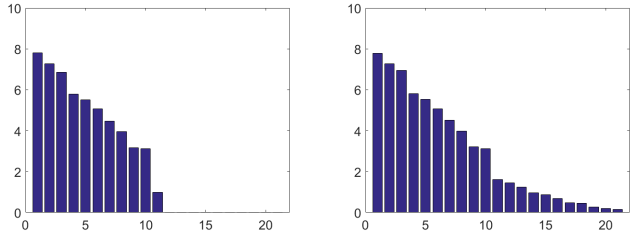
5 DISAGGREGATION OF GROUP-LEVEL FEATURE VALUES

For many applications in climate and ecology domains, the coarse-level features of the nested data are often computed by aggregating the corresponding feature values at finer spatial scales. While the coarse-scale value helps to summarize the feature information from adjacent locations into a single value, it also introduces collinearity into the data, a problem illustrated by the example shown in Table 2. To alleviate this problem, this section presents a novel framework for learning the disaggregated value of the group-level features. The disaggregated values are then incorporated into a multi-task multi-level learning framework called MTML_imputation.

5.1 Feature Disaggregation

To understand the rationale for disaggregating the group-level features, consider the nested data shown in Table 2(a). Since the average HUC temperature is identical for all lakes in the same region, the cross-scale interaction feature becomes correlated with the maximum depth feature (shaded cells in the table), which leads to a rank deficiency problem in the design matrix. As a result, existing MTL approaches will not be able to utilize the group-level features effectively in their formulation. By disaggregating the group-level features in a manner that is consistent with characteristics of the domain, it is possible to obtain a full-rank design matrix that can be better utilized by the MTL algorithm. For example, consider a nested data that contains 10 local and 1 group-level features. Analogous to Table 2(b), we created a synthetic dataset that contains the local, group-level, and cross-scale interaction features. We then applied

[1]For brevity, we have ignored the regularization term in the objective function as such regularization can also be applied during the second stage of the two-stage MLM.



(a) Group-level feature (b) Disaggregated group-level feature

Figure 1: Comparison between the singular value distribution of design matrices using group-level features (left) and disaggregated group-level features (right) for a given region. Both matrices contain 10 local, 1 group-level (or disaggregated group-level), and 10 cross-scale interaction features.

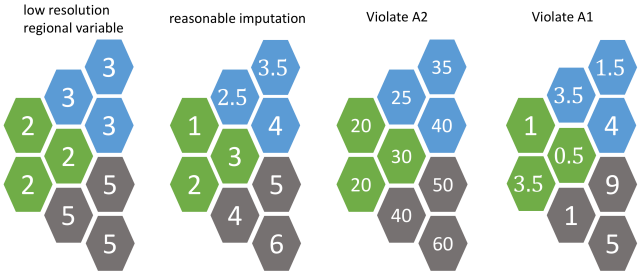


Figure 2: Comparison of different disaggregation results. Each hexagon is a sample, whose color represents its region and number represents its disaggregated value. The leftmost figure simply uses the group-level feature value. The second figure preserves both A1 and A2 assumptions. The last two figures violate assumptions A2 and A1, respectively.

singular value decomposition to the design matrix. Although there were 21 features in the design matrix, the rank of the matrix is only 11 as shown by the distribution of their singular values in Figure 1(a). Using our feature disaggregation approach, it is possible to increase the rank of the matrix by replacing the group-level feature values with their disaggregated values (see Figure 1(b)).

Although there may be other ways to increase the rank of the design matrix for MTL, e.g., by adding Gaussian noise to perturb the group-level features, their disaggregated values do not add any new information, and thus, may not help to improve the predictive performance of MTL. In principle, since there are many possible solutions to the feature disaggregation problem, realistic assumptions are needed to constrain the solution space in a way that is consistent with the domain expectation. For example, the spatial autocorrelation present in the data may provide useful guidance to aid the recovery of the disaggregated feature values.

For nested spatial data, we present a feature disaggregation approach based on the following two realistic assumptions:

- Assumption 1 (**A1**): *The disaggregated values should preserve the spatial continuity properties.*
- Assumption 2 (**A2**): *The disaggregated values should be close to its original group-level feature value.*

Assumption **A1** is a reflection of Tobler's first law of geography [24], which states that near things are related, and thus, should be more similar to each other. Assumption **A2** constrains the disaggregated values to be close to the original group-level feature value. Figure 2 illustrates the intuition behind our assumptions. Each hexagon represents a sampled observation, which is assigned to a group (region) based on its color. The leftmost diagram shows the original group-level feature value for each observation, which is the approach used in MLM and MTL. The second figure shows a reasonable recovery of the disaggregated values, which preserve both A1 and A2 assumptions. The disaggregation results shown in the third figure violates the A2 assumption, since the disaggregated values are quite different from their group-level feature value. Finally, the rightmost disaggregation violates assumption A1, since some neighboring hexagons have very different values.

