Interpretable Time Series Autoregression for Periodicity Quantification with Robustness

Xinyu Chen, Vassilis Digalakis Jr, Lijun Ding, Dingyi Zhuang, and Jinhua Zhao

Abstract—Time series autoregression (AR) is a classical statistical tool for modeling auto-correlations and periodic structures in real-world systems. We revisit this model from an interpretable machine learning perspective by introducing sparse autoregression (SAR), where ℓ_0-norm constraints are used to isolate dominant periodicities and Huber loss function over autoregressive residuals is adopted to handle heavy-tailed noises. We reformulate the optimization of SAR models with Huber losses as an exact Mixed-Integer Quadratic Programming (MIQP) and introduce a scalable and hybrid Iteratively Reweighted Least Squares (IRLS) algorithm. In the IRLS algorithm, we present a two-stage optimization scheme, including global support set selection via MIQP and local coefficient estimation via quadratic programming. These methods enable efficient model inference on real-world spatiotemporal datasets. We validate our SAR framework on large-scale and high-dimensional mobility and climate data. On hourly NYC ridesharing data, our model reveals interpretable daily and weekly periodicities as well as the accurate prediction via the use of Huber losses and ℓ_0-norm. On monthly climate datasets, our model uncovers the evolving spatial patterns of temperature and precipitation seasonality across four decades in North America and detects global sea surface temperature dynamics, including El Niño. Together, our results demonstrate the interpretability, robustness, and scalability of SAR for periodicity pattern discovery in complex spatiotemporal systems.

Index Terms—Interpretable machine learning, time series analysis, sparse autoregression, periodicity quantification, mixed-integer quadratic programming, iteratively reweighted least squares, human mobility, climate systems

1 Introduction

Many real-world systems represented by multivariate or multidimensional time series exhibit complex spatiotemporal patterns and temporal regularities, including periodicity, seasonality, and anomalies. Understanding these patterns is essential for anticipating system behavior, identifying disruptions, and supporting operational decision-making. In dynamic systems such as urban transportation and climate, periodic patterns can shift due to policy interventions, demand changes, service disruptions, environmental variability, and extreme events, necessitating interpretable and adaptive data-driven methods. A central challenge is to automatically identify dominant periodic components from spatially- and temporally-varying systems, track their evolution over a long-term time period, and distinguish true structural changes from random variability.

Urban transportation systems display strong temporal regularity driven by commuting travels, business cycles, and travel demand. In common sense, mobility regularity patterns are often governed by daily, weekly, and multiple periodic cycles. However, such patterns are not fixed as they evolve with infrastructure changes, economic conditions, and disruptive events. For example, the COVID-19 pandemic in 2020 led to a collapse in established mobility periodicity due to lockdown and shifts to remote work. These disruptions raise key questions: *How do periodic structures evolve over time? Can we systematically quantify such changes in a temporally-varying system? How to identify periodicity patterns?* Addressing these questions is critical for planning, forecasting, and adaptive resource allocation in transportation systems.

Climate systems also exhibit periodic and seasonal patterns that shape temperature, precipitation, ocean circulation, and atmospheric dynamics. Yet these patterns evolve over time due to long-term variability and climate change. Traditional time series decomposition methods [1], [2], [3] assume fixed seasonal structure and often ignore gradual or region-specific shifts. Accurately altering spatiotemporal patterns and monitoring their changes are essential for understanding climate dynamics, anticipating extreme events, managing agriculture production, and supporting policy-making. This calls for *interpretable and data-driven models that can robustly uncover the dominant seasonal components and track their evolution across space and time.*

To address these needs, we develop a unified interpretable machine learning framework for quantifying spatially- and temporally-varying periodicity and structural shifts in real-world time series data. Our models build on classical autoregression (AR) [4], [5] but incorporate sparse structure via ℓ_0-norm induced sparsity constraints to enhance interpretability. Inspired by recent advances in sparse regression [6] and time series convolution [7], we formulate interpretable AR models capable of isolating dominant periodic patterns over space and time. Our novelty is not the introduction of sparse AR, which has extensive prior foundations, but rather the development of exact sparse AR with

- *Xinyu Chen, Dingyi Zhuang, and Jinhua Zhao are with the Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, USA (e-mail: chenxy346@gmail.com; dingyi@mit.edu; jinhua@mit.edu).*
- *Vassilis Digalakis Jr is with the Questrom School of Business, Boston University, Boston, MA 02115, USA (e-mail: vvdigalakis@gmail.com).*
- *Lijun Ding is with the Department of Mathematics at the University of California, San Diego, La Jolla, CA 92093, USA (e-mail: l2ding@ucsd.edu).*

(Corresponding author: Jinhua Zhao)

mixed-integer quadratic programming (MIQP) for large-scale spatiotemporal datasets. Overall, this work makes the following contributions:

- **Huber Sparse Autoregression (Huber-SAR)**: We introduce a robust and interpretable machine learning framework to identify dominant auto-correlations from a sequence of time lags by reformulating AR with ℓ_0-norm induced sparsity constraints. To handle the heavy-tailed noises in time series, we propose a flexible Huber-SAR model in which Huber loss function uses a squared ℓ_2-norm for small residuals but transitions to an ℓ_1-norm for large residuals. The optimization is solved by MIQP techniques, providing more accurate periodicity quantification than conventional greedy methods.
- **High-Dimensional Huber-SAR**: We propose a high-dimensional Huber-SAR model in which the support sets across different time series are enforced to be consistent, improving interpretability. To address the computational burden, we develop a hybrid Iteratively Reweighted Least Squares (IRLS) algorithm and present a two-stage optimization scheme, including global support set selection via MIQP and local coefficient estimation via quadratic programming. The two-stage optimization makes the high-dimensional Huber-SAR method tractable and efficient in the case of millions of decision variables.
- **Extensive Real-World Validation**: We demonstrate the effectiveness of our models on large-scale transportation and climate datasets. In terms of human mobility prediction, Huber-SAR outperforms both AR and ℓ_2-SAR models, highlighting the importance of ℓ_0-norm induced sparsity constraints and Huber losses. In terms of periodicity quantification, our model reveals the spatial patterns of daily/weekly periodicity in New York City (NYC) and evolving spatial patterns of yearly seasonality in North America climate variables and global sea surface temperature. In addition, climate events such as El Niño can be clearly identified from the less seasonal regions.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces notation and background on AR models. Sections 4 and 5 elaborate on the Huber-SAR and high-dimensional Huber-SAR models, respectively. Section 6 reports empirical results. Finally, we conclude this study in Section 7.

2 Literature Review

2.1 Classical Time Series Autoregression

Time series modeling is a foundational tool in transportation, climate science, econometrics, and other fields [8], [9], [10]. Classical models such as AR, moving average (MA), and their combinations (e.g., ARIMA) are widely used to capture temporal dependencies and seasonal structures in univariate time series data [4], [5], [11]. Among these, AR models remain popular due to their simplicity and interpretability. More recently, applications have demanded models that adapt to non-stationary and temporally-varying systems. Temporally-varying AR allows coefficients to change over time [12], often incorporating structural or smoothness constraints to ensure interpretability and stability [13]. AR has also been generalized to multivariate settings, such as Vector Autoregressive (VAR) models [14], [15]. These ideas have been successfully applied in finance, neuroscience, and dynamical systems.

2.2 Sparse Autoregression and Interpretability

Traditional AR models include all lagged terms up to a fixed order, which can lead to overfitting and obscure the most meaningful temporal dependencies particularly in high-order settings [16]. Sparse AR addresses this limitation by selecting a subset of informative lags and imposing sparsity on the model parameter space, thereby improving both parsimony and interpretability. LASSO-based methods [17] are widely used to induce sparsity in AR models, offering robustness in high-dimensional settings [18], [19]. Beyond unstructured sparsity, recent work has explored structured sparse AR models, including interpretable formulations for nonlinear and dynamical systems [20], [21], [22] and graph-based sparsity for slowly evolving regression coefficients [6]. Relatedly, [23] proposes a sparsity-controlled VAR framework that enables multi-dimensional sparsity tuning for interpretable causal discovery.

While these approaches offer greater flexibility and improved predictive accuracy over classical LASSO-based alternatives, they suffer from limited scalability, as solving the resulting nonconvex optimization problems becomes computationally prohibitive in high-dimensional or large-scale settings. Moreover, reliance on heuristic or approximate solvers precludes guarantees of exact support recovery and global optimality. Overall, these advances move beyond classical LASSO by incorporating domain-specific sparsity structures; however, existing methods typically do not enforce structured sparsity across time and space simultaneously, nor do they leverage exact combinatorial optimization. Our work bridges this gap by introducing a framework for structured sparse AR with exact support set control over spatiotemporal dimensions.

2.3 Robust Regression & Autoregression

Robust regression is an important statistical framework that reformulates standard regression objectives to mitigate the influence of outliers and heavy-tailed noises. While traditional models such as ordinary least squares are highly sensitive to extreme values, robust methods use alternative loss functions to maintain the integrity of the model's parameters. Most robust AR methods are direct extensions of these principles, adapting them to temporal data structures where anomalies can otherwise distort the identification of true auto-correlations. A foundational tool in robust modeling is the Huber loss function [24], which provides a compromise between the ℓ_1- and ℓ_2-norm. The loss function of quantile regression stems from ℓ_1-norm, imposing weights on the positive and negative residuals of regression [25], [26]. Comparing to ℓ_1-norm, the loss function with pth power of ℓ_p-norm is also a robust formula for mitigating the learning biases resulted from heavy-tailed noises [27], [28], [29], [30].

2.4 Exact Sparse Regression

Sparse regression via ℓ_0-norm regularization, also known as best subset selection, has long been recognized for its statistical optimality and interpretability, but was historically limited by its combinatorial complexity. Early heuristic methods include Orthogonal Matching Pursuit (OMP) and CoSaMP [31], [32], [33], while convex relaxations such as the LASSO [17] and non-convex penalties such as SCAD and MCP offered tractable alternatives [34], [35]. The solution quality of these methods often negatively impact the interpretability of sparse structures. Recent breakthroughs in optimization have made it practically feasible to solve ℓ_0-norm regularized problems exactly using mixed-integer programming (MIP). Following the seminal work of [36], several studies have scaled MIP-based sparse regression to large datasets [37], [38], [39]. These methods retain the interpretability of exact sparse models while achieving near-LASSO speed.

Our work builds directly on this line, leveraging the MIP machinery developed in [6] to solve structured sparse regression with controlled support set consistency. However, unlike prior work which focused on static or graph-based regression, we apply these methods to time series with dynamic and spatial variation, developing the first exact sparse AR framework for periodicity quantification in real-world large-scale systems.

3 Preliminaries

3.1 Notation and Definition

In Table 1, we summarize the basic symbols and notation used throughout the paper. Notably, $\mathbb{R}$ denotes the set of real numbers, and $\mathbb{Z}^+$ denotes the set of positive integers.