The disaggregated values will be used in place of the original group-level feature values when fitting the local linear regression models to the nested data. Since the disaggregated values are no longer identical for all individuals in the same group, it is possible to fit the models to learn $\mathbf{w}_j^L$, $\mathbf{w}_R$, $\mathbf{G}$, and γ_{0j} from the data. Furthermore, the feature disaggregation and model building can be performed simultaneously, as will be discussed in the next section.

5.2 Proposed Formulation

Our proposed framework, called *MTML_Imputation*, for joint feature disaggregation and multi-task multi-level learning from nested spatial data can be stated as the following optimization problem:

$$\begin{aligned}\underset{\{\tilde{\mathbf{z}}_j, \mathbf{w}_j^L\}, \mathbf{w}_j^R, \mathbf{G}}{\arg\min} \quad & \sum_{j=1}^{M} \|\mathbf{y}_j - \mathbf{X}_j \mathbf{w}_j^L - \tilde{\mathbf{Z}}_j \mathbf{w}_j^R - \text{diag}(\mathbf{X}_j \mathbf{G} \tilde{\mathbf{Z}}_j^T)\|_2^2 \\ + \quad & \lambda_1 \Omega([W^L, W^R]) + \lambda_2 ||G||_1 \\ + \quad & \lambda_3 \sum_{p<q}^{N} D_{p,q} \|\tilde{\mathbf{z}}_p - \tilde{\mathbf{z}}_q\|_2^2 \\ + \quad & \lambda_4 \sum_{p=1}^{N} \|\tilde{\mathbf{z}}_p - \mathbf{z}_p\|_2^2, \end{aligned} \tag{14}$$

where $N = \sum_{j=1}^{M} n_j$ is the total number of observations, $\tilde{\mathbf{z}}_p$ is the disaggregated value of the group-level feature for the p-th observation, and $\mathbf{z}_p$ is its corresponding group-level feature value. Furthermore, $\tilde{\mathbf{Z}}_j$ denote a matrix consisting of all the disaggregated values of the group-level features in region j. The first term in the objective function measures the residual errors of the prediction models, whereas the second and third terms are regularization penalties to ensure model sparsity. The regularization term Ω applies to both $\mathbf{w}_j^L$ and $\mathbf{w}_j^R$, and can be an L_1, L_{21}, or L_2 norm. The fourth term of the objective function is based on Assumption A1, where $D_{p,q}$ is the spatial proximity between the p^{th} and q^{th} observations. The last term is used to enforce Assumption A2. Note that our proposed formulation allows $\mathbf{w}_j^R$ to vary by region since they can be applied to the different $\tilde{\mathbf{Z}}_j$ matrices unlike previous methods.

5.3 Optimization

We employ the block coordinate descent algorithm [25] to solve the optimization problem given in Eq. (14). To simplify the discussion, we adopt the following notations in the remainder of this section. If $\{\mathbf{p}_j\}_{j=1}^M$ is a set that contains M vectors, then $\mathbf{p}_{\text{vec}} = [\mathbf{p}_1; \cdots; \mathbf{p}_M]$ is a column vector obtained by concatenating the M column vectors of $\mathbf{p}_j$. Thus, $\mathbf{y}_{\text{vec}}$, W_{vec}^L and W_{vec}^R are obtained by concatenating the M vectors of $\{\mathbf{y}_j\}$, $\{\mathbf{w}_j^L\}$ and $\{\mathbf{w}_j^R\}$, respectively. Furthermore, let $\mathbf{G}_{rep}$ be the Kronecker product between an $M \times M$ matrix of 1's and matrix $\mathbf{G}$, i.e.:

$$\mathbf{G}_{\text{rep}} = \mathbf{1}_M \mathbf{1}_M^T \otimes \mathbf{G} = \begin{bmatrix} G & \dots & G \\ \vdots & \ddots & \vdots \\ G & \dots & G \end{bmatrix}.$$

Also, let $\mathbf{X}_{\text{diag}}$ be a matrix obtained by concatenating the set of matrices, $\{\mathbf{X}_j\}_{j=1}^M$, along the block diagonal while $\mathbf{X}_{\text{stack}}$ is obtained by stacking the matrices on top of one another, i.e.,

$$\mathbf{X}_{\text{diag}} = \begin{bmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_M \end{bmatrix}, \ \mathbf{X}_{\text{stack}} = \begin{bmatrix} X_1 \\ \ddots \\ X_M \end{bmatrix}.$$