TABLE 1: Summary of basic notation.

Notation	Description
$x \in \mathbb{R}$	Scalar
$\boldsymbol{x} \in \mathbb{R}^n$	Vector of length n
$\boldsymbol{X} \in \mathbb{R}^{m\times n}$	Matrix of size $m \times n$
$[i]$	Integer set $\{1, 2, \ldots, i\}$, $i \in \mathbb{Z}^+$
$[i, j]$	Integer set $\{i, i+1, \ldots, j\}$, $i < j$
$\|\cdot\|_0$	ℓ_0-norm (i.e., number of nonzero entries)
$\|\cdot\|_1$	ℓ_1-norm (i.e., sum of absolute values)
$\|\cdot\|_2$	ℓ_2-norm (i.e., Euclidean norm)
$\mathrm{tr}(\cdot)$	Trace of a square matrix
$\mathrm{supp}(\cdot)$	Support set (i.e., indices of nonzero entries)
$\mathcal{N}(\cdot)$	Gaussian distribution
$\mathbb{E}[\cdot]$	Expectation
$\cup, \cap$	Union and intersection of sets, respectively

In contrast to the loss function with ℓ_2-norm, Huber loss function acts like a squared ℓ_2-norm for small residuals but transitions to an ℓ_1-norm for large residuals [24], making regression more robust than least squares in the presence of outliers or missing values. Given any residual $\varepsilon \in \mathbb{R}$, the Huber loss function is defined as follows,

$$\rho_\delta(\varepsilon) = \begin{cases} \varepsilon^2/2, & \text{if } |\varepsilon| \le \delta, \\ \delta(|\varepsilon| - \delta/2), & \text{otherwise}, \end{cases} \tag{1}$$

with the Huber threshold $\delta > 0$, see Fig. 1 for illustration, allowing one to tune the loss function's sensitivity to outliers. The ℓ_2-norm induced loss function can be simplified as $\varepsilon^2/2$ when $\delta \to \infty$ in Eq. (1).

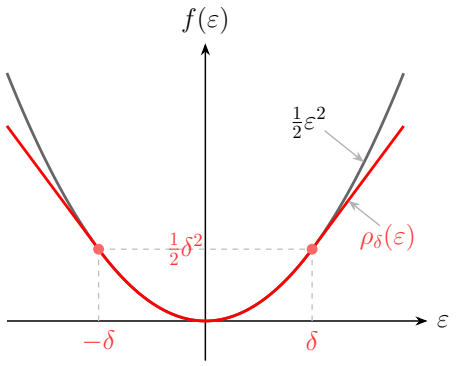


Fig. 1: Illustration of Huber loss function with threshold δ.

3.2 Time Series Autoregression

AR is a widely used statistical method for modeling temporal dependencies in time series [4], [5]. It expresses each observation of time series as a linear combination of its past values, plus Gaussian noise ε_t. The classical order-d AR model minimizes the sum of squared residuals such that

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \quad \frac{1}{2} \sum_{t\in[d+1,T]} \Big(x_t - \sum_{k\in[d]} w_k x_{t-k}\Big)^2, \tag{2}$$

where $d \in \mathbb{Z}^+$ is the AR order, leading to the notion as AR(d). The coefficient w_k in $\boldsymbol{w} = (w_1, w_2, \cdots, w_d)^\top \in \mathbb{R}^d$ captures the linear dependence between x_t and its k-lagged value x_{t-k} for some $k \in [d]$. As the residual at time t, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ typically denotes the Gaussian noise for the variance $\sigma^2 > 0$. For notational convenience, we define the design matrix $\boldsymbol{A} \in \mathbb{R}^{(T-d)\times d}$ and target vector $\tilde{\boldsymbol{x}} \in \mathbb{R}^{T-d}$ as:

$$\boldsymbol{A} \triangleq [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \cdots, \tilde{\boldsymbol{x}}_d], \quad \tilde{\boldsymbol{x}} \triangleq \{x_t\}_{t\in[d+1,T]}, \tag{3}$$

respectively. In terms of columns of the design matrix $\boldsymbol{A}$, we let $\tilde{\boldsymbol{x}}_k \triangleq \{x_{t-k}\}_{t\in[d+1,T]} \in \mathbb{R}^{T-d}$ for any $k \in [d]$. Then, Problem (2) takes the objective as an ℓ_2-norm induced loss function over $\tilde{\boldsymbol{x}} \approx \boldsymbol{A}\boldsymbol{w}$, leading to the standard least squares solution: $\hat{\boldsymbol{w}} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\tilde{\boldsymbol{x}}$, assuming $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible. In more general settings, this solution can also be expressed using the Moore-Penrose pseudoinverse (denoted by $\cdot^\dagger$) as $\hat{\boldsymbol{w}} = \boldsymbol{A}^\dagger\tilde{\boldsymbol{x}}$.

A time series $\boldsymbol{x}$ is said to exhibit periodicity with period $\Delta t \in \mathbb{Z}^+$ if $x_t \approx x_{t+\Delta t}$ for many t, typically reflected by high auto-correlation at lag Δt, i.e., $\mathrm{Cov}(x_t, x_{t+\Delta t}) \gg 0$. Seasonality refers to periodic patterns tied to calendar cycles (e.g., daily, weekly, and yearly) and can be modeled as a component s_t satisfying $s_t = s_{t+\Delta t}$. In the AR framework, strong periodicity at lag Δt manifests as a large positive coefficient $w_{\Delta t}$, allowing us to infer dominant cycles directly from the estimated sparse coefficient vector $\boldsymbol{w}$.

3.3 Sparse Autoregression

In traditional settings, autoregression with least squares typically yields dense solutions that fail to isolate dominant periodicities. This limitation motivates our transition to SAR, where we impose sparsity constraints on the coefficient vector $\boldsymbol{w}$ to identify dominant auto-correlations. The general form of SAR on the time series $\boldsymbol{x}$, with an aforementioned

data pair $\{\tilde{\boldsymbol{x}}, \boldsymbol{A}\}$, is given by $\min_{\boldsymbol{w}} \frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{w}\|_2^2 + \rho_\lambda(\boldsymbol{w})$. The use of ℓ_0-norm results in the penalty $\rho_\lambda(\boldsymbol{w}) = \frac{\lambda^2}{2}\mathbb{1}\{\boldsymbol{w} \neq 0\}$ [34] where $\mathbb{1}\{\cdot\}$ is the indicator function. The penalty $\rho_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$ with ℓ_1-norm refers to the LASSO [17], [18], [20], allowing a direct extension from univariate autoregression to vector autoregression with group Lasso [40], [41]. Other penalties such as SCAD and MCP have also been leveraged to time series [34], [42].

4 Huber Sparse Autoregression

This section introduces the Huber Sparse Autoregression (Huber-SAR) model with ℓ_0-norm induced sparsity constraints for identifying dominant auto-correlations in time series. We first describe the Huber-SAR model with MIQP reformulation. Then, we present the IRLS method as an efficient way to solve the resulting optimization problem.

4.1 Model Formulation

Huber loss function with a given threshold $\delta > 0$ in Eq. (1) uses a linear form for large residuals and squared form for relatively small residuals, making it less sensitive to heavy-tailed noises. This motivates us to reformulate the order-d AR model with Huber loss function. To identify the dominant auto-correlations (e.g., periodicity and seasonality), we introduce an ℓ_0-norm of the coefficient vector $\boldsymbol{w} \in \mathbb{R}^d$, which is defined as the number of nonzero entries, i.e., $\|\boldsymbol{w}\|_0 = |\operatorname{supp}(\boldsymbol{w})| \leq d$. In particular, $|\operatorname{supp}(\boldsymbol{w})| = \sum_{k\in[d]} \mathbb{1}\{w_k \neq 0\}$ is the cardinality of support set $\operatorname{supp}(\boldsymbol{w})$. As a result, we propose a Huber-SAR model in which sparsity and non-negativity constraints are imposed on $\boldsymbol{w}$, yielding the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \sum_{t\in[d+1,T]} \rho_\delta(\varepsilon_t) \\ \text{s.t.} \quad & \varepsilon_t = x_t - \sum_{k\in[d]} w_k x_{t-k}, \quad \forall t \in [d+1, T], \\ & 0 \leq \boldsymbol{w} \leq \mathcal{M} \cdot \mathbb{1}_d, \quad \|\boldsymbol{w}\|_0 \leq \tau, \end{aligned} \tag{4}$$

where $\mathcal{M} \in \mathbb{R}_+$ is a sufficiently large value as the upper bound of $\boldsymbol{w}$ with $\mathbb{1}_d$ denoting the vector of d ones, and $\tau \in \mathbb{Z}^+$ controls the maximum number of nonzero coefficients. The non-negativity constraint encourages interpretability by focusing on positive auto-correlations, as is typical for periodicity in time series. The objective function is written as the sum of Huber loss functions over residuals $\varepsilon_t, t \in [d+1, T]$.

However, Problem (4) is nonconvex due to the ℓ_0-norm induced sparsity constraint and is generally NP-hard to solve exactly. In principle, enumerating all support sets of cardinality τ would incur a combinatorial cost of $\mathcal{O}(d^\tau)$, making exact search infeasible for large d or τ. Modern MIQP solvers bypass this challenge using branch-and-bound techniques and cutting-plane methods (e.g., [6]), which allow them to find globally optimal solutions efficiently in practice. Compared to greedy methods such as subspace pursuit [33], MIQP yields higher-quality solutions with provable optimality guarantees. The use of binary variables enables precise support set control, which is essential for model interpretability in settings where dominant lag selection matters.

4.2 Mixed-Integer Quadratic Programming

Due to the combinatorial nature of the ℓ_0-norm induced constraint, Problem (4) with Huber loss function cannot be directly solved MIQP solvers. However, we can equivalently express the problem as an MIQP problem by introducing binary variables that encode the support set of $\boldsymbol{w}$. Let $\boldsymbol{z} \in \{0,1\}^d$ be a binary vector, where $z_k = 1$ indicates that w_k is allowed to be nonzero. We rewrite Problem (4) as:

$$\begin{aligned} \min_{\boldsymbol{w},\, \boldsymbol{z}\in\mathcal{C}} \quad & \sum_{t\in[d+1,T]} \rho_\delta(\varepsilon_t) \\ \text{s.t.} \quad & \varepsilon_t = x_t - \sum_{k\in[d]} w_k x_{t-k}, \quad \forall t \in [d+1, T], \\ & 0 \leq w_k \leq \mathcal{M} \cdot z_k, \qquad\qquad \forall k \in [d], \end{aligned} \tag{5}$$

where we define the feasible set

$$\mathcal{C} = \left\{ \boldsymbol{z} \in \{0,1\}^d \;\middle|\; \|\boldsymbol{z}\|_1 = \sum_{k\in[d]} z_k \leq \tau \right\} \subseteq \mathbb{R}^d, \tag{6}$$

related to the decision vector $\boldsymbol{z}$. The binary support constraint $\|\boldsymbol{z}\|_1 \leq \tau$ ensures that at most τ coefficients in $\boldsymbol{w}$ are nonzero.