We also define the following vector and matrix:

$$\boldsymbol{\ell} = \mathbf{y}_{\text{vec}} - \mathbf{X}_{\text{diag}} \mathbf{w}_{\text{vec}}^L - \tilde{\mathbf{Z}}_{\text{diag}} \mathbf{w}_{\text{vec}}^R - \text{diag}(\mathbf{X}_{\text{diag}} \mathbf{G}_{rep} \tilde{\mathbf{Z}}_{\text{diag}}^T),$$

$$\mathbf{R} = \begin{bmatrix} \sqrt{D_{12}} & \sqrt{D_{13}} & \cdots & 0 & 0 & \cdots & 0 \\ -\sqrt{D_{12}} & 0 & \cdots & \sqrt{D_{23}} & \sqrt{D_{24}} & \cdots & 0 \\ 0 & -\sqrt{D_{13}} & \cdots & -\sqrt{D_{23}} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \sqrt{D_{N-1,N}} \\ 0 & 0 & 0 & 0 & 0 & \cdots & -\sqrt{D_{N-1,N}} \end{bmatrix},$$

where D_{pq} is the spatial proximity between p^{th} and q^{th} samples. The block coordinate descent algorithm would iteratively update the regression coefficients, the disaggregated values $\tilde{\mathbf{Z}}_j$'s and the cross-scale interaction matrix $\mathbf{G}$. During initialization, $\tilde{Z}$ is set to Z while the rest of the parameters are initialized randomly. The update formula for these parameters are summarized below.

a) Update formula for $\mathbf{W}^L$ and $\mathbf{W}^R$. Let $\mathbf{w}_{\text{vec}}$ be the vector obtained by concatenating the vectors in $\{\mathbf{w}_j^L\}$ and $\{\mathbf{w}_j^R\}$. By keeping only terms that depend on $\mathbf{w}_{\text{vec}}$, the objective function reduces to:

$$\min_{\mathbf{w}_{\text{vec}}} \|\mathbf{y}_{\text{diag}} - \mathbf{X}_{\text{diag}}^{LR} \mathbf{w}_{\text{vec}} - \text{diag}(\mathbf{X}_{\text{diag}} \mathbf{G}_{rep} \tilde{\mathbf{Z}}_{\text{diag}}^T)\|_2^2 + \lambda_1 \Omega(\mathbf{w}_{\text{vec}}),$$

where $\mathbf{X}_{\text{diag}}^{LR} = [\mathbf{X}, \tilde{\mathbf{Z}}]_{\text{diag}}$ is the concatenation of the local and disaggregated feature values. Since the objective function above is a convex function, it can be solved using a proximal gradient descent algorithm such as FISTA [3, 2, 15, 19]. The update formula for $\mathbf{w}_{\text{vec}}$ is given as follows:

$$\mathbf{w}_{\text{vec}}^{(s)} \leftarrow prox_\lambda(\mathbf{w}_{\text{vec}}^{(s-1)} - \lambda \nabla g(\mathbf{w}_{\text{vec}}^{(s-1)})),$$

where

$$\nabla g(\mathbf{w}_{\text{vec}}) = 2(\mathbf{X}_{\text{diag}}^{LR})^T \left(\mathbf{y}_{\text{vec}} - \mathbf{X}_{\text{diag}}^{LR} w_{\text{vec}} - \text{diag}\left(\mathbf{X}_{\text{diag}} \mathbf{G}_{rep} \mathbf{Z}_{\text{diag}}^T \right) \right)$$

Note that $g(w_{\text{vec}})$ denote the smooth part of the objective function for $\mathbf{w}_{\text{vec}}$. The choice of proximal operator depends on the form of the regularization function Ω.

b) Update formula for G. The terms in the objective function that depend on G are as follows:

$$\min_{\mathbf{G}} \|\mathbf{y}_{\text{vec}} - \mathbf{X}^{LR}_{\text{diag}}\mathbf{w}_{\text{vec}} - \text{diag}(\mathbf{X}_{\text{stack}}\mathbf{G}\tilde{\mathbf{Z}}^T_{\text{stack}})\|_2^2 + \lambda_2\|\mathbf{G}\|_1.$$

Since $\|\mathbf{G}\|_1$ is not a differentiable function, we can also apply the accelerated proximal gradient algorithm to minimize the equation. The update formula for G can be written as:

$$\mathbf{G}^{(s)} \leftarrow prox_\lambda(\mathbf{G}^{(s-1)} - \lambda\nabla g(\mathbf{G})), \quad (15)$$
$$\nabla g(\mathbf{G}) = 2\mathbf{X}^T_{\text{stack}}\ell\mathbf{Z}_{\text{stack}},$$

where $g(\mathbf{G})$ is the smooth part of the objective function for G.

c) Update formula for $\tilde{\mathbf{Z}}$. The objective function for $\tilde{\mathbf{Z}}$ is:

$$\min_{\tilde{\mathbf{Z}}_{\text{stack}}} \|\boldsymbol{\ell}\|^2 + \lambda_3\|\tilde{\mathbf{Z}}^T_{\text{stack}}\mathbf{R}\|_F^2 + \lambda_4\|\tilde{\mathbf{Z}}_{\text{stack}} - \mathbf{Z}\|_F^2.$$

Since it is a smooth function, we can apply Nesterov accelerated gradient method [19] to update $\tilde{\mathbf{Z}}_{\text{stack}}$. Taking its partial derivative with respect to $\tilde{\mathbf{Z}}_{\text{diag}}$, we obtain the following update formula:

$$\tilde{\mathbf{Z}}^{(s)}_{\text{stack}} \leftarrow \tilde{\mathbf{Z}}^{(s-1)}_{\text{stack}} - \lambda\nabla_{\tilde{\mathbf{Z}}}, \quad (16)$$
$$\nabla_{\tilde{\mathbf{Z}}} = \nabla_{\boldsymbol{\ell}} - 2\lambda_3\mathbf{H}_{\text{diag}} - 2\lambda_4\mathbf{V}_{\text{diag}},$$

where $\mathbf{H}_{\text{stack}} = \mathbf{R}(\tilde{\mathbf{Z}}^T_{\text{stack}}\mathbf{R})^T$, $\mathbf{V}_{\text{stack}} = \tilde{\mathbf{Z}}_{\text{stack}} - \mathbf{Z}$, and $\nabla_{\boldsymbol{\ell}}$ is a matrix whose i-th row is given by $\nabla_{\ell,i} = 2\boldsymbol{\ell}_i((\mathbf{w}^R_{\text{vec}})^T + \mathbf{X}_{\text{diag},i}G_{\text{rep}})$, where $\mathbf{X}_{\text{diag},i}$ is the i^{th} row of $\mathbf{X}_{\text{diag}}$.

5.4 Learning Framework

One of the challenges in applying the framework is the need to know the disaggregated feature value for observations that belong to the test data. Towards this end, we developed a semi-supervised version of our framework. The main difference between the supervised and semi-supervised approach is that the regularization term for the disaggregated values involve examples from both training and test sets, whereas the least square loss function involves only the training set. The spatial proximity matrix D is computed based on pairwise distance for all examples in the training and test sets. Thus, the gradient term in the update formula for the disaggregated values in the test set $\tilde{\mathbf{Z}}^*$ becomes:

$$\nabla_{\tilde{\mathbf{Z}}^*} = 2\lambda_3\mathbf{H}^*_{\text{diag}} - 2\lambda_4\mathbf{V}^*_{\text{diag}}, \quad (17)$$

where $\mathbf{H}^*_{\text{stack}} = \mathbf{R}^*(\tilde{\mathbf{Z}}^{*T}_{\text{stack}}\mathbf{R}^*)^T$ and $\mathbf{V}^*_{\text{stack}} = \tilde{\mathbf{Z}}^*_{\text{stack}} - \mathbf{Z}^*$.

6 EXPERIMENTAL EVALUATION

We have performed extensive experiments to evaluate the performance of our proposed framework, MTML_imputation. The code and datasets used in our experiments are available at https://github.com/illidanlab/region-Disaggregation

6.1 Datasets

We applied MTML_imputation on both synthetic and real-world datasets. The synthetic data can be used to demonstrate the efficacy of the approach when the true disaggregated values of the group-level features and cross-scale interaction matrix G are known.

Table 3: Summary statistics for LAGOS-NE lake ecology data

Response variable	TP	TN	Chla	Secchi
# regions	47	24	55	55
# instances	4009	1500	5314	5492
# instances/region	20-369	20-236	23-575	21-583

6.1.1 Synthetic Data. The nested data generated has $M = 10$ groups, with 30 samples per group. We first randomly generate a 10×2 matrix from a uniform distribution. For each row $\boldsymbol{v}$ of the matrix, we randomly sample 30 2-dimensional coordinate vectors from a Gaussian distribution, with a mean vector equals to $\boldsymbol{v}$ and a covariance given by a 2×2 identity matrix. A spatial proximity matrix D is then computed based on the 2-d coordinates of the 300 samples. To ensure that the group-level features are spatially autocorrelated, we applied a smooth function on the spatial coordinates to generate the disaggregated values $\tilde{\mathbf{Z}}_j$ for each group. We then compute the mean of the disaggregated values for each group as its group-level feature value, $\mathbf{z}_j$. In addition, the local feature values for all 300 samples, $\{\mathbf{X}_j\}$, are randomly generated from a Gaussian distribution with mean 0 and variance 1. The true values of $\mathbf{w}^L_j$, $\mathbf{w}^R$, and G are also generated randomly from a Gaussian distribution. These matrices are used to determine the true values of the response variable for samples in each group based on the following equation: $\mathbf{y}_j = \mathbf{X}_j\mathbf{w}^L_j + \tilde{\mathbf{Z}}^T_j\mathbf{w}^R_j + \text{diag}(\mathbf{X}_j\mathbf{G}\tilde{\mathbf{Z}}_j) + \epsilon$, where ϵ is a Gaussian noise term.