In theory, the objective function with Huber loss function is equivalent to a quadratic programming [24]. Thus, the optimization problem in Eq. (5) is indeed an MIQP, which can be reformulated as follows,

$$\begin{aligned} \min_{\boldsymbol{w},\, \boldsymbol{z}\in\mathcal{C},\, \boldsymbol{\alpha},\, \boldsymbol{\beta}} \quad & \frac{1}{2} \sum_{t\in[d+1,T]} (\beta_t^2 + 2\delta\alpha_t) \\ \text{s.t.} \quad & \varepsilon_t = x_t - \sum_{k\in[d]} w_k x_{t-k}, \quad \forall t \in [d+1, T], \\ & -\alpha_t \leq \varepsilon_t - \beta_t \leq \alpha_t, \quad \forall t \in [d+1, T], \\ & \alpha_t \geq 0, \qquad\qquad \forall t \in [d+1, T], \\ & 0 \leq w_k \leq \mathcal{M} \cdot z_k, \qquad \forall k \in [d], \end{aligned} \tag{7}$$

or equivalently,

$$\begin{aligned} \min_{\boldsymbol{w},\, \boldsymbol{z},\, \boldsymbol{\alpha},\, \boldsymbol{\beta}} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \delta\|\boldsymbol{\alpha}\|_1 \\ \text{s.t.} \quad & \boldsymbol{\varepsilon} = \tilde{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{w}, \\ & -\boldsymbol{\alpha} \leq \boldsymbol{\varepsilon} - \boldsymbol{\beta} \leq \boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \geq 0, \\ & 0 \leq \boldsymbol{w} \leq \mathcal{M} \cdot \boldsymbol{z}, \qquad \boldsymbol{z} \in \mathcal{C}, \end{aligned} \tag{8}$$

where $\boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathbb{R}^{T-d}$ are slack vectors, which capture the quadratic and linear regions, respectively. The MIQP solver could optimize both slack vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously, and the linear constraints $-\boldsymbol{\alpha} \leq \boldsymbol{\varepsilon} - \boldsymbol{\beta} \leq \boldsymbol{\alpha}$ ensure the solver finds the optimal split.

4.3 Hybrid IRLS Algorithm

Instead of solving one massive MIQP problem, we develop a hybrid IRLS algorithm, which breaks the original MIQP into a sequence of simpler sub-problems. The use of IRLS can improve efficiency dramatically in the case of thousands of decision variables. Specifically, IRLS is used to solve certain optimization problems by an iterative method in which each

step involves solving a *weighted least squares* or *weighted SAR* problem with linear constraints of the MIQP form:

$$\begin{aligned}\{\boldsymbol{w}^{(\kappa)}, \boldsymbol{z}^{(\kappa)}\} = \underset{\boldsymbol{w},\,\boldsymbol{z}}{\arg\min} \quad & \frac{1}{2}\sum_{t\in[d+1,T]} \vartheta_t^{(\kappa)}\varepsilon_t^2 \\ \text{s.t.} \quad & \boldsymbol{\varepsilon} = \tilde{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{w}, \\ & 0 \le \boldsymbol{w} \le \mathcal{M}\cdot\boldsymbol{z},\ \boldsymbol{z}\in\mathcal{C},\end{aligned} \tag{9}$$

with the *weights* such that

$$\vartheta_t^{(\kappa)} = \min\{1, \delta/|\varepsilon_t^{(\kappa)}|\}, \tag{10}$$

for all $t \in [d+1, T]$, referring to Lemma 1.

Lemma 1. *For the Huber loss function $\rho_\delta(\varepsilon)$ on the residual ε, the IRLS method assumes the consistency between the first-order derivatives of weighted least squares and Huber loss function [29]:*

$$\frac{\mathrm{d}\,(\vartheta\varepsilon^2/2)}{\mathrm{d}\,\varepsilon} = \frac{\mathrm{d}\,\rho_\delta(\varepsilon)}{\mathrm{d}\,\varepsilon} \tag{11}$$

$$\Rightarrow \quad \vartheta\cdot\varepsilon = \begin{cases}\varepsilon, & \text{if } |\varepsilon|\le\delta,\\ \delta\cdot\operatorname{sgn}(\varepsilon), & \text{otherwise,}\end{cases} \tag{12}$$

with the sign function $\operatorname{sgn}(\varepsilon) = \varepsilon/|\varepsilon|$. Thus, the weight is

$$\vartheta = \min\{1, \delta/|\varepsilon|\} = \begin{cases}1, & \text{if } |\varepsilon|\le\delta,\\ \delta/|\varepsilon|, & \text{otherwise.}\end{cases} \tag{13}$$

In each step κ, we solve the weighted SAR in Eq. (9) for decision vectors $\boldsymbol{w}$ and $\boldsymbol{z}$, avoiding the extra slack variables required in Eq. (8). When handling a long time series with a large T and a relatively small $d \ll T$, MIQP solver can provide an efficient implementation and high-quality sparse solution that rigorously respects $0 \le \boldsymbol{w} \le \mathcal{M}\cdot\boldsymbol{z}$. Algorithm 1 shows the implementation of Huber-SAR with a hybrid IRLS method. The initialization with uniform weights $\vartheta_t^{(0)} = 1, \forall t$ implies that the optimization at the first step is actually an SAR with ℓ_2-norm losses.

Algorithm 1 Huber-SAR with IRLS

1: **Input:** Time series $\boldsymbol{x} \in \mathbb{R}^T$; AR order $d \in \mathbb{Z}^+$; sparsity level $\tau \in \mathbb{Z}^+$; Huber threshold $\delta > 0$.
2: Initialize the iteration $\kappa = 0$ and the weights as $\vartheta_t^{(0)} = 1$ for any $t \in [d+1, T]$.
3: **while** not converged **do**
4: Optimize the coefficient vector $\boldsymbol{w}^{(\kappa)}$ and the binary vector $\boldsymbol{z}^{(\kappa)}$ in Eq. (9) via MIQP solver.
5: Update $\kappa := \kappa + 1$.
6: Update the weights $\vartheta_t^{(\kappa)}$.
7: **end while**
8: Return coefficient vector $\boldsymbol{w}$.

5 High-Dimensional Huber Sparse Autoregression

This section introduces a unified framework for robust SAR on high-dimensional time series, designed to isolate dominant periodicities while remaining resilient to outliers. We develop the high-dimensional Huber-SAR model, as a generalization of the aforementioned Huber-SAR. To solve the MIQP problem efficiently, we present a hybrid IRLS algorithm for optimizing Huber-based autoregressive errors across different time series. To accelerate the MIQP, we develop a two-stage optimization scheme that separates the learning processes of the global support set and individual coefficient vectors.

5.1 Model Formulation

On the high-dimensional time series $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n\in[N]} \in \mathbb{R}^{N\times T}$ with N variables and T time steps, we model each time series by an independent AR(d) process with coefficient vector $\boldsymbol{w}_n \in \mathbb{R}^d$. The high-dimensional AR(d) is given by

$$x_{n,t} = \sum_{k\in[d]} w_{n,k}x_{n,t-k} + \varepsilon_{n,t},\ \forall n\in[N],\ t\in[d+1,T]. \tag{14}$$

To ensure temporal consistency, we impose the following structural constraints: i) each $\boldsymbol{w}_n$ is sparse and non-negative, and ii) all $\boldsymbol{w}_n$, $n \in [N]$ share the same support set. The optimization can be formulated as follows,

$$\begin{aligned}\min_{\{\boldsymbol{w}_n\}_{n\in[N]}} \quad & \sum_{n\in[N]}\sum_{t\in[d+1,T]} \rho_\delta(\varepsilon_{n,t}) \\ \text{s.t.} \quad & \varepsilon_{n,t} = x_{n,t} - \sum_{k\in[d]} w_{n,k}x_{n,t-k}, && \forall n,t, \\ & 0 \le \boldsymbol{w}_n \le \mathcal{M}\cdot\mathbb{1}_d, \quad \|\boldsymbol{w}_n\|_0 \le \tau, && \forall n, \\ & \operatorname{supp}(\boldsymbol{w}_n) = \operatorname{supp}(\boldsymbol{w}_{n+1}), && \forall n,\end{aligned} \tag{15}$$

where the autoregressive errors are parameterized by Huber threshold $\delta > 0$. In the constraint, $\tau \in \mathbb{Z}^+$ controls sparsity.

Remark 1. Imposing a common support set across all N time series, $\operatorname{supp}(\boldsymbol{w}_1) = \operatorname{supp}(\boldsymbol{w}_2) = \cdots = \operatorname{supp}(\boldsymbol{w}_N)$, introduces temporal consistency by ensuring that the dominant lag structure remains invariant. This refers to a zero-variation instance of the sparsely varying support model proposed in [6], in which consecutive support sets are allowed to show only within a bounded symmetric difference:

$$\begin{aligned}&|\operatorname{supp}(\boldsymbol{w}_n)\Delta\operatorname{supp}(\boldsymbol{w}_{n+1})| \\ =&|\operatorname{supp}(\boldsymbol{w}_n)\cup\operatorname{supp}(\boldsymbol{w}_{n+1})| \\ &-|\operatorname{supp}(\boldsymbol{w}_n)\cap\operatorname{supp}(\boldsymbol{w}_{n+1})| \le \tilde{\tau}, \quad \forall n\in[N-1],\end{aligned} \tag{16}$$

where Δ denotes the symmetric set difference. Setting $\tilde{\tau} = 0$ recovers the shared-support formulation used in this work. In this sense, our temporal regularity assumption corresponds to the most restrictive and interpretable form of the general slowly-varying sparsity framework.

To reformulate Eq. (15) as an MIQP problem, we again introduce a binary decision vector $\boldsymbol{z} \in \{0,1\}^d$ to represent the global support set in the constraint. The objective is converted into a quadratic form with slack vectors $\{\boldsymbol{\alpha}_n, \boldsymbol{\beta}_n\}_{n\in[N]}$. The resulting MIQP is given by

$$\begin{aligned}\min_{\{\boldsymbol{w}_n,\,\boldsymbol{\alpha}_n,\,\boldsymbol{\beta}_n\}_{n\in[N]},\,\boldsymbol{z}} \quad & \frac{1}{2}\sum_{n\in[N]}\left(\|\boldsymbol{\beta}_n\|_2^2 + 2\|\boldsymbol{\alpha}_n\|_1\right) \\ \text{s.t.} \quad & \boldsymbol{\varepsilon}_n = \tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}_n, && \forall n, \\ & -\boldsymbol{\alpha}_n \le \boldsymbol{\varepsilon}_n - \boldsymbol{\beta}_n \le \boldsymbol{\alpha}_n, && \forall n, \\ & \boldsymbol{\alpha}_n \ge 0, && \forall n, \\ & 0 \le \boldsymbol{w}_n \le \mathcal{M}\cdot\boldsymbol{z}, \quad \boldsymbol{z}\in\mathcal{C}, && \forall n,\end{aligned} \tag{17}$$

where the binary variable $z_k = 1$ if lag k is selected, enforcing a common support set across all time series.