6.1.2 Lake Ecology Data. We also employ the LAGOS-NE lake ecology dataset [22] for our experiments. The dataset contains various lake hydrogeomorphic and land use/land cover variables for a study region spanning 17 states in the United States. Our goal is to predict lake water quality variables such as total phosphorus (tp), total nitrogen (tn), chlorophyll-a (chla), and Secchi depth (secchi), Altogether, we selected 13 local and 8 regional (group-level) features as our predictor variables. Each feature was standardized to have zero mean and unit standard deviation. Due to skewness of their values, the response variables were log-transformed before standardization. A brief summary of the data is given in Table 3.

6.2 Experimental Setup

For each dataset, we perform a nested 10-fold cross-validation for hyperparameter tuning and model evaluation.

Baseline Algorithms. We compare the performance of our framework against the following baseline algorithms:

- **global_XZ**: A global lasso regression model trained on both local and group-level features.
- **global_X**: A global lasso regression model trained on the local features only.
- **Cross-level interaction**: A two-level MLM approach with L_1 regularization.
- **Independent_lasso**: A local lasso regression model is trained for each region using only the local features.
- **Least_L21**: An MTL approach based on joint feature selection with group lasso [30, 20] using only the local features.
- **Least_Lasso**: An MTL approach based on L_1 regularization, using only the local features.
- **MTML**: An MTL approach based on Least_Lasso using the local, regional, and cross-scale interaction features.

Table 4: Experiment result for synthetic data. The first two columns measure the prediction accuracy while the last two columns evaluate the cross-scale interactions (CSIs). NA means the method does not provide an estimate of the CSIs.

method/different metric	rmse	R-square	F1_sparse(G)	Acc_G_sign
global_XZ	7.20 ± 2.36	0.05 ± 0.01	NA	NA
global_X	7.06 ± 2.27	0.01 ± 0.01	NA	NA
Cross level interaction	4.27 ± 1.02	0.52 ± 0.29	0.7 ± 0.16	0.04 ± 0.13
Independent Lasso	4.14 ± 1.26	0.59 ± 0.16	0.65 ± 0.05	0.20 ± 0.21
Least_L21	5.17 ± 1.85	0.40 ± 0.21	0.75 ± 0.11	0.53 ± 0.32
Least_Lasso	4.56 ± 1.30	0.47 ± 0.30	0.69 ± 0.08	0.32 ± 0.27
MTML	4.64 ± 0.88	0.46 ± 0.29	0.60 ± 0.10	0.01 ± 0.32
MTML_noise	4.65 ± 1.16	0.45 ± 0.27	0.40 ± 0.20	0.38 ± 0.32
MTML_noimpute	3.74 ± 0.97	0.66 ± 0.14	0.72 ± 0.12	0.49 ± 0.30
MTML_imputation	**3.31 ± 0.96**	**0.74 ± 0.09**	**0.95 ± 0.07**	**0.93 ± 0.10**

We also consider two variations of our framework: (1) **MTML_noimpute**, in which the group-level feature values are used instead of disaggregated values for model building, and (2) **MTML_noise**, in which the disaggregated values are equal to the group-level value perturbed by some random Gaussian noise. Only cross-level interaction, MTML, MTML_imputation, and its variations are designed to produce the cross-scale interaction matrix **G**. For independent lasso, least_L21, and least_lasso, we can apply a subsequent post-processing step to estimate the cross-scale interaction matrix **G**.

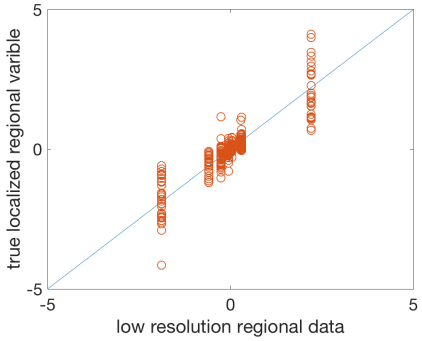
Evaluation Metrics. To assess the overall predictive performance of the various algorithms, we employ the following two metrics:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{1}{N}(y_i - y_i^{\text{pred}})^2}, \quad R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - y_i^{\text{pred}})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

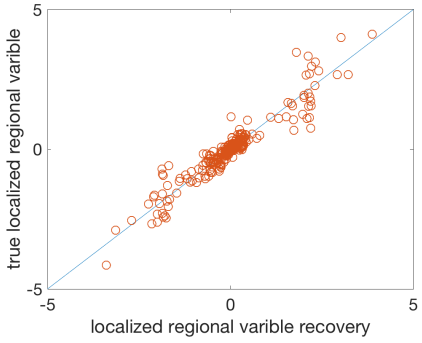
where $\bar{y}$ is the mean value of response variable. In addition, the following two metrics were used to determine whether the competing algorithms can correctly identify the sign and sparse structure of the cross-scale interaction matrix **G** when applied to the synthetic data: *1) F1_sparse*: This metric is computed based on the precision and recall values of algorithms to correctly identify the non-zero elements of the cross-scale interaction matrix **G**. *2) Acc_G_sign*: This metric is used to evaluate how well the algorithms correctly identify the sign of the elements within the matrix **G**.