5.2 Hybrid IRLS Algorithm

The MIQP formulated in Eq. (17) is computationally challenging due to the simultaneous presence of discrete binary variables and robust Huber losses. To address this problem, we propose a hybrid IRLS algorithm, which separates the support set selection from the robust weight estimation (i.e., identifying which observations are outliers). This approach decouples the optimization from the combinatorial constraints by iteratively approximating the Huber loss as a weighted least squares. Furthermore, we can leverage high-performance MIQP solvers to navigate the discrete ℓ_0-norm related search space while simultaneously adapting to the noise profile of the time series.

The algorithm initializes with uniform weights, $\vartheta_{n,t}^{(0)} = 1, \forall n, t$, and proceeds through the following steps for iteration $\kappa + 1$. The choice of this initialization strategy refers to $\delta \to \infty$, which is the optimization of ℓ_2-norm losses. The procedure of the hybrid IRLS algorithm is summarized as:

- **Coefficient estimation**: We solve for the sparse coefficient vectors $\{\boldsymbol{w}_n\}_{n\in[N]}$ and the shared binary vector $\boldsymbol{z}$ by optimizing the *weighted SAR* such that

$$\begin{aligned} \min_{\{\boldsymbol{w}_n\}_{n\in[N]},\boldsymbol{z}} \quad & \frac{1}{2}\sum_{n\in[N]}\sum_{t\in[d+1,T]} \vartheta_{n,t}^{(\kappa)}\varepsilon_{n,t}^2 \\ \text{s.t.} \quad & \boldsymbol{\varepsilon}_n = \tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}_n, \qquad \forall n, \\ & 0 \le \boldsymbol{w}_n \le \mathcal{M}\cdot\boldsymbol{z}, \quad \boldsymbol{z}\in\mathcal{C}, \quad \forall n. \end{aligned} \tag{18}$$

- **Weight estimation**: Following the optimization of the coefficients, we update the weights $\vartheta_{n,t}$ based on the new residuals $\varepsilon_{n,t}^{(\kappa+1)}$. This step ensures robustness against outliers by down-weighting residuals that exceed the threshold δ (see Lemma 1):

$$\vartheta_{n,t}^{(\kappa+1)} = \min\{1, \delta/|\varepsilon_{n,t}^{(\kappa+1)}|\}. \tag{19}$$

5.3 Acceleration via Global Support Set Estimation

When dealing with large systems such as climate datasets covering thousands of grid cells across multiple decades, optimizing a high-dimensional Huber-SAR model for these time series becomes computationally infeasible. However, these time series often share common underlying periodicity and auto-correlation patterns. To exploit this structure, we first estimate a global Huber-SAR structure by fitting a single sparse coefficient vector $\boldsymbol{w} \in \mathbb{R}^d$ across all N time series. The weighted SAR in Eq. (18) can be simplified as:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{z}} \quad & \sum_{n\in[N]} (\tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w})^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})(\tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}) \\ \text{s.t.} \quad & 0 \le \boldsymbol{w} \le \mathcal{M}\cdot\boldsymbol{z}, \quad \boldsymbol{z}\in\mathcal{C}, \end{aligned} \tag{20}$$

where we let the objective function be $f(\boldsymbol{w})$. The weights of IRLS are $\boldsymbol{\vartheta}_n^{(\kappa)} = \{\vartheta_{n,t}^{(\kappa)}\}_{t\in[d+1,T]} \in \mathbb{R}^{T-d}$, which are updated by Eq. (19) with the autoregressive residuals such that

$$\boldsymbol{\varepsilon}_n^{(\kappa)} = \tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}^{(\kappa)}, \quad \forall n \in [N]. \tag{21}$$

The global sparsity pattern can be reused to identify a shared support set that generalizes across different time series. According to the property of matrix trace, we have

$$\begin{aligned} f(\boldsymbol{w}) &= \sum_{n\in[N]} (\tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w})^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})(\tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}) \\ &= \sum_{n\in[N]} \boldsymbol{w}^\top\boldsymbol{A}_n^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})\boldsymbol{A}_n\boldsymbol{w} \\ &\qquad - 2\sum_{n\in[N]} \boldsymbol{w}^\top\boldsymbol{A}_n^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})\tilde{\boldsymbol{x}}_n + C \\ &= \sum_{n\in[N]} \operatorname{tr}\left(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{A}_n^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})\boldsymbol{A}_n\right) \\ &\qquad - 2\sum_{n\in[N]} \boldsymbol{w}^\top\boldsymbol{A}_n^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})\tilde{\boldsymbol{x}}_n + C \\ &= \operatorname{tr}(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{\Phi}) - 2\boldsymbol{w}^\top\boldsymbol{\psi} + C, \end{aligned} \tag{22}$$

where $C \in \mathbb{R}$ is the constant term. In particular, we define matrix $\boldsymbol{\Phi} \in \mathbb{R}^{d\times d}$ and vector $\boldsymbol{\psi} \in \mathbb{R}^d$ such that

$$\begin{aligned} \boldsymbol{\Phi} &\triangleq \sum_{n\in[N]} \boldsymbol{A}_n^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})\boldsymbol{A}_n, \\ \boldsymbol{\psi} &\triangleq \sum_{n\in[N]} \boldsymbol{A}_n^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})\tilde{\boldsymbol{x}}_n, \end{aligned} \tag{23}$$

to rewrite the objective function $f(\boldsymbol{w})$. Following that form, we encode the sparsity using binary vector $\boldsymbol{z} \in \{0,1\}^d$ and convert Problem (20) into

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{z}} \quad & \operatorname{tr}(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{\Phi}) - 2\boldsymbol{w}^\top\boldsymbol{\psi} \\ \text{s.t.} \quad & 0 \le \boldsymbol{w} \le \mathcal{M}\cdot\boldsymbol{z}, \quad \boldsymbol{z}\in\mathcal{C}. \end{aligned} \tag{24}$$

Thus, the resulting global support set is given by

$$\Omega := \{k \in [d] \mid w_k > 0\} = \operatorname{supp}(\boldsymbol{w}) = \operatorname{supp}(\boldsymbol{z}). \tag{25}$$

In practice, learning the global support set is advantageous because i) the underlying periodicity structure can be quantified across N time series, and ii) the dominant indices of auto-correlations can be identified and estimated efficiently without having to find them explicitly for each individual coefficient vector.

5.4 Local Coefficient Estimation

To learn the sparse coefficient vectors $\{\boldsymbol{w}_n\}_{n\in[N]}$ within the given support set Ω in Eq. (25), the optimization objective is simplified as a non-negative weighted least squares problem. Given the weight vector $\boldsymbol{\vartheta}_n^{(\kappa)}$ at the κth IRLS iteration, the problem for each n is given by

$$\begin{aligned} \min_{\boldsymbol{w}_n} \quad & (\tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}_n)^\top \operatorname{diag}(\boldsymbol{\vartheta}_n^{(\kappa)})(\tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}_n) \\ \text{s.t.} \quad & \mathcal{P}_\Omega(\boldsymbol{w}_n) \ge 0, \quad \mathcal{P}_\Omega^\perp(\boldsymbol{w}_n) = 0, \end{aligned} \tag{26}$$

for all $n \in [N]$. The weights $\boldsymbol{\vartheta}_n^{(\kappa)}$ are updated by Eq. (19) with the autoregressive residuals such that

$$\boldsymbol{\varepsilon}_n^{(\kappa)} = \tilde{\boldsymbol{x}}_n - \boldsymbol{A}_n\boldsymbol{w}_n^{(\kappa)}, \quad \forall n \in [N]. \tag{27}$$

To solve Problem (26) efficiently, we employ orthogonal projections to isolate the active support: $\mathcal{P}_\Omega(\cdot)$ denotes the orthogonal projection supported on Ω. For any vector $\boldsymbol{w} \in \mathbb{R}^d$ with entries $\{w_k\}_{k\in[d]}$, the orthogonal projection takes $[\mathcal{P}_\Omega(\boldsymbol{w})]_k = w_k$ if $k \in \Omega$; otherwise, $[\mathcal{P}_\Omega^\perp(\boldsymbol{w})]_k = 0$

for any $k \notin \Omega$ in the complement of Ω. By applying these projections, we can simplify Eq. (26) into a standard constrained ℓ_2-norm minimization:

$$\min_{\mathcal{P}_\Omega(\boldsymbol{w}_n)\geq 0} \quad \left\| \operatorname{diag}\left(\sqrt{\boldsymbol{\vartheta}_n^{(\kappa)}}\right) \left(\tilde{\boldsymbol{x}}_n - \mathcal{P}_\Omega^{\text{col}}(\boldsymbol{A}_n)\mathcal{P}_\Omega(\boldsymbol{w}_n)\right) \right\|_2^2 \tag{28}$$

where $\mathcal{P}_\Omega^{\text{col}}(\boldsymbol{A}_n)$ stands for the column-wise projection of the design matrix $\boldsymbol{A}_n$ onto the support Ω. This reformulated problem is solved using the active set method, which is particularly advantageous here as it handles the non-negativity constraints by iteratively identifying the active boundary constraints (where $w_k = 0$), ensuring a high-precision solution within the subspace restricted by Ω.

5.5 Algorithmic Implementation

Algorithm 2 shows the implementation of high-dimensional Huber-SAR, which integrates the hybrid IRLS method with a non-negative active set solver to achieve robust estimation in the presence of outliers. The algorithm takes the design matrix $\boldsymbol{A}_n$ and target vector $\tilde{\boldsymbol{x}}_n$ as inputs. It immediately restricts the search space to the active columns of $\boldsymbol{A}_n$ corresponding to the support Ω. The robust weights $\vartheta_{n,t}$ are initialized to unity, effectively starting the process with a standard non-negative least squares objective. By decoupling the support selection from the weight refinement, the method maintains the global optimality of the ℓ_0-norm constraint while benefiting from the speed of active set methods for the final non-negative estimation.