6.3 Results on Synthetic Data

Table 4 shows a comparison between the performance of our proposed framework against other baseline methods on the synthetic data. First, observe that the global models have the worst predictive performance as they fail to account for spatial heterogeneity in the data. The independent lasso approach performs much better but is still worse than the cross-level interaction, MTL, and MTML approaches since the local models are trained independently without using the group-level features. The results further suggest that MTML_imputation not only outperforms all other baselines in terms of model accuracy, it also gives the best estimate of the cross-scale interaction matrix **G**. Although the traditional cross-level interaction model can detect some of the CSIs, it also generates quite significant false positives and false negatives, which leads to its lower accuracy in terms of F1_sparse and Acc_G_sign. The results also suggest that applying postprocessing to the independent lasso, least_L21, and least_lasso methods can identify the cross-scale interaction matrix **G** with comparable accuracy as the cross-level interaction model, but is still far worse than MTML_imputation.



(a) Coarse-scale group-level feature



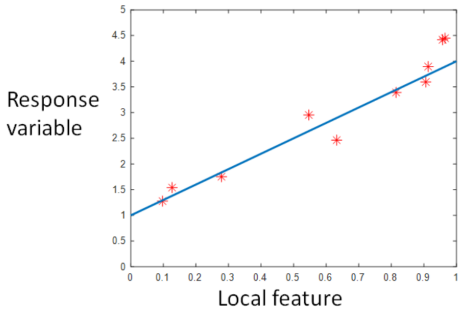
(b) Disaggregated group-level feature

Figure 3: Comparison between the disaggregated and group-level feature values. Both Y-axis corresponds to the high resolution values of the group-level features. The X-axis for the top diagram corresponds to the coarse-scale feature values, whereas the X-axis for the bottom diagram corresponds to the disaggregated values obtained by MTML_imputation. The diagonal line has slope equaling to 1, which represents perfect disaggregation.

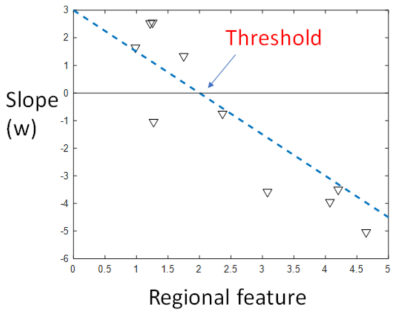
Furthermore, upon comparing the performance of different MTML approaches, we observe that MTML performs poorly since it cannot effectively utilize the group-level and the cross-level interaction features due to the rank deficiency problem illustrated in Table 2. For MTML_noimpute, which has the same objective function as our method but without the feature disaggregation step, its performance is worse than MTML_imputation, which shows the benefit of disaggregating the coarse-level feature values. The poor performance of MTML_noise also suggests that a simple disaggregation step of perturbing the group-level feature values with Gaussian noise neither improves the prediction accuracy nor our ability to detect CSIs. Finally, to evaluate the quality of the disaggregated values, Figure 3 shows the difference between using the group-level feature value to represent the true disaggregated values (top diagram) and using the disaggregated values obtained by MTML_imputation (bottom diagram). The results suggest that the disaggregated feature values of MTML_imputation are quite close to their true values.

Table 5: RMSE comparison for lake ecology data

method / response variable	tp	tn	chla	secchi
global_XZ	0.73 ± 0.03	0.65 ± 0.08	0.79 ± 0.06	0.77 ± 0.12
global_X	0.74 ± 0.03	0.68 ± 0.06	0.80 ± 0.04	0.78 ± 0.12
Cross level interaction	0.73 ± 0.03	0.61 ± 0.08	0.79 ± 0.14	0.73 ± 0.09
Independent_Lasso	7.65 ± 3.96	9.39 ± 1.99	5.81 ± 2.50	4.04 ± 2.51
Least_L21	0.65 ± 0.03	0.60 ± 0.07	0.70 ± 0.04	0.66 ± 0.04
Least_Lasso	0.66 ± 0.03	0.60 ± 0.07	0.71 ± 0.04	0.66 ± 0.04
MTML	0.79 ± 0.03	0.72 ± 0.08	0.86 ± 0.13	0.79 ± 0.07
MTML_noise	0.70 ± 0.06	0.64 ± 0.11	0.73 ± 0.03	0.77 ± 0.31
MTML_noimpute	0.68 ± 0.07	0.61 ± 0.09	0.71 ± 0.02	0.66 ± 0.03
MTML_imputation	**0.63 ± 0.03**	**0.58 ± 0.07**	**0.68 ± 0.02**	**0.64 ± 0.02**



(a) Response variable versus local feature



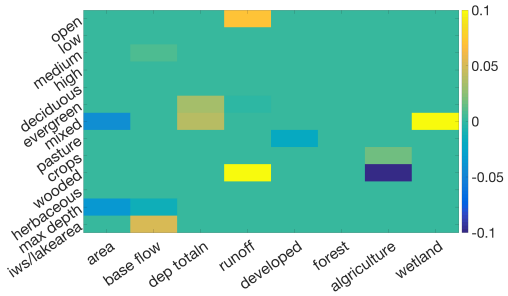
(b) Slope versus regional feature

Figure 4: An illustration of a negative CSI pattern.