Algorithm 2 High-Dimensional Huber-SAR

Input: Multivariate time series $\boldsymbol{X} \in \mathbb{R}^{N\times T}$; AR order $d \in \mathbb{Z}^+$; sparsity level $\tau \in \mathbb{Z}^+$; Huber threshold $\delta > 0$.
Initialize the iteration $\kappa = 0$ and the weights as $\vartheta_{n,t}^{(0)} = 1$ for any $n \in [N]$ and $t \in [d+1, T]$.
while not converged **do**
  Compute $\boldsymbol{\Phi}$ and $\boldsymbol{\psi}$ in Eq. (23).
  Optimize the global parameters $\boldsymbol{w}^{(\kappa)}$ and $\boldsymbol{z}^{(\kappa)}$ in Eq. (24) via MIQP solver.
  Update the weights $\vartheta_{n,t}^{(\kappa)}$.
  Update $\kappa := \kappa + 1$.
end while
Update the global support set Ω by Eq. (25).
Initialize the iteration $\kappa = 0$ and the weights as $\vartheta_{n,t}^{(0)} = 1$ for any $n \in [N]$ and $t \in [d+1, T]$.
while not converged **do**
  Update the coefficient vectors $\boldsymbol{w}_n^{(\kappa)}$ by Eq. (28).
  Update the weights $\vartheta_{n,t}^{(\kappa)}$.
  Update $\kappa := \kappa + 1$.
end while
Return coefficient vectors $\boldsymbol{w}_n$, $n \in [N]$.

6 Experiments

In this section, we evaluate the proposed Huber-SAR models through extensive experiments on real-world time series datasets. First, we analyze temporal patterns and predictability of hourly ridesharing trips in NYC across years. Second, we study large-scale climate datasets, including North America atmospheric variables and global sea surface temperatures. These datasets are both spatially and temporally structured, allowing one to reveal the geographical distribution of seasonal effects and track shifts in climate patterns over long time periods.

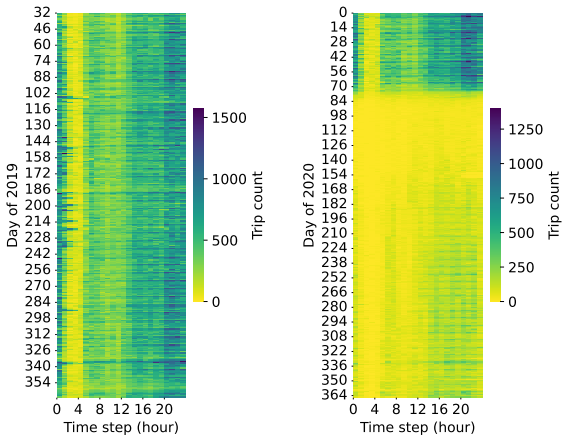
6.1 Human Mobility Periodicity

6.1.1 Dataset

This case study focuses on identifying and tracking periodic patterns in human mobility using ridesharing data in NYC from February 2019 to December 2020.[1] The original ridesharing trip records are collected with pickup/dropoff areas (i.e., 265 spatial areas in total) and time information. By aggregating these trip records with hourly resolution, we construct the trip time series across different spatial areas and hourly time steps. Due to daily and weekly commuting cycles and disruptions such as COVID-19, we demonstrate how the proposed Huber-SAR model captures the evolving temporal structure.

6.1.2 Motivation for SAR

We apply Huber-SAR in Algorithm 1 to NYC ridesharing data, focusing on human mobility prediction at John F. Kennedy (JFK) International Airport. The task is forecasting the ridesharing pickup trip time series of 2020 at JFK with the one-step prediction window. The training set for the autoregressive models is the pickup trip time series of 2019. To make an efficient prediction on these time series data, we remove the non-negativity constraint and relax the linear constraint in Eq. (4) as $-\mathcal{M} \cdot \mathbb{1}_d \leq \boldsymbol{w} \leq \mathcal{M} \cdot \mathbb{1}_d$ in SAR models, appropriately including both positive and negative auto-correlations of the given time series.



(a) 2019 (b) 2020

Fig. 2: Hourly time series of the ridesharing pickup trips at JFK from February 2019 to December 2020.

Fig. 3 demonstrates the prediction performance of AR, ℓ_2-SAR, and Huber-SAR models. The AR model is a special case of ℓ_2-SAR without any sparsity constraints on the coefficient vector $\boldsymbol{w}$, while ℓ_2-SAR is a special case of

1. The NYC ridesharing trip record dataset is available at https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

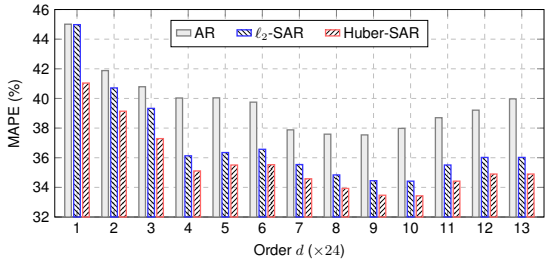


Fig. 3: Prediction performance of AR, ℓ_2-SAR, and Huber-SAR on the human mobility dataset. The different orders $d \in \{24, 48, \ldots, 312\}$ are used to test the model's robustness.

Huber-SAR when the Huber threshold $\delta \to \infty$ (i.e., a sufficiently large threshold). The comparison between AR and ℓ_2-SAR clearly shows the importance of ℓ_0-norm induced sparsity constraints. An appropriate sparsity level such as $\tau = 10$ in this case leads to a better temporal structure and more accurate predictions. With the increase of orders, e.g., $d \in \{24, 2 \times 24, \ldots, 9 \times 24\}$, the prediction errors of both AR and ℓ_2-SAR are reduced. However, the error decrease of ℓ_2-SAR is more remarkable than AR. When the order $d > 9 \times 24$, the prediction error of AR increases significantly, as ℓ_2-SAR has a marginal error increase. Time series prediction between Huber-SAR and ℓ_2-SAR with the same sparsity level $\tau = 10$ reveals some key results: i) The overall prediction performance of Huber-SAR is superior than ℓ_2-SAR mainly because of the robust Huber loss function; ii) When the order is relatively small, e.g., $d = 24$, AR and ℓ_2-SAR have the same level of prediction errors, but Huber-SAR outperforms both models significantly, verifying the importance of Huber loss function beyond ℓ_0-norm induced sparsity constraints; iii) Huber-SAR keeps the stable error decreases over ℓ_2-SAR when $d > 3 \times 24$, demonstrating the superiority of Huber loss function against ℓ_2-norm for characterizing non-Gaussian errors.

Table 2 summarizes the performance of human mobility prediction with different models of various parameters. Setting $\tau = d$ implies that no sparsity constraint would be imposed on AR models. Thus, Huber-SAR with without any sparsity constraint is indeed the Huber autoregressive (Huber-AR) model, and ℓ_2-SAR leads to the AR model. The prediction performance of AR is not as good as ℓ_2-SAR with $\tau = 10, 15$, and these results are consistent with Fig. 3. Huber-AR outperforms the standard AR model, but it is inferior than Huber-SAR with $\tau = 10, 15$. By setting the sparsity level $\tau \in \{5, 10, 15\}$, ℓ_2-SAR and Huber-SAR models perform best when $\tau = 10$, where the support set in Huber-SAR is $\text{supp}(\boldsymbol{w}) = \{1, 2, 24, 26, 48, 49, 72, 74, 167, 168\}$ (i.e., index set of auto-correlations), including hourly, daily, two-day, three-day, and weekly cycles. The Huber-SAR with threshold $\delta \in \{1, 5, 10, 15, 20\}$ always performs better than ℓ_2-SAR. To summarize, Huber-SAR with an appropriate sparsity level achieves better prediction performance than AR, Huber-AR, and ℓ_2-SAR.

6.1.3 Group Sparsity of Periodicities

To capture overlapping daily and weekly periodicities, we replace the sparsity constraint $\|\boldsymbol{w}\|_0 \leq \tau$ with a hierarchical

TABLE 2: Prediction performance (in MAPE) of ℓ_2-SAR and Huber-SAR models on the human mobility dataset. The order of autoregressive models is set as $d = 168$.

Model	$\tau = d$	$\tau = 15$	$\tau = 10$	$\tau = 5$
ℓ_2-SAR	37.88%	35.73%	35.54%	38.01%
ℓ_2-SAR (time-varying)	37.76%	36.32%	35.38%	37.39%
Huber-SAR, $\delta = 1$	36.92%	34.70%	**34.58%**	36.98%
Huber-SAR, $\delta = 5$	36.86%	34.67%	34.60%	36.92%
Huber-SAR, $\delta = 10$	36.81%	34.64%	34.61%	36.91%
Huber-SAR, $\delta = 15$	36.75%	34.63%	34.62%	36.92%
Huber-SAR, $\delta = 20$	36.73%	34.62%	34.63%	36.93%
ARIMA$(1, 2, 1)$	58.20%	-	-	-

structure. We let $\mathcal{Q}$ be a predefined set of candidate time lags, including periodic cycles such as $\mathcal{Q} = \{24, 168\}$ for hourly mobility data. To account for near-seasonal variance (e.g., time lags 167 and 168), we define a windowed group with radius $r \in \mathbb{Z}^+$:

$$G_q \triangleq \{q - r, \ldots, q + r\} \cap [d], \quad \forall q \in \mathcal{Q}, \tag{29}$$

where the local group is $G_1 = \{1, \ldots, r\}$, capturing short-term auto-correlations. To govern how many distinct periodic cycles the model is allowed to extract, we enforce the hard constraint such that

$$\sum_{k \in G_q} z_k \leq K, \quad \forall q \in \mathcal{Q}, \tag{30}$$

where $K \in \mathbb{Z}^+$ is the local sparsity level within G_q.

On the hourly pickup trip time series of 2019 in Fig. 2(a), we use the Huber-SAR model ($\delta = 1$) with the following setting: Sparsity $\tau = 8$, candidate time lags $\mathcal{Q} = \{1, 24, 168\}$, radius $r = 5$, and local sparsity $K = 1$. That means that only one time lag in a windowed group $G_q, q \in \mathcal{Q}$ could be selected. The results of coefficient vector $\boldsymbol{w}$ are given by

$$\mathcal{P}_\Omega(\boldsymbol{w}) = (0.27, 0.18, 0.02, 0.05, 0.04, 0.02, 0.05, 0.37)^\top$$

with support set $\Omega = \{1, 24, 47, 71, 72, 95, 96, 168\}$. In contrast, we have the results of Huber-SAR without group sparsity constraints as follows,

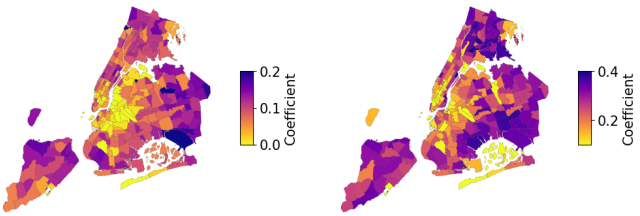
$$\mathcal{P}_\Omega(\boldsymbol{w}) = (0.26, 0.01, 0.17, 0.02, 0.07, 0.05, 0.12, 0.29)^\top$$

with support set $\Omega = \{1, 23, 24, 48, 72, 96, 167, 168\}$. The comparison demonstrates that the daily ($k = 24$) and weekly ($k = 168$) periodicities might be allocated to the nearby time lags (e.g., $k = 23, 167$) if there is no group sparsity constraint.