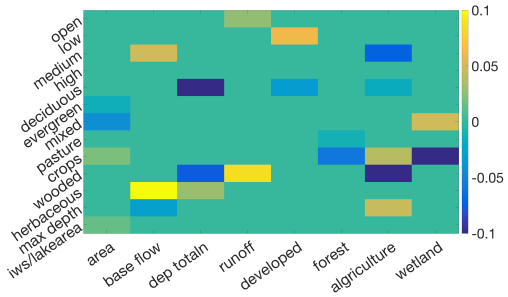
6.4 Results on Lake Ecology Data

Table 5 compares the performance of MTML_imputation against other baseline algorithms. The results show that our proposed framework outperforms the baseline algorithms for the prediction tasks of all 4 response variables (tp, tn, chla, and secchi). Unlike the synthetic data, the independent lasso model performs the worst on this dataset since the number of samples available in some regions is very small. The predictive performance of the cross level interaction model is also comparable to other global models in the data. MTML_imputation still has the lowest RMSE and highest R-square values, though the performance gains are not as high as those observed in the synthetic data. MTL approaches using only the local variables (least_L21 and least_lasso) perform better than the cross-level interaction model, though their accuracies are still lower than MTML_imputation.

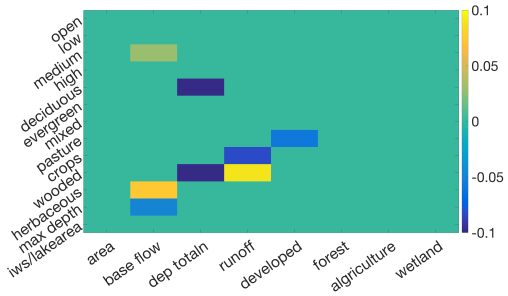
Since the true CSIs are unknown, we provide a qualitative analysis of the patterns found by the MTML_imputation algorithm. CSIs are important as they provide useful information about the coupling between the local and group-level features and their joint



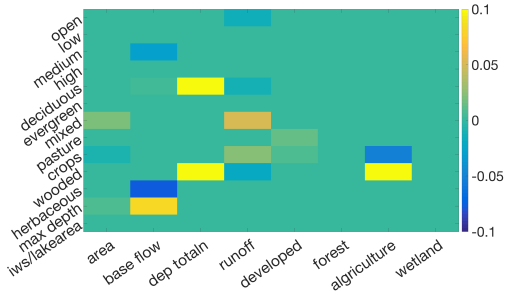
(a) CSI for TP



(b) CSI for TN



(c) CSI for Chla



(d) CSI for Secchi

Figure 5: CSI matrix (G) for TP, TN, Chla and Secchi. The x-axis corresponds to the regional features while the y-axis corresponds to the local features.

effect on the response variable. In particular, a non-zero CSI coefficient indicates the presence of a threshold effect on the nonlinear relationship between the local feature and the response variable [21, 23]. To illustrate this, Fig. 4 shows the relationship between a regional (group-level) feature and the slope of the local model relating the local feature and the response variable. A negative CSI means for some regions, there is a positive relationship between the local feature and the response variable, but for others, there is a negative relationship, depending on the regional feature. The threshold when the slope changes sign is important especially in ecology as it represents a tipping point of the system.

Figure 5 shows the CSI patterns found in this study. Some of patterns have positive signs while others are negative. An interesting negative CSI pattern found is that between local wetland (wooded) and regional agriculture (see Figure 5(a)). Such pattern has been reported in other previous studies [23, 8]. The pattern suggests a positive relationship between local wetland and TP when there is little agriculture, but changes sign when there is significant agricultural land use. The explanation here is that when there is too much agriculture, the wetlands may retain the phosphorous from getting into the lakes [31]. Other potentially interesting CSI patterns include the relationship between deciduous forest area, TN, and total nitrogen deposition of the region as well as the relationship between local cropland areas, Chla, and runoff.

7 CONCLUSION

This paper examines the problem of modeling nested spatial data using existing MLM and MTL approaches. We investigated their strengths and limitations, and showed the equivalence of their formulations under a mild condition. We also propose a novel framework called MTML_imputation to disaggregate the coarse-level group features and showed that the disaggregated values can be used to improve prediction accuracy and identification of cross-scale interactions in the data. The future works of this paper will be mainly focused on two aspects: **a)**. to extend the proposed framework from two-level to multi-level; and **b)** to extend the proposed framework from assuming mean aggregation of the regional feature values to other types of aggregation functions (e.g., max or min).