6.1.4 Periodicity of Ridesharing Trips

To discover the comparable periodicity of human mobility data across different spatial areas, we consider the hourly pickup trip time series of 2019 in NYC across 265 spatial areas. The hyperparameters of Huber-SAR are set as order $d = 168$, sparsity level $\tau = 6$, and Huber threshold $\delta = 1$. The support set is optimized as $\Omega = \{1, 23, 24, 143, 167, 168\}$, including daily and weekly cycles $k = 24, 168$. Fig. 4 shows the spatial patterns of the daily and weekly periodicities. The pickup trips of urban areas (e.g., Manhattan) in NYC are less daily/weekly periodic than the suburban areas because the travel behavior of urban areas is more complicated than

the suburban areas (e.g., commuting trips). These results highlight the importance of identifying spatial and temporal patterns from urban systems.



(a) Daily periodicity (b) Weekly periodicity

Fig. 4: Spatial patterns of the daily and weekly periodicities quantified by Huber-SAR with threshold $\delta = 1$ on the NYC ridesharing pickup trip time series in 2019.

6.2 Seasonality Patterns of North America Climate Variables

6.2.1 Model Setting

This case study investigates large-scale seasonal patterns in monthly climate variables across North America using long-term temperature and precipitation data (1980–2019). These data exhibit strong yearly seasonality and variability influenced by geography and climate change, and are both spatially and temporally structured. Considering a collection of time series arranged over an $M \times N$ spatial grid, we let $\boldsymbol{X}_{\gamma,t} \in \mathbb{R}^{M\times N}$ denote the spatial matrix at time $t \in [T_\gamma]$. Here, we partition the time series into $\Gamma \in \mathbb{Z}^+$ time segments, where each segment $\gamma \in [T_\gamma]$ contains $T_\gamma \in \mathbb{Z}^+$ time steps. Each grid cell (m, n) contains a a multivariate time series $\{x_{m,n,\gamma,t}\}_{t\in[T_\gamma],\gamma\in[\Gamma]}$, in which any γth univariate time series is of length T_γ. Following the same logic as the high-dimensional Huber-SAR, we model each time series with coefficient vector $\boldsymbol{w}_{m,n,\gamma} \in \mathbb{R}^d$, and impose a global support set across all grid cells and time segments.

Our goal is to uncover the geographical patterns of seasonality, track its evolution over decades, and assess the robustness of the learned patterns across spatial resolutions and time segment schemes. In climate systems, variables such as temperature and precipitation often exhibit strong seasonal structure. By setting the order $d = 12$, our model can select from time lags corresponding to past months in the year. For example, if $z_{12} = 1$ is the optimal solution, then it indicates that the data exhibits strong yearly seasonality, i.e., each month's value is relevant to its value one year ago. Other selected lags (e.g., z_1, z_{11}, etc.) can be interpreted as short-term auto-correlations or sub-seasonal effects. Since the climate variables with a monthly resolution are usually less noisy, we reduce our high-dimensional Huber-SAR model by choosing $\delta \to \infty$.

6.2.2 Dataset and Motivation for STV-SAR

We study the Daymet dataset from 1980 to 2019, which provides monthly long-term climate variables, including minimum/maximum temperature and precipitation, at a high spatial resolution of 1km×1km across North America [43].[2] Fig. 5 illustrates the minimum temperature data for January in the 2010s. These panels reveal strong seasonal structure and regional climate variability. Given the spatial scale and long-term coverage of this dataset, we aim to i) quantify yearly seasonality and its regional variation, and ii) examine how such patterns evolve across decades and spatial resolutions.

6.2.3 Global Support Set Verification

To enable scalable model training on high-dimensional spatiotemporal data, we aggregate the original 1 km×1 km dataset to a coarser resolution of 10 km×10 km. Even at this spatial resolution, the model must be fit across over 800,000 time series. To manage this computational burden, we employ the global support set selection outlined in Section 5.3. As shown in Table 3, the two-stage optimization scheme trains the full model across all time series in under 12 seconds, demonstrating its practical scalability for large-scale climate datasets.

We prescribe the sparsity level of our model as $\tau = 3$, aiming to capture dominant temporal dependencies that characterize yearly seasonality. For example, using the minimum temperature dataset, we construct monthly time series at each spatial cell over decadal segments (from 1980s to 2010s). Solving the global sparsity yields a common support set $\Omega = \{1, 11, 12\}$, which includes recent memory ($k = 1$), near-annual lags ($k = 11, 12$), and aligns with expected seasonal dynamics.

To assess the robustness of this global support set, we fit independent ℓ_2-SAR models to 204,153 time series at 10 km×10 km resolution using sparsity levels $\tau \in 3, 4, 5$. As shown in Fig. 6, even without enforcing any global constraint, the most frequently selected lags coincide with the global support set Ω. For $\tau = 3$, the dominant indices are $1, 11, 12$; for higher sparsity levels, the support expands to include nearby lags (e.g., $2, 10$), but $1, 11, 12$ remain prominent. This confirms that the global support set not only enhances computational efficiency but also captures shared temporal structures across space.

6.2.4 Results: Seasonality Patterns across Four Decades

Fig. 7 visualizes the yearly seasonality of monthly climate time series across North America from 1980 to 2019, using a spatial resolution of 10 km×10 km. Higher coefficients correspond to regions with stronger yearly periodicity. We begin by analyzing the seasonality patterns of minimum and maximum temperatures. Across the four decades, areas with stronger yearly seasonality are consistently concentrated in high-latitude regions, such as Canada and Alaska. In contrast, southern regions like Mexico exhibit weaker and more spatially variable seasonality. A notable expansion of highly seasonal regions occurs in the 2000s, suggesting an intensification of temperature-driven seasonal cycles. In the 2010s, two regions stand out: northern Canada (e.g., Nunavut) and the western United States. While the patterns in the 1980s and 1990s are relatively stable, the 2000s and 2010s show substantial shifts, highlighting evolving climate dynamics with potential ecological implications.

2. https://daac.ornl.gov/DAYMET

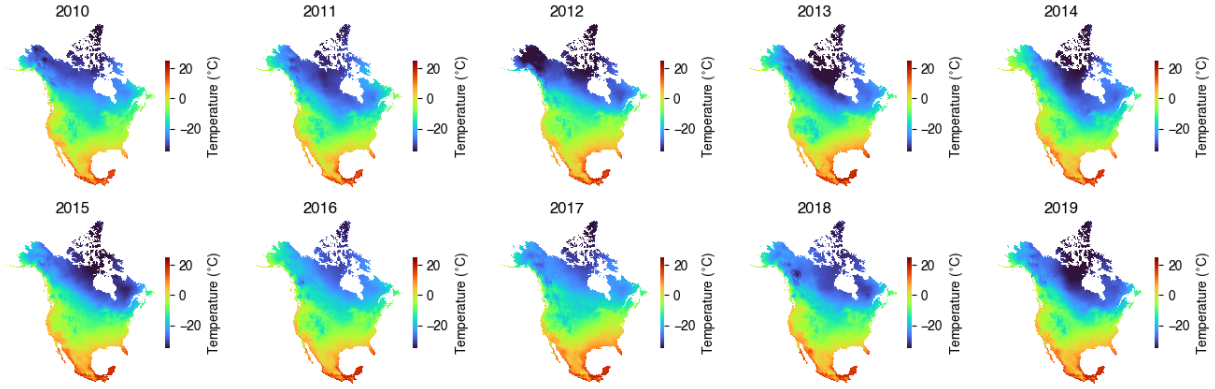
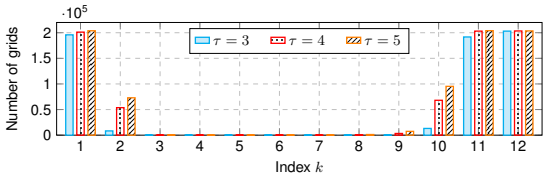
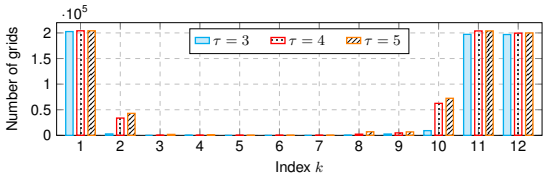


Fig. 5: Monthly aggregated minimum temperature across North America of January from 2010 to 2019. The color scale refers to the temperature in degrees Celsius, with blue and red tones indicating lower and higher temperatures, respectively.



(a) 2000s



(b) 2010s

Fig. 6: Support sets of independent SAR models with different sparsity levels tested on the minimum temperature data in 2000s and 2010s across 204,153 grids.

Next, we compare seasonal structures across different climate variables. The seasonality of precipitation exhibits significant spatial heterogeneity throughout the decades, even across small neighboring regions. Meanwhile, the seasonality of minimum and maximum temperatures remains relatively stable in the 1980s and 1990s. However, in the 2000s, minimum temperature seasonality intensifies and expands more broadly than that of maximum temperature. In the 2010s, Mexico shows higher seasonality in minimum temperature compared to maximum temperature, whereas the reverse holds in the western United States and central North America. Notably, in northern Canada, maximum temperature is less seasonal than minimum temperature. These contrasting patterns show various drivers and sensitivities of seasonal cycles across variables and geographies.

6.2.5 *Sensitivity Analysis: Spatial Resolution*

We now investigate the effect of spatial aggregation by comparing results at three different resolutions: 5 km×5 km, 10 km×10 km, and 20 km×20 km. Fig. 8 illustrates that the spatial pattern of yearly seasonality remains consistent across these resolutions. In particular, the northern Canada, the western United States, and the southeastern United States exhibit strong seasonal signals, while central North America displays substantially weaker seasonality. Temperature seasonality in Mexico shows high spatial variability and inconsistency among nearby areas, as highlighted in Fig. 9. Fine-grained resolution such as 5 km×5 km reveals localized structure that may be smoothed out at coarser scales. This demonstrates that our model captures robust patterns across resolutions while still preserving high-resolution detail where available.

TABLE 3: Runtime of the SAR model on the North America (minimum) temperature dataset with different spatial resolutions. Note that ν denotes the number of time series.

	5 km×5 km	10 km×10 km	20 km×20 km
Number ν	3,343,628	816,612	196,720
Runtime (s)	48.55	11.77	2.88

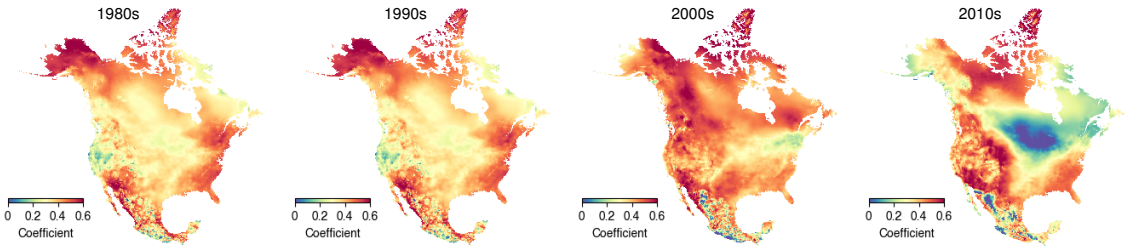
6.2.6 *Sensitivity Analysis: Phase Segmentation*

In our main analysis, we segmented the climate time series into decades to capture long-term trends. We now evaluate the sensitivity of our findings to the choice of segmentation granularity by re-estimating seasonality patterns using shorter, 5-year phases on the minimum temperature data from 2000 to 2019. Fig. 10(a) illustrates the spatial distribution of yearly seasonality from 2000 to 2004, which closely resembles the 1990s pattern shown in Fig. 7(a). In the 2005–2009 window (Fig. 10(b)), seasonality intensifies in several regions, aligning with the dominant pattern of the 2000s. From 2010 to 2014 (Fig. 10(c)), central North America shows noticeably weaker seasonal signals, whereas the period from 2015 to 2019 (Fig. 10(d)) marks a rebound in seasonality strength, particularly in Mexico. These subtle temporal shifts are less evident in the decade-level summary (Fig. 7), highlighting the value of finer segmentation for detecting short-term variations in seasonality patterns.

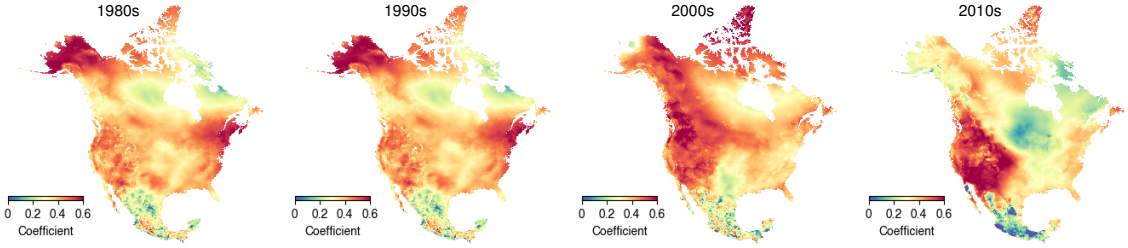
6.3 Seasonality Patterns of Sea Surface Temperature

6.3.1 *Dataset and Motivation for STV-SAR*

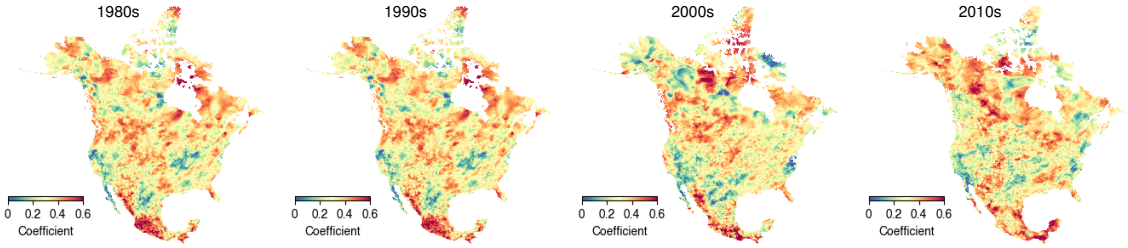
We study the NOAA Optimum Interpolation Sea Surface Temperature dataset, which provides monthly average temperatures over a $0.25° \times 0.25°$ global grid, comprising



(a) Minimum temperature

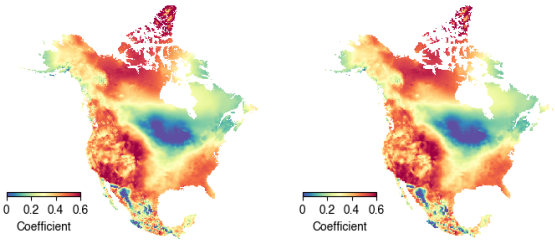


(b) Maximum temperature



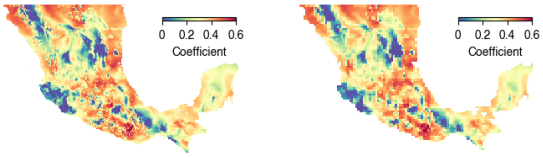
(c) Precipitation

Fig. 7: Spatial patterns of the strengths of yearly seasonality quantified by SAR on the North America climate data across the past four decades.



(a) 5 km × 5 km (b) 20 km × 20 km

Fig. 8: Spatial patterns of the strengths of yearly seasonality quantified by SAR on the minimum temperature data in 2010s.



(a) 5 km × 5 km (b) 20 km × 20 km

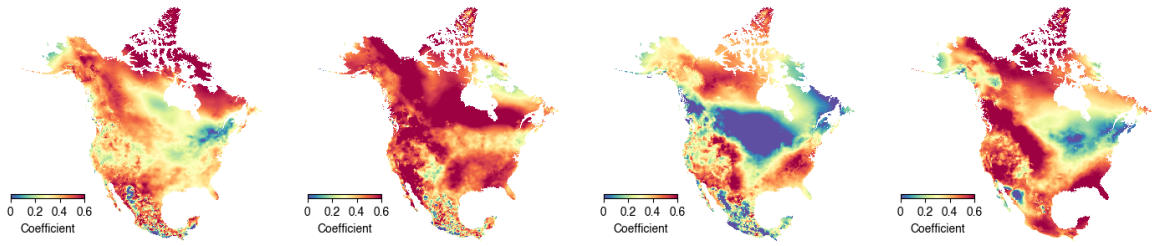
Fig. 9: Spatial patterns of the strengths of yearly seasonality on the minimum temperature data in 2010s within Mexico.

720×1440 cells (i.e., 1,036,800 spatial locations).[3] The dataset spans four decades, from January 1982 to December 2019. Fig. 11 presents both monthly and yearly average sea surface temperatures across this period. The monthly time series reveals strong seasonal fluctuations, while the yearly average trend shows a gradual warming signal consistent with global climate change. These dynamics highlight the importance of quantifying yearly seasonality, not only to understand recurring oceanic patterns but also to characterize the variability and potential instability in the climate system.

6.3.2 *Results: Spatial and Temporal Seasonality Patterns*

To analyze the temporal structure of sea surface temperatures, we fit the model separately on each decade of data. We

3. https://www.ncei.noaa.gov/data/sea-surface-temperature-optimum-interpolation/v2.1/access/avhrr/



(a) From 2000 to 2004 (b) From 2005 to 2009 (c) From 2010 to 2014 (d) From 2015 to 2019

Fig. 10: Spatial patterns of the strengths of yearly seasonality quantified by SAR on the minimum temperature data with a 5-year phase (i.e., $T_\gamma = 60$ months).

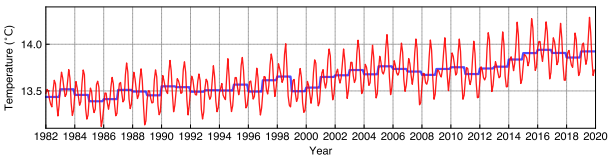


Fig. 11: Average values of the monthly sea surface temperature data from January 1982 to December 2019. The average temperature values in 1980s, 1990s, 2000s, and 2010s are 13.46°C, 13.54°C, 13.69°C, and 13.83°C, respectively. The blue and red curves correspond to yearly and monthly average temperature values, respectively.

fix the model order at $d = 12$ and sparsity level $\tau = 3$, which allows us to isolate and interpret coefficients corresponding to yearly periodicity. Fig. 12 shows the spatial distribution of the coefficient corresponding to lag $k = 12$, which serves as a proxy for yearly seasonality strength. Higher values (in red) indicate regions near Eastern and Central Asia, North America, Europe, and North Africa with stronger temperature seasonality. In contrast, equatorial regions such as the Pacific Ocean around the El Niño zone exhibit lower coefficients, reflecting irregular variation.

Despite the broad stability of global seasonal patterns across decades, we observe subtle temporal changes. For example, as shown in Fig. 13, the regions near Canada and the Arctic Ocean exhibit decreasing seasonality over time. This decline may signal evolving oceanic dynamics due to changes in ocean circulation, warming trends, or atmospheric variability. The Arctic Ocean, in particular, appears to have become notably less seasonal over the past two decades, potentially reflecting broader climate-induced disruptions in polar ocean systems.

7 Conclusion

This work introduces a unified Huber-SAR framework for quantifying periodicity in real-world time series data. We develop interpretable high-dimensional Huber-SAR model that incorporates temporally- and spatially-varying sparsity constraints to reveal structured periodic components. To address the computational challenges posed by the underlying MIQP problems, we propose scalable algorithmic solutions: an IRLS method and a two-stage optimization scheme. In particular, the IRLS method requires the estimation of both weights and weighted SAR, while the weighted SAR can be efficiently solved via the two-stage optimization scheme by assuming global sparsity. Applied to both human mobility and climate variable time series, our models successfully uncover meaningful, interpretable seasonal patterns and periodic structures. The learned spatial and temporal dynamics illustrate the practical relevance of periodicity quantification for analyzing complex systems. The Huber loss function and sparsity constraints also bring significant benefits for autoregressive prediction models.