8 ACKNOWLEDGEMENT

This research was supported in part by the NSF under grant EF-1638679, IIS-1615612 and IIS-1615597. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

[1] Herman Aguinis, Ryan K Gottfredson, and Steven Andrew Culpepper. 2013. Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, 39, 6, 1490–1528.
[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2007. Multi-task feature learning. In *Advances in NIPS*, 41–48.
[3] Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imag. Sc.*, 2, 1, 183–202.
[4] R Darrell Bock. 2014. *Multilevel analysis of educational data*. Elsevier.
[5] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.
[6] Jianhui Chen, Jiayu Zhou, and Jieping Ye. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proc of KDD*. ACM, 42–50.
[7] Avital Cnaan, Nan M Laird, and Peter Slasor. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, 16, 20, 2349–2380.
[8] C Emi Fergus, Patricia A Soranno, Kendra Spence Cheruvelil, and Mary T Bremigan. 2011. Multiscale landscape and wetland drivers of lake total phosphorus and water color. *Limnology and Oceanography*, 56, 6, 2127–2146.
[9] Donald E Farrar and Robert R Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92–107.
[10] Kelly-Ann Dixon Hamil, Basil V. Iannone III, Whitney K. Huang, Songlin Fei, and Hao Zhang. 2016. Cross-scale contradictions in ecological relationships. *Landscape Ecology*, 31, 1, 7–18.
[11] Joop J Hox, Mirjam Moerbeek, and Rens van de Schoot. 2010. *Multilevel analysis: Techniques and applications*. Routledge.
[12] Michael G Kenward and James H Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.
[13] Jeffery T. Leek and Roger D. Peng. 2015. What is the question? *Science*, 347, 6228, 1314–1315.
[14] Kaixiang Lin, Jianpeng Xu, Inci M Baytas, Shuiwang Ji, and Jiayu Zhou. 2016. Multi-task feature interaction learning. In *Proc. of KDD*, 1735–1744.
[15] Jun Liu, Shuiwang Ji, and Jieping Ye. 2009. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proc. of UAI*, 339–348.
[16] Aurelie C Lozano and Grzegorz Swirszcz. 2012. Multi-level lasso for sparse multi-task regression. In *Proc of ICML*, 595–602.
[17] Cora JM Maas and Joop J Hox. 2005. Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1, 3, 86.
[18] John E Mathieu, Herman Aguinis, Steven A Culpepper, and Gilad Chen. 2012. Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, 97, 5, 951.
[19] Yurii Nesterov. 2007. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076. Universite catholique de Louvain, Center for Operations Research and Econometrics (CORE).
[20] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. 2010. Efficient and robust feature selection via joint l2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, 1813–1821.
[21] D.P.C. Peters, R.A. Pielke, B.T. Bestelmeyer, C.D. Allen, S. Munson-McGee, and K.M. Havstad. 2004. Cross-scale interactions, nonlinearities, and forecasting catastrophic events. *PNAS*, 101, 42, 15130–15135.
[22] Patricia A Soranno et al. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. *GigaScience*, 4, 1, 28.
[23] Patricia A Soranno et al. 2014. Cross-scale interactions: quantifying multi-scaled cause–effect relationships in macrosystems. *Frontiers in Ecology and the Environment*, 12, 1, 65–73.
[24] W. R. Tobler. 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46, 234–240.
[25] Paul Tseng. 2001. Convergence of a block coordinate descent method for non-differentiable minimization. *J. of Optim. Theory and Appl.*, 109, 3, 475–494.
[26] Brandon K Vaughn. 2008. Data analysis using regression and multilevel/hierarchical models. *Journal of Educational Measurement*, 45, 1, 94–97.
[27] Tyler Wagner, C Emi Fergus, Craig A Stow, Kendra S Cheruvelil, and Patricia A Soranno. 2016. The statistical power to detect cross-scale interactions at macroscales. *Ecosphere*, 7, 7.
[28] Michael R. Willig, Christopher P. Bloch, Nicholas Brokaw, Christopher Higgins, Jill Thompson, and Craig R. Zimmermann. 2007. Cross-scale responses of biodiversity to hurricane and anthropogenic disturbance in a tropical forest. *Ecosystems*, 10, 824–838.
[29] Jianpeng Xu, Pang-Ning Tan, Lifeng Luo, and Jiayu Zhou. 2016. Gspartan: a geospatio-temporal multi-task learning framework for multi-location prediction. In *Proc of SDM*, 657–665.
[30] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 1, 49–67.
[31] Shuai Yuan, Jiayu Zhou, Pang-Ning Tan, Emi Fergus, Tyler Wagner, and Patricia Soranno. 2017. Multi-level multi-task learning for modeling cross-scale interactions in nested geospatial data. In *Proc. of ICDM*, 1153–1158.
[32] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Clustered multi-task learning via alternating structure optimization. In *Advances in Neural Information Processing Systems*, 702–710.
[33] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Malsar: multi-task learning via structural regularization. *Arizona State University*, 21.
[34] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. 2011. A multi-task learning formulation for predicting disease progression. In *Proc of KDD*. ACM, 814–822.