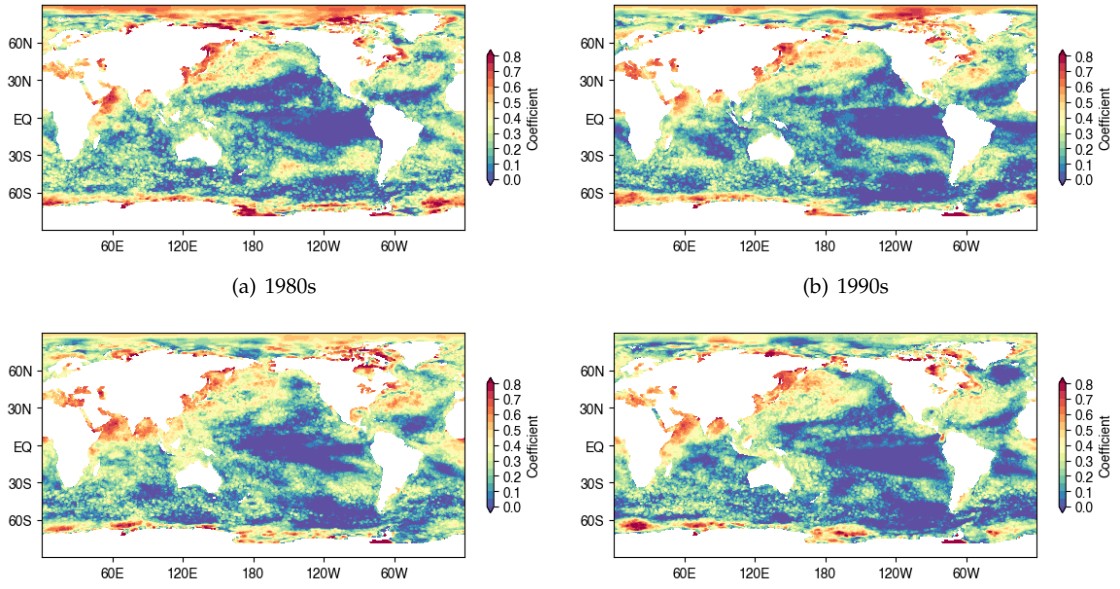
We point out the following directions for future improvement. The global support estimation strategy used in high-dimensional Huber-SAR is particularly effective when temporal structures are shared across time series, as in climate datasets (e.g., Fig. 6). However, for heterogeneous datasets with diverse or complex dynamics, the approximation introduced by the two-stage optimization may lead to estimation bias. Furthermore, heterogeneity in spatiotemporal data, e.g., varying density in NYC ridesharing data across different boroughs, can lead to a support set bias where the model over-prioritizes regions with higher signal-to-noise ratios. Addressing this gap potentially by integrating adaptive or hierarchical modeling techniques could further enhance the robustness and generality of the framework.

Acknowledgments

This research is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Vehicle Technology Program Award Number DE-EE0009211 and and DE-EE0011186. The Mens, Manus, and Machina (M3S) is an interdisciplinary research group (IRG) of the Singapore MIT Alliance for Research and Technology (SMART) center.

References

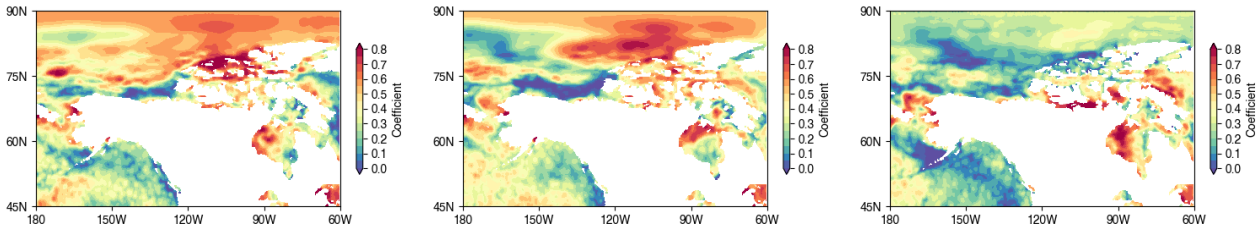
[1] S. Pezzulli, D. Stephenson, and A. Hannachi, "The variability of seasonality," *Journal of Climate*, vol. 18, no. 1, pp. 71–88, 2005.
[2] X. Feng, A. Porporato, and I. Rodriguez-Iturbe, "Changes in rainfall seasonality in the tropics," *Nature Climate Change*, vol. 3, no. 9, pp. 811–815, 2013.
[3] J. D. Tonkin, M. T. Bogan, N. Bonada, B. Rios-Touma, and D. A. Lytle, "Seasonality and predictability shape temporal species diversity," *Ecology*, vol. 98, no. 5, pp. 1201–1216, 2017.
[4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
[5] J. D. Hamilton, *Time series analysis*. Princeton University Press, 2020.



(a) 1980s (b) 1990s

(c) 2000s (d) 2010s

Fig. 12: Spatial patterns of the strengths of yearly seasonality quantified by SAR on the sea surface temperature data in the past four decades.



(a) 1980s (b) 1990s (c) 2010s

Fig. 13: Spatial patterns of the strengths of yearly seasonality on the sea surface temperature data with highlighted areas around Canada.

[6] D. Bertsimas, V. Digalakis Jr, M. L. Li, and O. S. Lami, "Slowly varying regression under sparsity," *Operations Research*, 2024.

[7] X. Chen, H. Cai, F. Liu, and J. Zhao, "Correlating time series with interpretable convolutional kernels," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[8] X. Chen, C. Zhang, X. Chen, N. Saunier, and L. Sun, "Discovering dynamic patterns from spatiotemporal data with time-varying low-rank autoregression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 2, pp. 504–517, 2024.

[9] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

[10] J. H. Cochrane, "Time series for macroeconomics and finance," 1997.

[11] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer Science & Business Media, 1991.

[12] J. M. Haslbeck, L. F. Bringmann, and L. J. Waldorp, "A tutorial on estimating time-varying vector autoregressive models," *Multivariate Behavioral Research*, vol. 56, no. 1, pp. 120–149, 2021.

[13] K. D. Harris, A. Aravkin, R. Rao, and B. W. Brunton, "Time-varying autoregression with low-rank tensors," *SIAM Journal on Applied Dynamical Systems*, vol. 20, no. 4, pp. 2335–2358, 2021.

[14] L. F. Bringmann, E. L. Hamaker, D. E. Vigo, A. Aubert, D. Borsboom, and F. Tuerlinckx, "Changing dynamics: Time-varying autoregressive models using generalized additive modeling." *Psychological Methods*, vol. 22, no. 3, p. 409, 2017.

[15] G. E. Primiceri, "Time varying structural vector autoregressions and monetary policy," *The Review of Economic Studies*, vol. 72, no. 3, pp. 821–852, 2005.

[16] Y. Nardi and A. Rinaldo, "Autoregressive process modeling via the lasso procedure," *Journal of Multivariate Analysis*, vol. 102, no. 3, pp. 528–549, 2011.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[18] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *The Annals of Statistics*, pp. 1535–1567, 2015.

[19] R. A. Davis, P. Zang, and T. Zheng, "Sparse vector autoregressive modeling," *Journal of Computational and Graphical Statistics*, vol. 25, no. 4, pp. 1077–1096, 2016.

[20] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[21] D. Bertsimas and W. Gurnee, "Learning sparse nonlinear dynamics via mixed-integer optimization," *Nonlinear Dynamics*, vol. 111, no. 7, pp. 6585–6604, 2023.

[22] J. Liu, S. Rosen, C. Zhong, and C. Rudin, "Okridge: Scalable optimal k-sparse ridge regression," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] E. Carrizosa, A. V. Olivares-Nadal, and P. Ramírez-Cobo, "A sparsity-controlled vector autoregressive model," *Biostatistics*, vol. 18, no. 2, pp. 244–259, 2017.

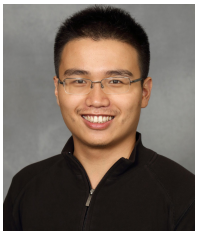
[24] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[25] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.
[26] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," *Advances in neural information processing systems*, vol. 32, 2019.
[27] D. Adil, R. Peng, and S. Sachdeva, "Fast, provably convergent irls algorithm for p-norm linear regression," *Advances in neural information processing systems*, vol. 32, 2019.
[28] D. Adil, R. Kyng, R. Peng, and S. Sachdeva, "Fast algorithms for p-regression," *Journal of the ACM*, vol. 71, no. 5, pp. 1–45, 2024.
[29] C. Kümmerle, C. Mayrink Verdun, and D. Stöger, "Iteratively reweighted least squares for basis pursuit with global linear convergence rate," *Advances in neural information processing systems*, vol. 34, pp. 2873–2886, 2021.
[30] L. Peng, C. Kümmerle, and R. Vidal, "Global linear and local superlinear convergence of irls for non-smooth robust regression," *Advances in neural information processing systems*, vol. 35, pp. 28 972–28 987, 2022.
[31] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.
[32] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.
[33] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.
[34] J. Fan, J. Lv, and L. Qi, "Sparse high-dimensional models in economics," *Annual Review of Economics*, vol. 3, no. 1, pp. 291–317, 2011.
[35] G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens, "A simple new approach to variable selection in regression, with application to genetic fine mapping," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 5, pp. 1273–1300, 2020.
[36] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *The Annals of Statistics*, vol. 44, no. 2, pp. 813–852, 2016.
[37] D. Bertsimas and B. Van Parys, "Sparse high-dimensional regression," *The Annals of Statistics*, vol. 48, no. 1, pp. 300–323, 2020.
[38] H. Hazimeh, R. Mazumder, and A. Saab, "Sparse regression at scale: Branch-and-bound rooted in first-order optimization," *Mathematical Programming*, vol. 196, no. 1, pp. 347–388, 2022.
[39] A. Atamturk and A. Gomez, "Rank-one convexification for sparse regression," *Journal of Machine Learning Research*, vol. 26, no. 35, pp. 1–50, 2025.
[40] S. Song and P. J. Bickel, "Large vector auto regressions," *arXiv preprint arXiv:1106.3915*, 2011.
[41] S. Basu, A. Shojaie, and G. Michailidis, "Network granger causality with inherent grouping structure," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 417–453, 2015.
[42] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.
[43] M. Thornton, R. Shrestha, Y. Wei, P. Thornton, and S.-C. Kao, "Daymet: Monthly climate summaries on a 1-km grid for north america, version 4 r1," 2022. [Online]. Available: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2131



Xinyu Chen is a Postdoctoral Associate at Massachusetts Institute of Technology, Cambridge, MA, United States. He received his Ph.D. degree from the University of Montreal, Montreal, QC, Canada. His current research centers on machine learning, spatiotemporal data modeling, intelligent transportation systems, and urban science.



Vassilis Digalakis Jr is an Assistant Professor of Operations & Technology Management at the O&T Management Department at Boston University's Questrom School of Business. He received his Ph.D. in Operations Research at the Operations Research Center at MIT. His research develops methods in trustworthy AI and machine learning, with applications in healthcare and sustainability.



Lijun Ding is an Assistant Professor at the Department of Mathematics at the University of California, San Diego. He earned his PhD in Operations Research from Cornell University in 2021. He specializes in continuous optimization, especially in semidefinite programming and applications of optimization in statistics.



Dingyi Zhuang is a Ph.D. student in Transportation Engineering at MIT Urban Mobility Lab. He received a B.S. degree in Mechanical Engineering from Shanghai Jiao Tong University in 2019, and an M.Eng. degree in Transportation Engineering from McGill University, in 2021. His research interests lie in deep learning, urban computing, and spatiotemporal data modeling.



Jinhua Zhao is currently the Professor of Cities and Transportation at MIT. He brings behavioral science and transportation technology together to shape travel behavior, design mobility systems, and reform urban policies. He directs the MIT Urban Mobility Laboratory and Public Transit Laboratory.